NEURO-SYMBOLIC FORENSIC REASONING FOR OPEN-GENERATOR AI-GENERATED VIDEO DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-generated video detection is critical for information integrity, yet remains hard in open-generator deployment: new generators shift artifact distributions, forensic cues are sparse and heterogeneous, and benchmark shortcuts can overstate progress. We argue that the root cause is the prevailing *binary classification* formulation $f(V) \to \{0, 1\}$, which learns generator-specific statistical shortcuts rather than transferable forensic principles, making detectors brittle to unseen generators and costly to retrain. We recast the problem as *forensic reasoning*: instead of a binary label, the system must produce a verdict together with a grounded evidence chain, $f(V) \to (\hat{y}, \mathcal{E})$. To instantiate this formulation, we propose FOVEA, a neuro-symbolic framework for **F**orensic **O**pen-generator **V**ideo **E**vidence **A**cquisition that trains no detector weights. A VLM planner forms forensic hypotheses and routes videos to specialized analysts, whose judgments are grounded by deterministic CV measurements including optical flow, frequency analysis, error-level analysis, temporal tracking, and geometric consistency. A reliability-aware judge fuses heterogeneous evidence via strong-evidence override, confidence-filtered voting, and disagreement-penalized aggregation to produce an interpretable verdict. On GenVideo, across ten unseen generators, FOVEA with Claude Opus 4.6 achieves 97.39% accuracy, outperforming the strongest supervised baseline, NSG-VD, by 5.93 points and reaching 0.99 AUROC, using only a one-time 100-video threshold calibration. The advantage persists after removing visible watermarks, on clips of at least 4 s, and on the GenVidBench benchmark.

1 INTRODUCTION

The rapid advancement of AI video generation, from GAN-based synthesizers to recent diffusion and transformer-based models such as Sora (OpenAI, 2024), Gen-2 (Runway Research, 2023), and Kling (Kuaishou Technology, 2024), has sharply lowered the barrier to producing photorealistic synthetic videos. While these models enable new forms of creative expression, they also make it possible to fabricate political speeches, legal evidence, news footage, and social-media content at scale, threatening information integrity and public trust (Vaccari & Chadwick, 2020; Chesney & Citron, 2019). Reliable detection is therefore increasingly urgent. In this work, we focus on *fully synthesized* videos, where every frame is generated from a text or image prompt, following the setting of the GenVideo benchmark (Chen et al., 2024). This setting spans diffusion-based models such as Sora, ModelScope, and Lavie; transformer-based systems such as OpenSora and Latte; and proprietary generators such as Gen-2 and Pika. We exclude face-swap deepfakes, localized editing, and traditional CGI/VFX, which involve different generation mechanisms and forensic signatures.

Open-generator AI-generated video detection. AI-generated video detection is especially challenging in the *open-generator* setting. Unlike closed-set classification, where training and test samples share similar artifact distributions, realistic deployment requires detecting videos produced by generators that were unavailable during detector training. New generators differ in architecture, training data, and post-processing, so the observable artifact distribution shifts rapidly. Moreover, the forensic evidence itself is sparse and heterogeneous: a fake video may reveal itself through semantic implausibility, distorted geometry, abnormal motion, high-frequency synthesis traces, temporal identity drift, generator watermarks, or none of these cues individually. Compounding this difficulty, existing benchmarks can contain shortcuts that inflate apparent progress; for example, we observe a duration

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

distribution bias in GenVideo where 8 of 10 test generators produce videos no longer than 4 s, while real videos typically span 10–30 s.

Existing methods and their limitations. Existing methods formulate AI-generated video detection as supervised binary classification $f_\theta(V) \to \{0, 1\}$ (Chen et al., 2024; Tan et al., 2024; Xu et al., 2023; Gu et al., 2021; Zhang et al., 2025). DeMamba (Chen et al., 2024) models temporal artifact patterns with a state-space architecture, while NSG-VD (Zhang et al., 2025) measures a physics-driven normalized spatiotemporal gradient with a pre-trained diffusion model. Despite their architectural differences, all such methods fit a parameterized decision boundary from labeled examples, making them vulnerable when new generators produce qualitatively different artifacts. This creates a structural asymmetry: generators can be released or updated continuously, while supervised detectors must repeatedly collect new labeled data, retrain, and redeploy. Moreover, the binary output $\{0, 1\}$ provides no forensic justification, limiting applicability in high-stakes scenarios. We therefore ask: *can neuro-symbolic forensic reasoning replace supervised binary classification* for open-generator AI-generated video detection?

Our proposal. We propose to recast the problem from binary classification to *forensic reasoning*: instead of learning a mapping $f_\theta(V) \to \{0, 1\}$, the system must produce a verdict together with its supporting evidence chain, $f(V) \to (\hat{y}, \mathcal{E})$. This neuro-symbolic instantiation eliminates the dependence on task-specific trainable parameters and requires every verdict to be traceable to concrete forensic observations. To instantiate this formulation, we introduce FOVEA, a neuro-symbolic framework for **F**orensic **O**pen-generator **V**ideo **E**vidence **A**cquisition. The key idea is to combine the semantic cognition of vision-language models (VLMs) (OpenAI, 2023; Gemini Team, Google, 2023) with the deterministic precision of classical computer vision (CV) algorithms. VLMs provide high-level visual understanding: they can reason about physical plausibility, scene semantics, artistic style, object interactions, and global inconsistencies. However, VLMs can hallucinate, may overlook subtle low-level artifacts, and often lack calibrated access to pixel-level or temporal measurements. Classical CV algorithms provide the complementary capability: optical flow, frequency-domain analysis, error-level analysis, feature tracking, and geometric consistency checks produce deterministic and interpretable measurements. FOVEA integrates these two modes so that neural hypotheses are checked against symbolic measurements and every verdict is traceable to the logged measurements behind it. FOVEA trains no task-specific model weights; its only labeled data is a one-time set of 100 calibration videos, drawn from generators disjoint from the test generators, used to fix a few scalar thresholds (Appendix A.4).

Challenges. Realizing this neuro-symbolic forensic reasoning idea is challenging. First, not every forensic dimension is relevant to every video; exhaustive analysis is inefficient because videos contain redundant frames, and the signal may concentrate in a single modality **(C1: evidence selection)**. Second, VLMs alone are unreliable forensic judges: they may confidently infer artifacts from unusual but real footage, or miss sub-pixel and frequency-domain traces that are invisible at the semantic level; every semantic claim must be anchored to deterministic measurements **(C2: evidence grounding)**. Third, the resulting evidence is heterogeneous and sometimes conflicting; for example, style cues may suggest synthesis while temporal motion appears natural, or a watermark cue may be strong but spatial evidence weak **(C3: evidence fusion)**.

FOVEA addresses these challenges through a coordinated neuro-symbolic pipeline (see Figure 1). A **Human Eyes** module first performs rapid screening on key frames for obvious high-confidence evidence such as anatomical collapse or visible generator watermarks (C1). When the evidence is not decisive, a **Planner** uses the initial observations and sampled frames to select only the forensic dimensions that are likely to be informative (C1). Five specialized **Analyst** modules then inspect complementary evidence types: style, physics, spatial consistency, temporal consistency, and watermark/text cues. Each analyst combines VLM reasoning with dynamically routed CV sub-skills, such as FFT spectral analysis, optical flow, ELA, boundary consistency, face detection, and feature tracking, and returns a structured evidence record grounded in deterministic measurements (C2). Finally, a **Judge** performs reliability-aware fusion using strong-evidence override, confidence-filtered majority voting, and disagreement-penalized weighted aggregation (C3), producing the final verdict-evidence pair $(\hat{y}, \mathcal{E})$. This design makes FOVEA more than a VLM prompt or a collection of tools: it is a neuro-symbolic forensic reasoning system that hypothesizes, measures, verifies, and fuses evidence at test time.

Contributions. Our notable contributions are summarized as follows:

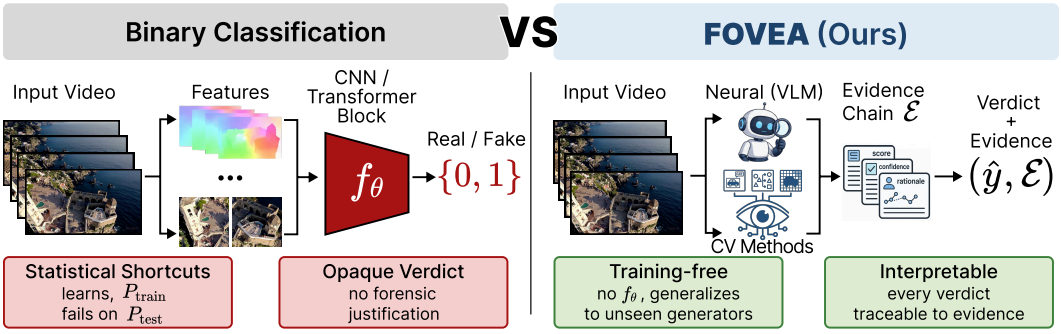

Figure 1: **From binary classification to neuro-symbolic forensic reasoning.** Supervised detectors learn a mapping $f_\theta(V) \to \{0, 1\}$ from generator-specific artifacts and must be retrained as new generators emerge. FOVEA instead produces $(\hat{y}, \mathcal{E})$: a verdict with a grounded evidence chain. A VLM planner forms forensic hypotheses (neural), symbolic CV tools verify them through deterministic measurements, and a reliability-aware judge fuses the evidence into an interpretable verdict.

1. We reformulate AI-generated video detection from binary classification $f_\theta(V) \to \{0, 1\}$ to *forensic reasoning* $f(V) \to (\hat{y}, \mathcal{E})$, requiring every verdict to be traceable to a measurement-linked evidence chain, and identify the neuro-symbolic constraint that effective evidence must satisfy.
2. We introduce FOVEA, a neuro-symbolic framework without detector training that instantiates this formulation. FOVEA combines VLM-based hypothesis formation with deterministic CV measurements through content-conditioned analyst and sub-skill routing.
3. We design a reliability-aware evidence fusion mechanism that resolves heterogeneous and conflicting forensic cues through strong-evidence override, confidence-filtered voting, and disagreement-penalized weighted aggregation.
4. Without detector training, FOVEA reaches 97.39% accuracy and 0.99 AUROC on GenVideo, matching the strongest recent supervised detectors in AUROC. Its advantage persists after watermark removal, on clips of at least 4 s, and on GenVidBench, and ablations show that CV grounding and sub-skill routing are essential.

2 Related Work

AI-Generated Video Detection. Recent supervised methods target diffusion-based generators: DeMamba (Chen et al., 2024) uses a state-space model for temporal artifacts; TALL (Xu et al., 2023) rearranges frames into thumbnail layouts; STIL (Gu et al., 2021) learns spatial-temporal inconsistency features; and NSG-VD (Zhang et al., 2025) uses a physics-driven normalized spatiotemporal gradient from a pre-trained diffusion model, reporting the strongest GenVideo results. Despite architectural diversity, all operate as *black-box pattern matchers* that learn generator-specific shortcuts rather than transferable forensic principles, requiring costly retraining when new generators emerge. Further recent supervised video detectors (Bai et al., 2024; Ma et al., 2025; None, 2025; Wang et al., 2026b;a) and training-free image and video detectors (Ricker et al., 2024; He et al., 2024; Cozzolino et al., 2024; Zheng et al., 2025; Ben Hayun et al., 2026) broaden this landscape; we compare with all ten using AUROC/AP (Appendix A.1). The broader media forensics field has produced methods for face swaps (Rössler et al., 2019) and generated images (Wang et al., 2020) using frequency analysis (Frank et al., 2020) and biological signals (Ciftci et al., 2020), but frame-level methods miss temporal artifacts unique to whole-video generation. Grounded and explainable image detectors localize artifacts with masks, boxes, or region crops (Kang et al., 2025; Ji et al., 2026b;a; Xiao et al., 2026); they are trained for still images and do not model temporal evidence (Appendix A.6).

Neuro-Symbolic Reasoning and VLM-Based Agents. Vision-language models (OpenAI, 2023; Gemini Team, Google, 2023) have enabled visual understanding beyond classification. LLMs have been deployed as reasoning and tool-using agents (Yao et al., 2023; Schick et al., 2023), and neuro-symbolic AI (d'Avila Garcez & Lamb, 2023) combines neural learning with symbolic reasoning. However, combining VLM reasoning with classical CV precision for video forgery detection has

received little attention. FOVEA addresses this gap through a neuro-symbolic agent framework that dynamically routes VLM hypotheses to symbolic CV tools.

Concurrent Work. Several concurrent works explore related directions. LAVID (Liu et al., 2025) proposes an agentic LVLM framework for video detection; TruthLens (Chakraborty et al., 2025) introduces training-free VLM-based image deepfake detection; RuleBoost (Raza et al., 2024) combines neuro-symbolic rules for deepfake detection; and VidGuard-R1 (Park et al., 2026) and Skyra (Li et al., 2025) apply reasoning MLLMs and independently note GenVideo's duration bias. FOVEA differs by: (i) linking analyst judgments to deterministic CV measurements via explicit sub-skill routing, without training a reasoning model or using a per-dataset labeled reference set; (ii) employing a multi-agent architecture with specialized analysts and a reliability-aware judge; and (iii) providing ablation evidence that CV grounding and adaptive routing are each essential (detailed comparison in Appendix A.5).

3 PROBLEM FORMULATION: FORENSIC REASONING FOR OPEN-GENERATOR AI-GENERATED VIDEO DETECTION

We focus on *fully synthesized* videos, where every frame is machine-generated from a text or image prompt. This setting includes three generator families: (i) *diffusion with U-Nets* (e.g., ModelScope (Wang et al., 2023), SVD, VideoCrafter); (ii) *diffusion with Transformers* (e.g., Sora (OpenAI, 2024), OpenSora, Latte); and (iii) *other or proprietary generators* (e.g., Gen-2 (Runway Research, 2023), Pika, CogVideo). We exclude face-swap deepfakes (Rössler et al., 2019), localized editing, and traditional CGI/VFX, which involve different generation mechanisms and forensic signatures.

3.1 THE LIMITATION OF BINARY CLASSIFICATION

Existing methods formulate AI-generated video detection as supervised binary classification. Given a test video $\tilde{V} = \{\tilde{I}_t\}_{t=1}^{T}$, they seek a learned mapping $f_\theta(\tilde{V}) \to \{0, 1\}$, where θ is fitted by minimizing a loss over a labeled training set drawn from a known generator set $\mathcal{G}_{\text{train}}$. This formulation has two structural deficiencies in the open-generator setting:

1. **Statistical shortcut dependence.** Because θ is optimized on the artifact distribution P_{train} of $\mathcal{G}_{\text{train}}$, the learned decision boundary often captures generator-specific shortcuts, such as characteristic resolutions, durations, or high-frequency noise patterns, rather than general forensic principles. When test generators $\mathcal{G}_{\text{test}}$ ($\mathcal{G}_{\text{train}} \cap \mathcal{G}_{\text{test}} = \emptyset$) produce a different artifact distribution P_{test}, these shortcuts fail.
2. **Opaque verdicts.** A binary output provides no justification for the decision. In high-stakes applications such as legal evidence authentication and misinformation detection, a bare $\{0, 1\}$ label is insufficient: the system must explain *what evidence* supports its conclusion and *how reliable* that evidence is.

3.2 FORENSIC REASONING FORMULATION

To address these limitations, we reformulate the problem as *forensic reasoning*. Rather than learning a parameterized mapping to $\{0, 1\}$, we define the detection task as producing a verdict together with its supporting evidence chain:

$$f(\tilde{V}) \;\to\; \big(\hat{y},\, \mathcal{E}\big), \tag{1}$$

where $\hat{y} \in \{\text{REAL}, \text{FAKE}\}$ is the final verdict and $\mathcal{E} = \{e_1, \ldots, e_m\}$ is the evidence chain: a set of structured forensic observations, each containing a forensic question, supporting measurements, a local score, a confidence estimate, and a textual rationale. This formulation shifts the burden from fitting statistical parameters θ to acquiring and reasoning over transferable forensic evidence at test time, and requires that every verdict be traceable to its supporting observations.

3.3 THE NEURO-SYMBOLIC INSTANTIATION CONSTRAINT

For the evidence chain $\mathcal{E}$ to be both *effective* and *trustworthy*, its generation must satisfy two complementary constraints:

1. **Semantic plausibility** (neural). Deciding *which* forensic dimensions to inspect requires high-level visual cognition: understanding scene semantics, physical plausibility, artistic style, and contextual coherence. VLMs provide this capability but can hallucinate and lack calibrated access to pixel-level measurements.
2. **Deterministic grounding** (symbolic). Each piece of evidence should preferably be anchored to verifiable, deterministic measurements $z = h(\tilde{V})$ produced by classical CV algorithms (optical flow, FFT, ELA, feature tracking, etc.), rather than relying solely on neural impressions. This ensures reproducibility and guards against hallucination. An exception applies to inherently semantic evidence (e.g., text content or logo recognition) where the forensic signal is categorical rather than pixel-level; in such cases, VLM reasoning alone may suffice.

These two constraints motivate a *neuro-symbolic* architecture: the neural side (VLMs) guides *what to examine* and *how to interpret* the results, while the symbolic side (CV tools) provides *the measurements themselves*. Section 4 presents FOVEA, our concrete instantiation of this neuro-symbolic forensic reasoning framework.

Evaluation protocol. Following GenVideo (Chen et al., 2024), supervised baselines are trained on one generator, Pika, and evaluated on ten unseen generators. FOVEA is evaluated on the same test protocol without task-specific detector training: all VLM backbones are frozen, CV tools are deterministic, and no generator-specific classifier is fitted. Since duration imbalance can act as a shortcut, we report both the standard cross-generator protocol and a duration-controlled protocol that removes the short-video bias. Performance is measured by Accuracy, Recall, and F1, and by threshold-free AUROC/AP computed from the continuous fake score.

4 THE FOVEA SOLUTION

4.1 OVERVIEW

FOVEA is a training-free neuro-symbolic framework for **F**orensic **O**pen-generator **V**ideo **E**vidence **A**cquisition. It instantiates the forensic reasoning formulation of Eq. 1: given a video $\tilde{V}$, FOVEA produces a verdict $\hat{y}$ together with a grounded evidence chain $\mathcal{E}$, rather than fitting a generator-specific decision boundary from labeled data.

As shown in Figure 2, FOVEA realizes the neuro-symbolic constraint (Section 3.3) through a coordinated pipeline. The *neural* component uses VLMs for semantic understanding, hypothesis formation, and planning; the *symbolic* component uses classical CV tools to produce deterministic measurements of motion, frequency, compression residuals, temporal tracks, and geometric consistency. The pipeline first builds an evidence cache from frames and metadata, then uses a **Human Eyes** module for rapid screening. If no decisive evidence is found, a **Planner** selects relevant **Analyst** modules, each of which combines VLM reasoning with routed CV sub-skills and returns a structured evidence record. Finally, a **Judge** performs reliability-aware fusion through strong-evidence override, confidence-filtered voting, and disagreement-penalized aggregation, yielding the final $(\hat{y}, \mathcal{E})$. Three design challenges arise: **(C1) evidence selection**, which forensic dimensions to inspect; **(C2) evidence grounding**, how to anchor semantic claims in deterministic measurements; and **(C3) evidence fusion**, how to resolve heterogeneous and conflicting evidence into a calibrated verdict.

4.2 EVIDENCE-ACQUISITION PIPELINE

Shortcut-controlled evidence extraction. Given $\tilde{V}$, FOVEA constructs an evidence cache $\mathcal{C} = \{\mathbf{V}, \mathcal{M}\}$: a uniformly sampled frame set $\mathbf{V}$ (4 fps for videos $\leq$60 s, 2 fps otherwise) and metadata $\mathcal{M}$ (duration, resolution, codec, frame rate). Metadata is used for sampling and logging but not as a decision cue; caching makes CV measurements reproducible.

Human Eyes: one-sided high-precision screening. The **Human Eyes** module performs a cheap first pass on three key frames (first, middle, last), searching for decisive fake-specific evidence along four dimensions: spatial artifacts, watermark presence, physics anomalies, and style incoherence. For each dimension d, it outputs $o_{\mathrm{HE}}^{(d)} = (s_{\mathrm{HE}}^{(d)}, c_{\mathrm{HE}}^{(d)}, r_{\mathrm{HE}}^{(d)})$ (fake-evidence score, confidence, rationale). If any dimension yields $s_{\mathrm{HE}}^{(d)} \geq \tau_{\mathrm{HE}}$ and $c_{\mathrm{HE}}^{(d)} \geq \gamma_{\mathrm{HE}}$ ($\tau_{\mathrm{HE}} = \gamma_{\mathrm{HE}} = 0.8$, fixed on calibration set), the

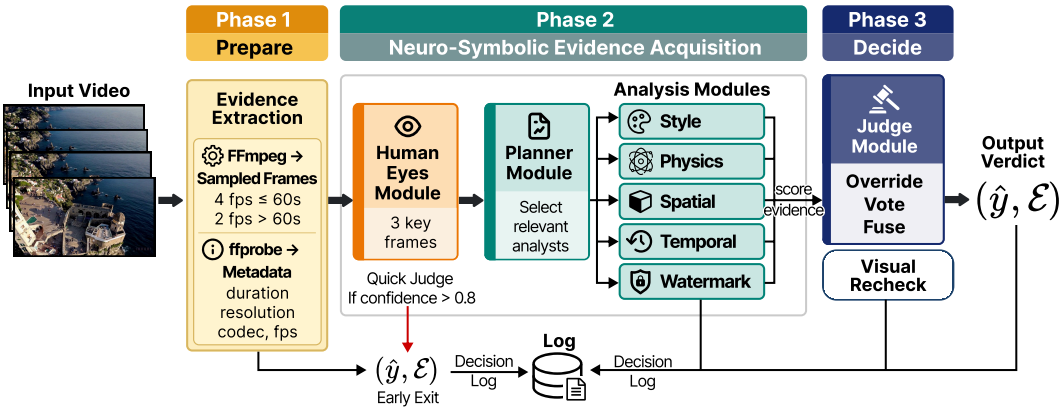

Figure 2: **Overview of FOVEA.** An input video is processed into an evidence cache (frames + metadata). The Human Eyes module performs early-exit screening. If the video is not obviously fake, the Planner selects relevant analyst modules, which run in parallel with dynamically routed CV sub-skills (symbolic). Each analyst combines VLM-based neural reasoning with symbolic CV measurements, outputting a structured evidence record. The Judge fuses all evidence into a final verdict-evidence pair $(\hat{y}, \mathcal{E})$.

video is classified as fake by early exit, catching 37.1% of fakes with high precision. Otherwise, the observation record $\mathcal{O}_{\text{HE}} = \{o_{\text{HE}}^{(d)}\}_{d \in \mathcal{D}}$ is forwarded to the Planner: sub-threshold cues are insufficient for a verdict but informative for routing.

Planner: content-conditioned analyst routing. The **Planner** addresses **(C1) evidence selection**. Given $\mathcal{C}$ and $\mathcal{O}_{\text{HE}}$, it selects a subset of analyst modules:

$$\mathcal{A}(\tilde{V}) = \{a \in \mathcal{A}_0 \mid \pi_a(\mathbf{V}, \mathcal{M}, \mathcal{O}_{\text{HE}}) = 1\},$$

where π_a is a VLM-based routing function reasoning over the Human Eyes record and the full frame set. For example, a sub-threshold style score activates the Style analyst for FFT analysis, while a static landscape with no faces deactivates Spatial and Temporal analysts. Activation frequencies are in Appendix G.3 (Table 19).

Analysts: measurement-grounded evidence contracts. The **Analyst** modules address **(C2) evidence grounding**. Five analysts target distinct forensic dimensions: **Style** (spectral/texture), **Physics** (motion/geometry), **Spatial** (ELA, boundaries, faces, blur), **Temporal** (phase coherence, feature tracking), and **Watermark** (logos/text). Each returns a structured evidence contract $\mathcal{E}_a = (q_a, \mathcal{R}_a, \{z_{a,k}\}_{k \in \mathcal{R}_a}, s_a, c_a, r_a)$, where q_a is the forensic question, $\mathcal{R}_a$ the selected CV sub-skill set, $z_{a,k} = h_k(\mathbf{V})$ deterministic measurements, $s_a \in [0,1]$ fake score, $c_a \in [0,1]$ confidence, and r_a a measurement-grounded rationale. Each analyst first routes sub-skills:

$$\mathcal{R}_a(\mathbf{V}, \mathcal{O}_{\text{HE}}) = \{k \in \mathcal{K}_a \mid \rho_a(q_a, \mathbf{V}, \mathcal{O}_{\text{HE}}, k) = 1\}, \tag{2}$$

executes the selected tools ($z_{a,k} = h_k(\mathbf{V})$), then combines them with VLM reasoning:

$$s_a,\ c_a,\ r_a = g_a\big(\mathbf{V}, q_a, \{(k, z_{a,k})\}_{k \in \mathcal{R}_a}\big), \tag{3}$$

where g_a is the VLM reasoning core. CV outputs are passed as JSON scalars (e.g., flow magnitude, FFT energy ratio) or rendered heatmap overlays (base64 images). This two-level routing avoids exhaustive execution and makes each analyst judgment traceable to the logged measurements it used. The full sub-skill inventory is in Table 20 (Appendix H).

Reliability calibration. Each analyst's confidence $c_a \in [0,1]$ is computed by a weakest-link rule:

$$c_a = \min\big(c_a^{\text{clarity}}, c_a^{\text{agree}}, c_a^{\text{quality}}\big), \tag{4}$$

where c_a^{clarity} measures evidence clarity, c_a^{agree} measures VLM–CV consistency, and c_a^{quality} accounts for occlusion, blur, and compression. This prevents high confidence when evidence is weak or measurements disagree. On a 100-video calibration set, predictions with $c_a \geq 0.8$ exceed 95% correctness; all thresholds are fixed before evaluation (Appendix A.4).

Judge agent: reliability-aware evidence fusion. The Judge addresses **(C3) evidence fusion**. Let $\mathcal{A}(\tilde{V})$ denote the active analyst set, each returning $(s_a, c_a, \mathcal{E}_a)$. The Judge uses a three-stage cascade:

1. **Strong-evidence override.** If any analyst reports high fake evidence with sufficient reliability,

$$\exists a \in \mathcal{A}(\tilde{V}) \quad s_a \geq \tau_{\text{ovr}} \text{ and } c_a \geq \gamma_{\text{ovr}}, \tag{5}$$

the Judge classifies the video as FAKE ($\tau_{\text{ovr}} = 0.85$, $\gamma_{\text{ovr}} = 0.6$). This rule is one-sided: absence of visible artifacts is not treated as evidence of being real.

2. **Confidence-filtered majority vote.** If no override applies, keep reliable analysts: $\mathcal{A}' = \{a \in \mathcal{A}(\tilde{V}) \mid c_a \geq \gamma\}$, $\gamma = 0.6$. Each casts a vote $\hat{y}_a = \mathbb{1}[s_a \geq 0.5]$; if $\geq$2 agree and form a strict majority, the Judge adopts it.

3. **Disagreement-penalized weighted fusion.** If no majority emerges (with $\mathcal{A}' \leftarrow \mathcal{A}(\tilde{V})$ as fallback when $\mathcal{A}'$ is empty; defaulting to REAL if all $c_a = 0$), compute:

$$\bar{s}_w = \frac{\sum_{a \in \mathcal{A}'} c_a s_a}{\sum_{a \in \mathcal{A}'} c_a}, \qquad \sigma_w = \sqrt{\frac{\sum_{a \in \mathcal{A}'} c_a (s_a - \bar{s}_w)^2}{\sum_{a \in \mathcal{A}'} c_a}}. \tag{6}$$

The final score is

$$s_{\text{final}} = \text{clip}\big(\bar{s}_w - \lambda \sigma_w,\, 0,\, 1\big), \qquad \lambda = 0.15. \tag{7}$$

The video is FAKE if $s_{\text{final}} \geq 0.5$, REAL otherwise. Final confidence:

$$c_{\text{final}} = \bar{c} \cdot (1 - \sigma_w), \qquad \bar{c} = \frac{\sum_{a \in \mathcal{A}'} c_a}{|\mathcal{A}'|}. \tag{8}$$

For Claude Opus 4.6, the Judge also receives frames alongside analyst reports for a final consistency review. This review is advisory only: it does *not* override the deterministic cascade.

4.3 IMPLEMENTATION DETAILS

FOVEA is implemented in Python using LangGraph (LangChain, Inc., 2024) with parallel analyst dispatch and typed JSON schemas. VLM reasoning uses Claude Opus 4.6 (Anthropic, 2026) as the primary backbone; ablations use GPT-5.2 (OpenAI, 2025) and Gemini 3.1 Pro (Google DeepMind, 2026). Frames are uniformly sampled, base64-encoded, and capped at 40 per VLM call. All symbolic sub-skills are deterministic CPU-only OpenCV (Bradski, 2000) routines. Results are cached (filesystem + SQLite) for reproducibility. All thresholds are fixed once on 100 labeled calibration videos from GenVideo's training generators, which are disjoint from the test generators, and are shared by all backbones and ablations; no task-specific parameters are trained (Appendix A.4). Full details in Appendix H and I.

5 EXPERIMENTS

We evaluate: **(Q1)** Does FOVEA outperform supervised detectors on unseen generators? **(Q2)** Does the advantage hold without ultra-short clips, without visible watermarks, and on another benchmark? **(Q3)** Which components drive the gains? **(Q4)** What are inference costs?

Benchmark and baselines. We evaluate on GenVideo (Chen et al., 2024): 8,588 fake videos from ten unseen generators and 10,000 real videos (MSR-VTT (Xu et al., 2016)). Supervised baselines train on 10,000 real/fake pairs (Kinetics-400 (Kay et al., 2017) + Pika) per Zhang et al. (2025); FOVEA trains no detector weights and uses frozen VLMs, deterministic CV tools, and fixed thresholds. Since eight of ten generators produce clips $\leq$4 s while real videos span 10–30 s, we report both standard and duration-controlled ($\geq$4 s) evaluations. We compare against DeMamba (Chen et al., 2024), NPR (Tan et al., 2024), TALL (Xu et al., 2023), STIL (Gu et al., 2021), and NSG-VD (Zhang et al., 2025) (strongest supervised baseline), and additionally with ten recent supervised and training-free detectors using AUROC/AP. Details in Appendix J and K.

Exp-1: Standard open-generator evaluation. Table 1 reports per-generator and average performance under the standard GenVideo protocol.

FOVEA achieves **97.39% accuracy** and **97.84% F1**, outperforming NSG-VD by 5.93 accuracy points and 6.97 F1 points. Micro-averaged accuracy across all 18,588 test videos is 97.11%, confirming the result is not driven by small-sample generators. Unlike prior methods, FOVEA introduces zero

Table 1: **Comparisons with baselines on the standard evaluation (%).** Methods marked † are **supervised** (trained on 10k Kinetics-400 + Pika pairs). FOVEA trains **no detector weights** (Claude Opus 4.6). †Baseline numbers from Zhang et al. (2025). Best average in **bold**.

Method	Metric	Model Scope	Morph Studio	Moon Valley	HotShot	Show1	Gen2	Crafter	Lavie	Sora	Wild Scrape	Avg.
DeMamba†	Recall	87.00	93.60	98.80	40.60	48.40	98.00	88.40	59.00	48.21	58.20	72.02
	Accuracy	91.70	95.00	97.60	68.50	72.40	97.20	92.40	77.70	72.32	77.30	84.21
	F1	91.29	94.93	97.63	56.31	63.68	97.22	92.08	72.57	63.53	71.94	80.12
NPR†	Recall	61.20	80.00	98.00	16.00	33.00	91.20	80.60	34.60	35.71	43.20	57.35
	Accuracy	79.80	89.20	98.20	57.20	65.70	94.80	89.50	66.50	67.86	70.80	77.96
	F1	75.18	88.11	98.20	27.21	49.03	94.61	88.47	50.81	52.63	59.67	68.39
TALL†	Recall	51.20	65.20	93.40	32.00	61.60	94.80	81.80	49.20	25.00	53.60	60.78
	Accuracy	75.10	82.10	96.20	65.50	80.30	96.90	90.40	74.10	61.61	76.30	79.85
	F1	67.28	78.46	96.09	48.12	75.77	96.83	89.50	65.51	39.44	69.34	72.63
STIL†	Recall	73.80	70.80	43.40	1.40	2.00	45.00	13.20	7.20	1.79	11.60	27.02
	Accuracy	86.90	85.40	71.70	50.70	51.00	72.50	56.60	53.60	50.89	55.80	63.51
	F1	84.93	82.90	60.53	2.76	3.92	62.07	23.32	13.43	3.51	20.79	35.82
NSG-VD†	Recall	68.33	98.33	100.00	92.50	87.50	80.00	98.33	94.17	78.57	82.50	88.02
	Accuracy	81.67	98.33	96.67	91.67	90.83	88.33	95.83	94.17	88.39	88.75	91.46
	F1	78.85	98.33	96.77	91.74	90.52	87.27	95.93	94.17	87.13	88.00	90.87
FOVEA (Ours)	Recall	100.00	100.00	100.00	92.10	89.97	100.00	96.13	96.30	100.00	84.95	**95.90**
	Accuracy	100.00	100.00	100.00	94.29	92.86	100.00	97.14	97.14	100.00	92.50	**97.39**
	F1	100.00	100.00	100.00	95.83	94.74	100.00	97.96	97.96	100.00	91.89	**97.84**

trainable parameters, requires no retraining when new generators appear, and produces interpretable evidence chains (paradigm comparison in Table 9, Appendix B). Threshold-free metrics agree: from its continuous fake score, FOVEA reaches 0.99 AUROC and 1.00 AP. This matches the strongest recent supervised detector, CMTA (0.99/0.99), exceeds ATSS (0.97/0.98), ReStraV (0.92/0.90), and DeCoF (0.92/0.82), and far exceeds every training-free detector (best: STALL, 0.80/0.80) (Table 3, Appendix A.1).

Exp-2: Short-clip robustness. To test whether the advantage survives removal of ultra-short clips, we exclude videos <4 s.

As shown in Table 8 (Appendix A.10), four generators pass the ≥ 4 s filter. FOVEA achieves **98.13% accuracy** and **97.97% F1**; NSG-VD achieves 89.76%, yielding an 8.37-point gap (vs. 5.93 under standard). NSG-VD's standard-protocol accuracy on these same four generators is 86.79% (from Table 1), so the filter does not degrade it; the widened margin reflects FOVEA's consistent performance. This is a short-clip robustness test rather than full duration deconfounding. Only four generators and 27.8% of fake videos remain, retained real videos still span 10–30 s, and the four-generator average is not comparable to Table 1 (Appendix A.11).

Exp-3: Watermark-removal control. Because 89% of Human Eyes early exits involve watermarked videos, visible watermarks could act as a shortcut. We crop the bottom-right corner (about 5% of the frame area) from every test video, manually verify that no visible watermark remains, and rerun full FOVEA with GPT-5.2. Pooled accuracy rises slightly from 89.47% to 91.03%, and mean per-generator recall from 0.90 to 0.91 (Table 4, Appendix A.2). FOVEA therefore does not rely on the visible watermark.

Exp-4: Cross-benchmark transfer. On the GenVidBench-143K test set (Ni et al., 2026), full FOVEA with GPT-5.2 reaches 85.71% mean Top-1 accuracy, the highest among the listed methods and 2.44 points above the strongest trained video backbone, MViTv2-B (Table 5, Appendix A.3).

Exp-5: Ablation study. We next examine which parts of FOVEA are responsible for its performance. To avoid ceiling effects from Claude Opus 4.6, which reaches 100% accuracy on several generators, we conduct ablations with GPT-5.2 as the VLM backbone. Table 2 reports accuracy changes relative to full FOVEA; full per-generator breakdowns are provided in Appendix C.

The ablations show that FOVEA is not merely a VLM prompt wrapped in a pipeline. Removing CV sub-skills reduces accuracy by 13.02 points (15.53 under duration control), showing symbolic

Table 2: **Ablation summary (%).** Accuracy change ΔAcc relative to full FOVEA (GPT-5.2). Full per-generator tables in Appendix C.

	Standard (10 gen.)		**Duration ≥4 s (4 gen.)**	
Configuration	Acc.	ΔAcc	Acc.	ΔAcc
FOVEA (full, GPT-5.2)	88.27	—	91.32	—
w/o Sub-skill routing	75.71	−12.56	79.93	−11.39
w/o CV sub-skills (VLM-only)	75.25	−13.02	75.79	−15.53
w/o Style Analyst	82.88	−5.39	85.61	−5.71
w/o Watermark Analyst	83.11	−5.16	82.94	−8.38
w/o Spatial Analyst	84.89	−3.38	83.99	−7.33
w/o Temporal Analyst	86.32	−1.95	91.88	+0.56
w/o Physics Analyst	88.93	+0.66	91.05	−0.27
w/o Planner routing	87.51	−0.76	89.59	−1.73
w/o Human Eyes (early-exit)	89.68	+1.41	89.77	−1.55
Pure VLM (no framework)	42.92	−45.35	39.31	−52.01

measurements are essential for grounding. Removing sub-skill routing causes a 12.56-point drop, confirming adaptive tool selection is accuracy-critical, not just an efficiency mechanism. Together, these validate **(C1) evidence selection** through routing and **(C2) evidence grounding** through deterministic CV measurements.

The Pure VLM baseline (42.92%) demonstrates that the improvement comes from the structured neuro-symbolic process, not the VLM backbone alone. Individual analyst ablations reveal that Style and Watermark are most influential under standard evaluation, while Watermark and Spatial become critical under duration control, consistent with the heterogeneous-evidence view that different generators expose different cues. Human Eyes and Physics help on the ≥4 s cohort but slightly hurt on the standard one, so their effect is cohort-dependent (Appendix A.9).

Exp-6: VLM backbone sensitivity. Table 13 (Appendix D) evaluates FOVEA with different VLM backbones. Claude Opus 4.6 provides the strongest performance (97.39% standard, 98.13% duration-controlled). GPT-5.2 is weaker but still outperforms supervised NSG-VD under duration control. Gemini 3.1 Pro achieves high recall but lower accuracy due to more false positives.

Exp-7: Inference cost. FOVEA has higher per-video cost than feed-forward detectors due to VLM API calls: 53.4 s and USD 0.25 per video at the evaluated operating point, of which local CV computation takes about 5 s. Videos are independent and can be processed concurrently subject to API rate limits. We did not time the baselines on identical hardware and make no efficiency-superiority claim (Appendices E and A.8).

6 CONCLUSION

This paper asked whether forensic reasoning can replace detector training for AI-generated video detection. FOVEA, a neuro-symbolic reasoning framework that trains no detector weights, achieves 97.39% accuracy and 0.99 AUROC on GenVideo with Claude Opus 4.6, outperforming the strongest supervised method NSG-VD (91.46%) by 5.93 percentage points. Its advantage persists on clips of at least 4 s, after watermark removal, and on GenVidBench. The framework integrates VLM semantic cognition (neural) with classical CV algorithms (symbolic) through dynamic sub-skill routing, producing verdicts that are traceable to logged forensic measurements.

Limitations. Performance is backbone-dependent (GPT-5.2: 88.27%; Gemini 3.1 Pro: 86.57%), the Human Eyes gate predominantly catches watermarked videos (89% of early catches), and per-video API cost ($0.25 avg.) limits real-time deployment. The evidence chains are traceable to logged measurements, but their correctness and causal relevance have not been independently validated. The ≥4 s control is not a duration-matched deconfounding. Open-source VLM backbones are untested, and exposure of proprietary VLMs to benchmark videos cannot be ruled out. Robustness to adversarial manipulation remains untested.

AI USE STATEMENT

Vision-language models are a core component of the proposed method; the exact models and versions are specified in Section 4.3, and sensitivity to the backbone is evaluated in Section 5. In addition, we used generative AI tools to assist with writing, editing, formatting, and reference management. All bibliographic entries were checked against authoritative sources (DBLP, arXiv, and publisher pages), and the authors reviewed all AI-assisted content. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only publicly available benchmarks and does not involve human subjects or crowdsourcing. FOVEA is a detection tool rather than a generative model; it requires no training and releases no pre-trained models, datasets, or scraped data. While defending against misinformation is our primary motivation, detection tools can be misused for censorship or discrediting authentic content; the interpretable evidence-chain design mitigates this risk by making every verdict auditable and contestable. Verdicts and evidence chains are decision-support artifacts: false positives should not trigger censorship, sanctions, or authenticity claims without human verification, and fluent rationales are not proof. Deployments that send frames to commercial VLM APIs should apply privacy, consent, retention, and access controls (Appendix A.11).

REPRODUCIBILITY STATEMENT

An anonymized code repository is available at `https://anonymous.4open.science/r/FOVEA-23FA`. The evaluation benchmark (GenVideo) is publicly available (Chen et al., 2024). Implementation details, including VLM backbones and CV tools, are given in Section 4.3 and Appendix H; the calibration set and every calibrated constant are specified in Appendix A.4. The execution log of the evaluated runs is included in the repository.

REFERENCES

Anthropic. Claude Opus 4.6 system card. `https://www.anthropic.com/claude-opus-4-6-system-card`, 2026.

Jianfa Bai, Man Lin, Gang Cao, and Zijie Lou. AI-generated video detection via spatial-temporal anomaly learning. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, volume 15040, pp. 460–470, 2024.

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

Omer Ben Hayun, Roy Betser, Meir Yossef Levi, Levi Kassel, and Guy Gilboa. Training-free detection of generated videos via spatial-temporal likelihoods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16299–16310, 2026.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning (ICML)*, volume 139, pp. 813–824, 2021.

Gary Bradski. The OpenCV library. *Dr. Dobb's Journal of Software Tools*, 2000.

Ritabrata Chakraborty, Rajatsubhra Chakraborty, Ali Khaleghi Rahimian, and Thomas MacDougall. TruthLens: A training-free paradigm for DeepFake detection. *arXiv preprint arXiv:2503.15342*, 2025.

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

Robert Chesney and Danielle Citron. Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107(6):1753–1820, 2019.

Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. FakeCatcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Davide Cozzolino, Giovanni Poggi, Matthias Nießner, and Luisa Verdoliva. Zero-shot detection of AI-generated images. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, volume 15076 of *Lecture Notes in Computer Science*, pp. 54–72. Springer, 2024. doi: 10.1007/978-3-031-72649-1_4.

Artur d'Avila Garcez and Luis C. Lamb. Neurosymbolic AI: The 3rd wave. *Artificial Intelligence Review*, 56(11):12387–12406, 2023. arXiv:2012.05876.

Christoph Feichtenhofer. X3D: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 200–210, 2020.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6201–6210, 2019.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3247–3258, 2020.

Gemini Team, Google. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Google DeepMind. Gemini 3.1 pro. `https://deepmind.google/models/model-cards/gemini-3-1-pro/`, 2026.

Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Spatiotemporal inconsistency learning for deepfake video detection. In *Proceedings of the 29th ACM International Conference on Multimedia (ACM MM)*, 2021.

Zhiyuan He, Pin-Yu Chen, and Tsung-Yi Ho. RIGID: A training-free and model-agnostic framework for robust AI-generated image detection. *arXiv preprint arXiv:2405.20112*, 2024.

Yikun Ji, Yan Hong, Bowen Deng, Jun Lan, Huijia Zhu, Weiqiang Wang, Liqing Zhang, and Jianfu Zhang. Locate-then-examine: Grounded region reasoning improves detection of AI-generated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19165–19175, June 2026a.

Yikun Ji, Yan Hong, Qi Fan, Jun Lan, Huijia Zhu, Weiqiang Wang, Liqing Zhang, and Jianfu Zhang. FakeXplain: AI-generated image detection via human-aligned grounded reasoning. In *International Conference on Learning Representations (ICLR)*, 2026b. URL `https://openreview.net/forum?id=UcpTOa8OnG`.

Hengrui Kang, Siwei Wen, Zichen Wen, Junyan Ye, Weijia Li, Peilin Feng, Baichuan Zhou, Bin Wang, Dahua Lin, Linfeng Zhang, and Conghui He. LEGION: Learning to ground and explain for synthetic image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18937–18947, October 2025.

Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Kuaishou Technology. Kling: AI video generation. `https://klingai.com`, 2024.

LangChain, Inc. LangGraph: Build resilient agents. `https://github.com/langchain-ai/langgraph`, 2024.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. UniFormerV2: Unlocking the potential of image ViTs for video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1632–1643, 2023.

Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. MViTv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4794–4804, 2022.

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

Ji Lin, Chuang Gan, and Song Han. TSM: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7082–7092, 2019.

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

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. EvalCrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22139–22149, 2024.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video Swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3192–3201, 2022.

Long Ma, Zhiyuan Yan, Qinglang Guo, Yong Liao, Haiyang Yu, and Pengyuan Zhou. Detecting AI-generated video via frame consistency. In *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2025.

Meta. Reels | Instagram help center. `https://help.instagram.com/270447560766967`, 2025. Maximum length: 3 minutes; Accessed: 2025-12-01.

Zhenliang Ni, Qiangyu Yan, Mouxiao Huang, Tianning Yuan, Yehui Tang, Hailin Hu, Xinghao Chen, and Yunhe Wang. GenVidBench: A 6-million benchmark for AI-generated video detection. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 15582–15590, 2026.

None. Quo vadis, action recognition? A new model and the Kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–4733, 2017.

None. AI-generated video detection via perceptual straightening. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

OpenAI. GPT-4V(ision) system card. `https://cdn.openai.com/papers/GPTV_System_Card.pdf`, 2023.

OpenAI. Sora: Creating video from text. `https://openai.com/sora`, 2024.

OpenAI. Update to GPT-5 system card: GPT-5.2. `https://openai.com/index/gpt-5-system-card-update-gpt-5-2/`, 2025. December 11, 2025.

Kyoungjun Park, Yifan Yang, Juheon Yi, Shicheng Zheng, Yifei Shen, Dongqi Han, Caihua Shan, Muhammad Muaz, and Lili Qiu. VidGuard-R1: AI-generated video detection and explanation via reasoning MLLMs and RL. In *International Conference on Learning Representations (ICLR)*, 2026.

Muhammad Anas Raza, Khalid Mahmood Malik, and Ijaz Ul Haq. RuleBoost: A neuro-symbolic framework for robust deepfake detection. In *IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–10, 2024. doi: 10.1109/IJCB62174.2024.10744498.

Jonas Ricker, Denis Lukovnikov, and Asja Fischer. AEROBLADE: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9130–9140, 2024.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Runway Research. Gen-2: The next step forward for generative AI. `https://runway.com/research/gen-2`, 2023.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 68539–68551, 2023.

Hao Shao, Shengju Qian, and Yu Liu. Temporal interlacing network. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 11966–11973, 2020.

Chuangchuang Tan, Huan Liu, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28130–28139, 2024.

TikTok. TikTok newsroom — product updates and platform statistics. `https://newsroom.tiktok.com/`, 2024. Accessed: 2025-12-01.

Cristian Vaccari and Andrew Chadwick. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, 6(1), 2020. doi: 10.1177/2056305120903408.

Hang Wang, Chao Shen, Chenhao Lin, Minghui Yang, Lei Zhang, and Cong Wang. CMTA: Leveraging cross-modal temporal artifacts for generalizable AI-generated video detection. *arXiv preprint arXiv:2605.00630*, 2026a.

Hang Wang, Chao Shen, Lei Zhang, and Zhi-Qi Cheng. ATSS: Detecting AI-generated videos via anomalous temporal self-similarity. *arXiv preprint arXiv:2604.04029*, 2026b.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. ModelScope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Wenhao Wang and Yi Yang. VidProM: A million-scale real prompt-gallery dataset for text-to-video diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Yao Xiao, Weiyan Chen, Jiahao Chen, Zijie Cao, Weijian Deng, Binbin Yang, Ziyi Dong, Xiangyang Ji, Wei Ke, Pengxu Wei, and Liang Lin. Unveiling perceptual artifacts: A fine-grained benchmark for interpretable AI-generated image detection. In *International Conference on Learning Representations (ICLR)*, 2026. URL `https://openreview.net/forum?id=Tk8ujiOgHM`.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. TALL: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 588–597, 2020.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

YouTube. YouTube about press — YouTube by the numbers. `https://blog.youtube/press/`, 2024a. Accessed: 2025-12-01.

YouTube. Get started creating YouTube Shorts. `https://support.google.com/youtube/answer/10059070`, 2024b. Maximum length: 3 minutes; Accessed: 2025-12-01.

Shuhai Zhang, Zihao Lian, Jiahao Yang, Daiyuan Li, Guoxuan Pang, Feng Liu, Bo Han, Shutao Li, and Mingkui Tan. Physics-driven spatiotemporal modeling for AI-generated video detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. Spotlight; arXiv:2510.08073.

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

Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *European Conference on Computer Vision (ECCV)*, volume 11205, pp. 831–846, 2018.

A Additional Evaluations and Clarifications

This section reports additional evaluations: threshold-free metrics against recent detectors, a watermark-removal control, cross-benchmark transfer, the calibration protocol, comparisons with the closest reasoning-based and grounded detectors, and the precise scope of our evidence and duration claims.

A.1 Threshold-Free Metrics and Recent Detectors

FOVEA produces a continuous fake score in addition to its verdict, so threshold-free metrics can be computed directly. Table 3 reports AUROC and AP on the standard ten-generator GenVideo protocol, with fake as the positive class. We compare with training-free detectors (AEROBLADE (Ricker et al., 2024), RIGID (He et al., 2024), ZED (Cozzolino et al., 2024), D3 (Zheng et al., 2025), STALL (Ben Hayun et al., 2026)) and recent supervised video detectors (AIGVDet (Bai et al., 2024), DeCoF (Ma et al., 2025), ReStraV (None, 2025), ATSS (Wang et al., 2026b), CMTA (Wang et al., 2026a)).

Among training-free detectors, FOVEA is the only one above 0.80 average AUROC. With Claude Opus 4.6 it matches the strongest supervised detector (CMTA, 0.99/0.99) without fitting detector weights. With GPT-5.2 it trails the three strongest supervised detectors in AUROC. Gemini 3.1 Pro attains 0.99 AUROC even though its fixed-threshold accuracy is the lowest of the three backbones (Table 13): suggesting that its errors stem mainly from threshold placement (false positives at $s_{\text{final}} \geq 0.5$) rather than from poor ranking.

A.2 Watermark-Removal Control

Visible generator watermarks could act as a dataset shortcut rather than a transferable forensic cue. To test this, we audited GenVideo, applied the same bottom-right crop to *every* test video (real and fake), and manually verified that no visible watermark remained. The crop covers about 5% of the image area, leaves the central content intact, and preserves the full paired cohort. We then reran full FOVEA with GPT-5.2.

FOVEA remains effective without the visible cue: pooled accuracy rises slightly from .8947 to .9103, and mean recall from .90 to .91. This input-level control is stricter than the Watermark-Analyst ablation of Table 2 (88.27→83.11 standard, 91.32→82.94 duration-controlled), because in that

Table 3: **AUROC and AP on the standard GenVideo protocol** (decimals; fake is the positive class). "Det. training" indicates whether task-specific detector weights are fitted. ✗ does not mean data-independent: ZED uses a learned real-only encoder, STALL uses real-only calibration, and FOVEA uses a one-time 100-video scalar-threshold calibration without weight updates (Appendix A.4). AIGVDet, DeCoF, ReStraV, ATSS, and CMTA fit task-specific detector weights; ReStraV's "zero-shot" setting means testing on generators unseen during fitting. Baseline values are as reported by Ben Hayun et al. (2026) (Table 1; AEROBLADE, RIGID, ZED, D3, STALL), Wang et al. (2026b) (Tables I–II; AIGVDet, DeCoF, ReStraV, ATSS), and Wang et al. (2026a) (Tables I–II; CMTA), all on the same ten GenVideo generators.

Method	**Det. training**	**Metric**	MS	Morph	Moon	Hot	Show	Gen2	Craft	LaVie	Sora	Wild	**Avg.**
AEROBLADE	✗	AUROC	.60	.74	.75	.31	.48	.56	.64	.58	.73	.49	.59
		AP	.60	.73	.78	.39	.50	.59	.65	.59	.70	.53	.61
RIGID	✗	AUROC	.62	.74	.72	.64	.53	.70	.71	.77	.49	.61	.65
		AP	.59	.69	.66	.65	.52	.67	.66	.76	.48	.59	.63
ZED	✗	AUROC	.61	.60	.63	.47	.45	.51	.55	.39	.71	.55	.55
		AP	.57	.60	.72	.46	.47	.58	.56	.42	.79	.57	.57
D3-L2	✗	AUROC	.63	.66	.81	.56	.76	.88	.79	.68	.75	.65	.72
		AP	.64	.71	.82	.64	.80	.90	.82	.68	.75	.69	.74
D3-cos	✗	AUROC	.60	.64	.81	.54	.75	.88	.76	.67	.74	.64	.70
		AP	.63	.69	.82	.62	.79	.90	.79	.68	.74	.69	.74
STALL	✗	AUROC	.78	.83	.72	.79	.82	.88	.82	.85	.79	.72	.80
		AP	.78	.84	.75	.78	.80	.89	.80	.84	.80	.68	.80
AIGVDet	✓	AUROC	.68	.79	.59	.75	.70	.72	.70	.80	.61	.68	.70
		AP	.71	.68	.56	.52	.73	.90	.76	.89	.66	.65	.70
DeCoF	✓	AUROC	.95	.97	1.00	.90	.78	.99	.97	.85	.96	.88	.92
		AP	.91	.92	.99	.76	.49	.98	.95	.78	.56	.74	.82
ReStraV	✓	AUROC	.97	.98	.97	.99	.80	1.00	.77	.98	.97	.82	.92
		AP	.95	.97	.97	.99	.75	1.00	.71	.97	.98	.74	.90
ATSS	✓	AUROC	.93	.99	1.00	.99	1.00	1.00	1.00	.98	.85	.98	.97
		AP	.95	.99	1.00	1.00	1.00	1.00	1.00	.99	.86	.98	.98
CMTA	✓	AUROC	.99	1.00	.98	.98	.99	1.00	1.00	.99	1.00	.99	.99
		AP	.99	1.00	.96	.97	.98	1.00	1.00	.99	1.00	.99	.99
FOVEA–GPT-5.2	✗	AUROC	.94	.95	.97	.77	.82	.92	.88	.83	.92	.87	.89
		AP	.92	.92	.95	.73	.77	.88	.82	.78	.89	.82	.85
FOVEA–Gemini 3.1 Pro	✗	AUROC	1.00	.99	1.00	.98	.96	.99	.98	.97	1.00	.98	.99
		AP	1.00	.99	1.00	.98	.97	.99	.98	.98	1.00	.98	.99
FOVEA–Claude Opus 4.6	✗	AUROC	1.00	1.00	1.00	.99	.98	1.00	1.00	1.00	1.00	.97	**.99**
		AP	1.00	1.00	1.00	.99	.98	1.00	1.00	1.00	1.00	.98	**1.00**

MS/Morph/Moon/Hot/Show/Craft/Wild: ModelScope/MorphStudio/MoonValley/HotShot/Show-1/Crafter/WildScrape.

Table 4: **Watermark-removal control** (FOVEA–GPT-5.2). *Top:* pooled metrics over all test videos. *Bottom:* per-generator fake recall. Pooled accuracy (0.8947 on the uncropped cohort) differs from the generator-macro accuracy of Table 2 (88.27%).

Input (pooled)	Accuracy	Precision	Recall	Fake F1
Original	.8947	.9927	.8967	.9423
Uniform crop	.9103	.9891	.9113	.9486

Recall	ModelScope	MorphStudio	MoonValley	HotShot	Show-1	Gen2	Crafter	LaVie	Sora	WildScrape	**Mean**
Original	.94	.96	1.00	.66	.80	.98	.96	.84	.94	.90	.90
Cropped	.88	.94	1.00	.82	.90	.98	.98	.84	.89	.88	.91

ablation Human Eyes still sees the raw frames. Cropping also removes non-watermark pixels, so we do not attribute the positive change solely to watermark removal. The control tests cropping of the observed cue; masking, blurring, inpainting, and re-encoding are not tested.

A.3 CROSS-BENCHMARK EVALUATION ON GENVIDBENCH

To test transfer beyond GenVideo, we ran full FOVEA with GPT-5.2 on the GenVidBench-143K test set (Ni et al., 2026). Its training and test splits differ in both real-video source and generators; the test set pairs real HD-VG-130M videos with image-to-video (MuseV, SVD) and text-to-video (CogVideo, Mora) generations derived from the same source content. Table 5 reports Top-1 accuracy against the

video backbones benchmarked by Ni et al. (2026): I3D (None, 2017), SlowFast (Feichtenhofer et al., 2019), TPN (Yang et al., 2020), TIN (Shao et al., 2020), X3D (Feichtenhofer, 2020), TRN (Zhou et al., 2018), TSM (Lin et al., 2019), UniFormerV2 (Li et al., 2023), TimeSformer (Bertasius et al., 2021), VideoSwin (Liu et al., 2022), and MViTv2-B (Li et al., 2022).

Table 5: **Top-1 accuracy (%) on the GenVidBench-143K test set.** MuseV, SVD, CogVideo, and Mora are generated subsets; HD-VG is the real subset (HD-VG-130M). The five subsets have equal size, so Top-1 is their mean. Baseline rows are from Table 7 of Ni et al. (2026); FOVEA uses GPT-5.2 and no detector training.

Method	MuseV	SVD	CogVideo	Mora	HD-VG	**Top-1**
I3D	8.15	8.29	60.11	59.24	93.99	45.96
SlowFast	12.25	12.68	38.34	45.93	93.63	40.57
TPN	37.86	8.79	68.25	90.04	97.34	60.46
TIN	33.78	21.47	81.59	79.44	97.88	62.83
X3D	92.39	37.27	65.72	49.60	97.51	68.50
TRN	38.92	26.64	91.34	93.98	93.97	68.97
TSM	70.37	54.70	78.46	70.37	96.76	74.13
UniFormerV2	20.05	14.81	45.21	99.21	96.89	55.23
TimeSformer	73.14	20.17	74.80	39.40	92.32	59.97
VideoSwin	49.11	68.48	88.47	81.42	98.55	77.21
MViTv2-B	76.34	98.29	47.50	96.62	97.58	83.27
FOVEA–GPT-5.2	73.79	84.75	96.25	85.00	88.75	**85.71**

FOVEA obtains the highest mean Top-1 accuracy, 2.44 points above MViTv2-B, without being trained on GenVidBench. We do not report detection accuracy on EvalCrafter (Liu et al., 2024), VideoPhy (Bansal et al., 2025), or VidProM (Wang & Yang, 2024): they evaluate generation quality, physical faithfulness, or prompts rather than real-versus-generated detection, and converting them into detection benchmarks would require a new protocol with a matched real-video set, making the results incomparable to their published tasks.

A.4 CALIBRATION PROTOCOL

FOVEA trains no task-specific model weights. Its only use of labeled data is a one-time calibration of scalar thresholds on 100 labeled videos. These were sampled proportionally from the official GenVideo *training* pool: real videos from Kinetics-400 and Youku-mPLUG, and fake videos from ZeroScope, I2VGen-XL, SVD, AnimateDiff, VideoCrafter, Pika, DynamiCrafter, Stable Diffusion, SEINE, Latte, OpenSora, and CogVideo-5B. GenVideo's generator-disjoint split separates these generators from the ten test generators, and no test video was used.

Labels were used to inspect outcomes near candidate decision boundaries and to fix the following global scalars: Human Eyes evidence/confidence thresholds $\tau_{\mathrm{HE}} = \gamma_{\mathrm{HE}} = 0.8$; analyst audit confidence 0.8; Judge strong-evidence score/confidence $\tau_{\mathrm{ovr}} = 0.85$, $\gamma_{\mathrm{ovr}} = 0.6$; reliability filter $\gamma = 0.6$; analyst-vote and final-score thresholds 0.5; and disagreement penalty $\lambda = 0.15$. Selection was manual and prioritized a high-precision Human Eyes early exit; no aggregate metric, learned objective, or model weight was optimized. On the calibration set, analyst predictions with confidence ≥ 0.8 exceeded 95% correctness. The same constants are frozen for GPT-5.2, Gemini 3.1 Pro, Claude Opus 4.6, and all ablations. The calibration videos are neither inference inputs nor a recurring deployment stage: a new video or generator requires no labeled support set and no update.

A.5 CLOSEST REASONING-BASED VIDEO DETECTORS

Table 6 contrasts FOVEA with the closest reasoning-based detectors. Published numbers provide context only; they are not controlled rankings, because protocols and training data differ.

LAVID and FOVEA both use frozen LVLMs and tools. FOVEA differs in four ways: it decomposes the investigation among specialized analysts, routes analysts and sub-skills by content, tests hypotheses with deterministic spatiotemporal measurements, and fuses the evidence with a reliability-aware Judge. This measurement-linked multi-agent architecture, rather than generic tool use, is the specific

Table 6: **Closest reasoning-based detectors.** Metrics on GenVideo are as reported by the respective papers under their own protocols.

Method	Task-specific adaptation	GenVideo status / relation to FOVEA
LAVID (Liu et al., 2025)	Frozen LVLM, but 25% of each dataset is a labeled reference set for tool selection, and labeled batches select prompt rewrites by F1 (an adaptation the paper runs on the test set).	Closest conceptually, but no GenVideo result and no released code, data, or per-video predictions. Selects transformed views for one LVLM and OR-combines predictions.
TruthLens (Chakraborty et al., 2025)	Training-free.	Image-only VQA/artifact-description pipeline evaluated on face images; no temporal analysis or GenVideo protocol.
VidGuard-R1 (Park et al., 2026)	Qwen2.5-VL-7B SFT, then DPO or GRPO, using 30K CoT and 100K RL samples.	GenVideo mean Recall/F1 = .92/.91 zero-shot after GenVidBench pretraining and .96/.96 after GenVideo training (different protocol).
Skyra (Li et al., 2025)	Full-parameter SFT on ViF-CoT-4K, then GRPO; the GenVideo variant is initialized from the SFT checkpoint and RL-trained for one epoch on 2.2K GenVideo samples.	Acc./Recall/F1 = .7178/.4560/.5900 without GenVideo adaptation and .9100/.8766/.9000 with it.
FOVEA (ours)	No task-specific model-weight training; one-time 100-video scalar calibration (Appendix A.4).	Claude Opus 4.6 Acc./Recall/F1 = .9739/.9590/.9784 under the protocol of Table 1.

novelty. TruthLens lacks temporal analysis, while VidGuard-R1 and Skyra train specialized reasoning models.

A.6 GROUNDED AND EXPLAINABLE IMAGE-LEVEL DETECTORS

Several recent works ground AI-generated *image* detection in localized regions. Table 7 summarizes how they relate to FOVEA.

Table 7: **Grounded image-level detectors versus FOVEA.** None of the cited works reports GenVideo results.

Method	Input / grounding	Task-specific training	Temporal evidence
LEGION (Kang et al., 2025)	Image; pixel masks	Yes	No
FakeXplain (Ji et al., 2026b)	Image; boxes	Yes	No
Locate-Then-Examine (Ji et al., 2026a)	Image; boxes/crops	Yes	No
X-AIGD (Xiao et al., 2026)	Image benchmark; human masks	Yes (proposed detectors)	No
FOVEA (ours)	Video; measurement-linked chain	No weights; scalar calibration	Yes

These methods learn or evaluate spatial grounding for still images. FOVEA instead targets synthesized video and links its decisions to deterministic spatial and inter-frame measurements (optical flow, tracking, phase coherence, FFT, ELA). Its contribution is temporal measurement grounding and reliability-aware multi-agent fusion, not pixel-mask localization.

A.7 RESOURCE REGIME AND BACKBONE DEPENDENCE

The comparison with supervised baselines concerns *target-task adaptation*, not total prior knowledge. FOVEA uses frozen general-purpose VLMs with broad visual priors, whereas supervised baselines fit detector weights under the Pika/Kinetics-400 protocol. The pretraining corpora of proprietary VLMs are undisclosed, so exposure to benchmark videos cannot be ruled out independently. This exposure is, however, held constant within the GPT-5.2 ablation: Pure VLM obtains 42.92% accuracy versus 88.27% for full FOVEA, isolating a 45.35-point gain from the pipeline for that backbone. We do not extrapolate the size of this gain to Claude Opus 4.6, Gemini 3.1 Pro, or open-source VLMs.

Any VLM whose API accepts multiple images and returns structured output can be connected; only the provider-specific message and parser adapter changes, and the same orchestration and evidence schema already run unchanged across GPT, Gemini, and Claude. Our accuracy evidence, however, covers only these three backbones; open-source support is an interface property, not an evaluated accuracy claim.

A.8 Efficiency and Deployment

At the evaluated operating point, FOVEA takes 53.4 s per video on average (52.5 s median), including about 5 s of local CPU work (Table 18). The critical path consists of Human Eyes ($\sim$3.5 s), Planner ($\sim$4.0 s), the parallel analysts ($\sim$28.0 s), the Judge ($\sim$3.5 s), and API/network overhead ($\sim$14.4 s). Videos are independent, so with a concurrency budget C the ideal batch wall time is about N/C times the single-video critical path, subject to provider rate limits; API calls and cost remain linear in N. The evaluated operating point cost USD 0.25 per video. With newer APIs, token pricing, and discounts, we estimate the current deployment cost at about USD 0.01 per video. FOVEA targets evidence-producing forensic screening rather than minimum-latency classification. We did not time the baselines on identical hardware and inputs, and make no efficiency-superiority claim.

A.9 Interpreting Component Ablations

The large architectural effects are consistent across protocols: removing CV grounding or sub-skill routing reduces standard accuracy from 88.27% to 75.25%/75.71%, and Pure VLM falls to 42.92%. The Human Eyes and Physics effects are smaller and cohort-dependent. Removing them changes standard accuracy by $+1.41/+0.66$ points but duration-controlled accuracy by $-1.55/-0.27$ points. Both modules therefore help on the $\geq$4 s cohort but are not uniformly beneficial on every distribution, and the aggregate results do not support a general noise mechanism. Dependence on the Planner is bounded: removing Planner routing changes standard/duration-controlled accuracy from 88.27/91.32 to 87.51/89.59 ($-0.76/-1.73$ points). Deterministic measurements and multi-source Judge fusion therefore remain the central components.

A.10 Short-Clip Robustness Results

Table 8 gives the per-generator results of the $\geq$4 s evaluation summarized in Exp-2. The four-generator averages are not comparable to the ten-generator averages of Table 1.

A.11 Scope of the Evidence and Duration Claims

Evidence chains. FOVEA stores evidence at two levels. The execution log records the full decision process: sub-skill routing, analyst scores and confidences, Judge validation, and the decision rule responsible for the final verdict. A structured database preserves the evidence payload produced by the sub-skills. For example, one `face_collapse_moderate` record identifies frames 2–5, reports an anomaly score of 0.60 against a 0.50 threshold, and localizes anomalies to face bounding boxes in frames 3 and 4. The log also records, for instance, a strong override triggered by the Physics analyst immediately before the corresponding verdict, and it shows algorithmic cues being rejected or down-weighted when they are not visually corroborated. The claim these records support is *traceability*: every verdict can be traced to logged measurements and the rule that used them. We have not independently validated that the localized cues are correct, specific to AI generation, or causally responsible for the verdict; such validation would require ground-truth artifact annotations and is left to future work.

Duration. The $\geq$4 s evaluation is a short-clip robustness test, not a full duration deconfounding. After filtering, 2,048 of the 2,391 retained fake videos come from ModelScope and Gen-2, whose typical duration is 4 s, whereas the 9,200 retained real videos (MSR-VTT) still span 10–30 s. The experiment shows that the advantage persists once clips shorter than 4 s are excluded, but it does not remove all duration-related information. Duration-matched trimming of real videos and duration-, resolution-, metadata-, and source-only baselines were not run. Duration metadata is used only for frame sampling and logging, never as a decision feature.

Table 8: **Duration-controlled evaluation ($\geq$4 s, %).** Methods marked † are **supervised**; FOVEA trains **no detector weights**. Test videos <4 s are excluded; four generators retain sufficient samples. Best average results are highlighted in **bold**.

Method	Metric	ModelScope($\geq$ 4 s)	Gen-2($\geq$ 4 s)	Sora($\geq$ 4 s)	WildScrape($\geq$ 4 s)	Avg.
DeMamba†	Recall	86.10	93.36	30.95	29.49	59.97
	Accuracy	90.95	94.58	63.38	62.64	77.89
	F1	90.49	94.51	45.80	44.12	68.73
NPR†	Recall	60.42	76.26	19.80	24.69	45.29
	Accuracy	78.46	86.38	58.15	60.59	70.89
	F1	73.72	84.85	32.12	38.52	57.30
TALL†	Recall	51.27	86.28	5.83	30.53	43.48
	Accuracy	74.23	91.74	51.52	63.87	70.34
	F1	66.55	91.26	10.73	45.80	53.58
STIL†	Recall	73.12	32.21	0.27	0.33	26.48
	Accuracy	84.96	64.50	48.53	48.56	61.64
	F1	82.94	47.57	0.52	0.64	32.92
NSG-VD†	Recall	67.82	87.73	78.72	84.08	79.59
	Accuracy	80.67	91.57	94.08	92.71	89.76
	F1	78.72	91.47	71.15	85.44	81.69
FOVEA (Ours)	Recall	100.00	100.00	100.00	85.00	**96.25**
	Accuracy	100.00	100.00	100.00	92.50	**98.13**
	F1	100.00	100.00	100.00	91.89	**97.97**

Responsible use. The verdict and evidence chain of FOVEA are decision-support artifacts, not autonomous grounds for high-stakes action. False positives should not trigger censorship, sanctions, or authenticity claims without human verification, and fluent rationales are not proof. Deployments that send frames to external APIs should apply privacy, consent, retention, and access controls, and should use controlled endpoints for sensitive data.

B PER-GENERATOR COMPARISON

Table 9: **Paradigm comparison.** FOVEA fits no detector weights (only 100 labeled videos for one-time scalar-threshold calibration), requires no retraining, and provides interpretable evidence.

Property	STIL	NPR	TALL	DeMamba	NSG-VD	FOVEA
Labeled data (train/calib.)	10k	10k	10k	10k	10k	**100***
Trainable Params (M)	21.63	1.37	82.89	119.89	0.25	**0**
GPU Required	✓	✓	✓	✓	✓	✗
New Gen. Adapt.	Retrain	Retrain	Retrain	Retrain	Retrain	**None**
Interpretable	✗	✗	✗	✗	✗	✓
Avg. Accuracy (%)	63.51	77.96	79.85	84.21	91.46	**97.39**

*Calibration only, from generators disjoint from the test set (Appendix A.4).

Table 10 provides the full per-generator comparison between FOVEA and NSG-VD. FOVEA outperforms NSG-VD in accuracy on all 10 generators and in recall on 7 of 10, achieving perfect scores (100%/100%) on ModelScope, MorphStudio, MoonValley, Gen-2, and Sora. On the two most challenging generators, FOVEA still leads in accuracy: WildScrape (92.50% vs. 88.75%) and HotShot (94.29% vs. 91.67%). The only cases where NSG-VD retains a slight edge are recall on HotShot (92.50% vs. 92.10%) and Crafter (98.33% vs. 96.13%), both within 2.5%. FOVEA also achieves near-perfect real-video accuracy (precision $\approx$99.9%), compared to NSG-VD's 95.65%.

Table 10: **Per-generator comparison between FOVEA and NSG-VD.** Accuracy and fake recall (%) on each generator. NSG-VD is trained on 10k pairs; FOVEA uses zero training data. Green cells indicate where FOVEA outperforms NSG-VD.

	ModelScope		MorphStudio		MoonValley		HotShot		Show-1	
Method	Acc.	Rec.	Acc.	Rec.	Acc.	Rec.	Acc.	Rec.	Acc.	Rec.
NSG-VD†	81.67	68.33	98.33	98.33	96.67	100.00	91.67	92.50	90.83	87.50
FOVEA (Ours)	100.00	100.00	100.00	100.00	100.00	100.00	94.29	92.10	92.86	89.97
	Gen-2		**Crafter**		**Lavie**		**Sora**		**WildScrape**	
Method	Acc.	Rec.	Acc.	Rec.	Acc.	Rec.	Acc.	Rec.	Acc.	Rec.
NSG-VD†	88.33	80.00	95.83	98.33	94.17	94.17	88.39	78.57	88.75	82.50
FOVEA (Ours)	100.00	100.00	97.14	96.13	97.14	96.30	100.00	100.00	92.50	84.95

C ABLATION STUDY RESULTS

Tables 11 and 12 present the full per-generator breakdown for each ablation configuration discussed in Section 5. All experiments use GPT-5.2 as the VLM backbone (temperature=0, fixed random seeds).

D FULL VLM BACKBONE RESULTS

Tables 14 and 15 present the complete per-generator breakdown.

E INFERENCE COST ANALYSIS

FOVEA has a higher per-video wall-clock time than supervised baselines, averaging 53.4 s per video (median 52.5 s). However, local CV sub-skill execution and orchestration consume under 5 s; the remaining time is network round-trips through a third-party API proxy. In a direct-API benchmark from a low-latency region, median end-to-end time drops to approximately 12–15 s per video (local compute $\sim$5 s plus sequential VLM calls at 1–3 s each). Three factors further reduce the practical gap: (i) FOVEA trains no detector weights; (ii) no retraining is needed when new generators appear; and (iii) VLM per-token prices and latencies continue to drop year over year. For latency-critical deployments, FOVEA can serve as a high-accuracy second-stage verifier. See Table 18 for a latency breakdown.

Monetary cost. At current Claude Opus 4.6 pricing, the average per-video API cost is \$0.25 (range: \$0.15–0.40 depending on video length and number of analysts activated). Evaluating the full GenVideo test set (18,588 videos) costs approximately \$4,647, substantial but comparable to GPU-hour costs of training supervised baselines on large datasets. Importantly, this is a one-time inference cost with no training investment; supervised baselines require repeated GPU training (typically 8–24 GPU-hours per configuration) whenever new generators emerge. VLM pricing follows a deflationary trend (input token costs dropped $\sim$10$\times$ between 2023–2025); with newer APIs, token pricing, and discounts, we estimate the current deployment cost at about \$0.01 per video (Appendix A.8).

F ERROR ANALYSIS

The main text summarizes key error patterns. Here we present the full breakdown for completeness.

False negatives (missed fakes). FOVEA misses some AI-generated videos from two generators: HotShot (recall: 92.10%) and WildScrape (recall: 84.95%). HotShot's 1-second clips contain so few frames that temporal analysis is largely ineffective, and the low resolution (672$\times$384) limits the visibility of spatial artifacts. WildScrape is a mixed-source collection where some videos are produced by high-quality proprietary generators that leave minimal detectable artifacts at the resolution provided

Table 11: **Ablation study — full per-generator results (standard evaluation, %).** All configurations are **training-free** FOVEA variants using GPT-5.2 as the VLM backbone.

Configuration	Metric	Model Scope	Morph Studio	Moon Valley	HotShot	Show1	Gen2	Crafter	Lavie	Sora	Wild Scrape	Avg.
FOVEA (full)	Recall	93.95	96.32	100.00	66.15	79.91	98.11	95.91	83.91	94.29	90.05	**89.83**
	Accuracy	91.43	92.86	95.71	71.43	81.43	94.29	92.86	84.29	90.91	87.50	**88.27**
	F1	93.62	95.05	97.09	76.74	86.02	96.08	95.05	88.42	92.96	87.80	**90.92**
w/o Human Eyes	Recall	93.66	95.89	100.00	75.80	86.06	97.82	97.72	78.29	89.95	75.01	**89.10**
	Accuracy	92.86	94.29	97.14	79.72	87.14	95.71	95.71	81.43	89.89	82.50	**89.68**
	F1	94.95	96.02	98.04	84.44	90.53	97.03	97.03	85.71	89.77	81.08	**91.48**
w/o Planner routing	Recall	98.08	93.88	100.00	65.94	73.88	98.37	98.00	83.79	88.57	85.16	**88.56**
	Accuracy	94.29	91.43	95.71	71.43	77.14	94.29	94.29	84.29	87.27	84.76	**87.51**
	F1	96.08	94.04	97.09	76.74	82.22	96.08	96.08	88.42	89.86	84.61	**90.16**
w/o Sub-skill routing	Recall	89.73	92.04	100.00	52.15	48.03	94.98	85.94	53.70	84.86	69.91	**77.20**
	Accuracy	82.50	87.14	92.86	58.57	55.71	85.21	82.86	60.07	79.65	72.50	**75.71**
	F1	83.72	91.09	95.24	64.20	60.76	86.36	87.76	65.85	80.95	71.79	**78.77**
w/o CV sub-skills	Recall	95.06	93.92	100.00	57.86	70.12	80.21	88.19	63.83	89.94	80.07	**81.90**
	Accuracy	80.20	85.71	89.90	59.96	68.57	72.50	81.43	64.29	77.50	72.50	**75.25**
	F1	82.61	90.38	93.46	67.44	76.09	74.42	87.13	71.91	79.78	74.42	**79.79**
w/o Style Analyst	Recall	91.76	92.16	100.00	58.27	69.99	78.20	76.07	57.87	88.57	55.00	**76.76**
	Accuracy	92.86	92.86	98.57	68.57	77.14	82.86	81.43	68.57	90.91	75.31	**82.88**
	F1	94.85	94.85	99.01	72.50	81.40	86.67	85.39	72.50	92.54	68.75	**84.84**
w/o Physics Analyst	Recall	95.99	94.31	100.00	73.50	86.16	100.00	98.02	79.94	92.50	85.02	**90.55**
	Accuracy	92.86	91.43	95.71	77.14	85.71	95.71	94.29	81.43	89.60	84.96	**88.93**
	F1	95.05	94.07	97.09	82.22	89.58	97.09	96.08	86.02	92.50	85.30	**91.46**
w/o Spatial Analyst	Recall	94.90	97.84	100.00	61.90	86.18	95.07	91.89	86.10	85.02	80.19	**87.90**
	Accuracy	87.50	92.86	94.29	67.14	84.29	87.50	88.57	84.29	82.50	79.86	**84.89**
	F1	88.37	95.15	96.15	72.94	88.66	88.37	91.93	88.66	82.93	79.92	**87.32**
w/o Temporal Analyst	Recall	94.71	98.06	100.00	56.05	64.00	100.00	83.95	63.72	89.92	69.93	**82.10**
	Accuracy	94.84	97.14	98.57	67.14	72.86	97.50	87.14	72.86	92.50	82.50	**86.32**
	F1	94.97	98.08	99.01	70.89	77.11	97.56	90.32	77.11	92.31	80.38	**87.73**
w/o Watermark Analyst	Recall	100.00	96.03	100.00	60.05	79.99	94.62	100.00	81.99	75.01	95.49	**88.30**
	Accuracy	87.50	89.96	92.86	64.29	78.57	85.06	92.86	79.99	74.77	85.23	**83.11**
	F1	88.89	93.20	95.24	70.59	84.21	86.36	95.24	85.42	75.15	86.36	**86.05**
Pure VLM (no framework)	Recall	50.96	46.23	56.27	18.43	35.42	52.44	50.78	38.98	52.68	40.54	**44.27**
	Accuracy	48.29	47.34	51.63	24.39	33.20	44.90	44.39	42.58	48.71	43.78	**42.92**
	F1	48.58	48.66	52.08	26.77	41.73	49.15	49.65	43.03	49.08	39.34	**44.81**

in the benchmark. In both cases, the errors are concentrated on clips where the VLM finds insufficient visual evidence to differentiate from real content.

False positives (real classified as fake). With Claude Opus 4.6, FOVEA achieves near-perfect real-video accuracy (precision ≈99.9%, only a handful of false positives across the 10 000-video MSR-VTT real test pool) on the GenVideo test set. However, with weaker backbones, false positives occur: Gemini 3.1 Pro produces 17.5% false positives, typically on real videos with unusual visual characteristics (e.g., low-quality webcam footage, heavily compressed content, or artistic/stylized camera work) that the VLM misinterprets as AI artifacts.

Generator-specific observations.

- **ModelScope, MorphStudio, MoonValley, Gen-2, Sora**: Perfect detection (100% accuracy) with Claude Opus 4.6. These generators all leave identifiable artifacts (frequency-domain signatures, spatial inconsistencies, or temporal flickering) that the neuro-symbolic pipeline reliably detects.
- **Crafter, Lavie**: Near-perfect detection (97.14% accuracy). The few missed cases involve ultra-short clips where frame counts are insufficient for reliable temporal analysis.
- **WildScrape**: The most challenging case (92.50% accuracy), reflecting the heterogeneous nature of this mixed-source collection.

Table 12: **Ablation study — full per-generator results (duration-controlled, ≥4 s, %).** All configurations are **training-free** FOVEA variants using GPT-5.2.

Configuration	Metric	ModelScope(≥ 4 s)	Gen-2(≥ 4 s)	Sora(≥ 4 s)	WildScrape(≥ 4 s)	Avg.
FOVEA (full)	Recall	93.95	98.28	93.96	89.29	**93.87**
	Accuracy	91.43	95.10	91.87	86.88	**91.32**
	F1	93.62	95.98	93.85	87.55	**92.75**
w/o Human Eyes	Recall	93.66	97.53	90.91	74.24	**89.09**
	Accuracy	92.86	95.36	89.27	81.57	**89.77**
	F1	94.95	97.64	89.06	80.74	**90.60**
w/o Planner routing	Recall	98.08	98.90	89.50	85.80	**93.07**
	Accuracy	94.29	93.47	86.33	84.26	**89.59**
	F1	96.08	95.87	89.13	83.78	**91.22**
w/o Sub-skill routing	Recall	89.73	94.54	84.36	70.08	**84.68**
	Accuracy	82.50	85.57	79.76	71.90	**79.93**
	F1	83.72	86.04	81.80	71.51	**80.77**
w/o CV sub-skills	Recall	95.06	79.84	90.47	80.26	**86.41**
	Accuracy	80.20	72.98	78.06	71.91	**75.79**
	F1	82.61	74.66	80.35	73.96	**77.89**
w/o Style Analyst	Recall	91.76	78.74	88.80	55.68	**78.75**
	Accuracy	92.86	82.82	90.66	76.10	**85.61**
	F1	94.85	86.02	93.50	69.00	**85.84**
w/o Physics Analyst	Recall	95.99	100.00	92.27	84.17	**93.11**
	Accuracy	92.86	96.18	89.25	85.91	**91.05**
	F1	95.05	97.27	92.62	84.75	**92.42**
w/o Spatial Analyst	Recall	94.90	95.85	84.70	79.23	**88.67**
	Accuracy	87.50	86.70	81.97	79.80	**83.99**
	F1	88.37	87.80	82.54	79.95	**84.67**
w/o Temporal Analyst	Recall	94.71	100.00	89.96	69.34	**88.50**
	Accuracy	94.84	97.53	92.56	82.58	**91.88**
	F1	94.97	97.85	92.45	80.57	**91.46**
w/o Watermark Analyst	Recall	100.00	94.19	74.05	96.14	**91.09**
	Accuracy	87.50	85.70	73.80	84.78	**82.94**
	F1	88.89	86.20	74.87	85.87	**83.96**
Pure VLM (no framework)	Recall	39.99	45.84	41.06	36.91	**40.95**
	Accuracy	35.50	44.17	42.52	35.05	**39.31**
	F1	39.28	44.03	39.09	33.93	**39.08**

Table 13: **VLM backbone comparison (%).** All backbones use the **training-free** FOVEA framework. Average performance on standard and duration-controlled (≥4 s) evaluations. Best results in **bold**.

	Standard (10 gen.)			**Duration-Controlled (≥4 s)**		
VLM Backbone	Recall	Acc.	F1	Recall	Acc.	F1
GPT-5.2	89.83	88.27	90.92	93.87	91.32	92.75
Gemini 3.1 Pro	99.40	86.57	89.81	100.00	82.28	85.63
Claude Opus 4.6	**95.90**	**97.39**	**97.84**	**96.25**	**98.13**	**97.97**

- **Show-1, HotShot**: The second and third most challenging generators (92.86% and 94.29% accuracy, respectively). HotShot's difficulty stems from its extreme brevity (1 s) and low resolution; Show-1 similarly produces short clips with limited temporal evidence.

Table 14: **VLM backbone comparison — full per-generator results (standard evaluation, %).** Per-generator Accuracy is computed on a balanced eval pool that pairs each generator's fake set with an equally sized real subsample drawn from MSR-VTT, so Gemini's 17.5% aggregate false-positive rate translates into a per-generator Accuracy below the (Recall, Specificity) midpoint suggested by the aggregate figure.

VLM Backbone	Metric	Model Scope	Morph Studio	Moon Valley	HotShot	Show1	Gen2	Crafter	Lavie	Sora	Wild Scrape	Avg.
GPT-5.2	Recall	93.95	96.32	100.00	66.15	79.91	98.11	95.91	83.91	94.29	90.05	**89.83**
	Accuracy	91.43	92.86	95.71	71.43	81.43	94.29	92.86	84.29	90.91	87.50	**88.27**
	F1	93.62	95.05	97.09	76.74	86.02	96.08	95.05	88.42	92.96	87.80	**90.92**
Gemini 3.1 Pro	Recall	100.00	100.00	100.00	100.00	96.16	100.00	100.00	97.82	100.00	100.00	**99.40**
	Accuracy	83.33	90.28	89.72	90.12	87.14	82.22	90.44	88.57	82.78	81.67	**86.57**
	F1	85.71	93.64	93.14	93.54	91.43	84.90	93.52	92.45	85.31	84.52	**89.81**
Claude Opus 4.6	Recall	100.00	100.00	100.00	92.10	89.97	100.00	96.13	96.30	100.00	84.95	**95.90**
	Accuracy	100.00	100.00	100.00	94.29	92.86	100.00	97.14	97.14	100.00	92.50	**97.39**
	F1	100.00	100.00	100.00	95.83	94.74	100.00	97.96	97.96	100.00	91.89	**97.84**

Table 15: **VLM backbone comparison — full per-generator results (duration-controlled, $\geq$4 s, %).** Same per-generator balanced-pool accounting as Table 14.

VLM Backbone	Metric	ModelScope	Gen-2	Sora	WildScrape	Avg.
GPT-5.2	Recall	93.95	98.28	93.96	89.29	**93.87**
	Accuracy	91.43	95.10	91.87	86.88	**91.32**
	F1	93.62	95.98	93.85	87.55	**92.75**
Gemini 3.1 Pro	Recall	100.00	100.00	100.00	100.00	**100.00**
	Accuracy	83.33	81.77	82.10	81.94	**82.28**
	F1	85.71	85.38	86.11	85.32	**85.63**
Claude Opus 4.6	Recall	100.00	100.00	100.00	85.00	**96.25**
	Accuracy	100.00	100.00	100.00	92.50	**98.13**
	F1	100.00	100.00	100.00	91.89	**97.97**

G Additional Experimental Results

G.1 Qualitative Case Studies

We present three Sora-generated video cases analyzed by FOVEA, illustrating different detection difficulty levels. For each case, we show sampled frames and the per-agent evidence summary.

Case 1: Easy detection (sora_28). As shown in Figure 3, the Human Eyes agent immediately detects a clearly visible multicolor OpenAI/Sora swirl watermark at the bottom-right corner (confidence 0.88), triggering an early-exit classification as **Fake** without activating further analysis.



Figure 3: **Case 1 (sora_28): Easy detection via watermark shortcut.** A visible OpenAI/Sora watermark in the bottom-right corner is caught by the Human Eyes agent (conf. 0.88), bypassing all downstream analysis.

Case 2: Multi-agent collaboration (sora_16). A close-up portrait of an older man by a window with no obvious artifacts at first glance (Figure 4). The Human Eyes agent does not trigger early-exit. Four analyst agents contribute complementary evidence (Table 16):

Table 16: **Agent evidence summary for Case 2 (sora_16).**

Agent	Score	Conf.	Key Evidence
Style	0.88	0.72	Airbrushed/overly smooth skin, plastic beard microtexture, poreless look; FFT shows ≈0 high-frequency energy
Watermark	0.85	0.80	Faint vertical tick marks in bottom-right; later frames show OpenAI swirl-like logo
Temporal	0.84	0.74	Per-frame micro-jitter and texture redraw; keypoint jerk ratio 98.4%
Spatial	0.62	0.55	ELA boundary halos (halo_score=2.02; boundary_bg_ratio up to 3.02)



Figure 4: **Case 2 (sora_16): Multi-agent collaboration.** No obvious artifacts visible to the Human Eyes agent. The Style analyst detects AI-characteristic oversmoothing (score 0.88), triggering a **Strong Override** → **Fake** (conf. 0.72).

Verdict: Strong Override triggered by Style agent → **Fake** (score 0.88, confidence 0.72).

Case 3: Challenging detection (sora_45). Figure 5 shows a wall of vintage CRT televisions displaying various images, a visually complex scene where AI artifacts are subtle. The Human Eyes agent passes (confidence 0.62, below threshold). Agent analysis (Table 17) shows high disagreement:

Table 17: **Agent evidence summary for Case 3 (sora_45).**

Agent	Score	Conf.	Key Evidence
Style	0.82	0.67	TV screens show smeared/placeholder faces; FFT indicates extremely low HF energy
Spatial	0.74	0.62	221 faces detected on TV screens with avg. anomaly_score ≈0.60; elevated boundary_bg_ratio (2.1–2.2)
Watermark	0.30	0.68	Small magenta/pink spiral logo in bottom-right, classified as producer logo, not AI watermark
Temporal	0.40	0.62	Physical TV stack stays consistent; screen changes are expected hard-cuts, not AI artifacts

Verdict: Dynamic Weighted Fusion → **Fake** ($\bar{s}_w$ = 0.564, σ_w = 0.222, s_{final} = 0.531, confidence 0.504). This is a near-threshold detection, illustrating the system's limitation on visually complex scenes where inherent noise masks AI artifacts.

G.2 Inference Latency Breakdown

Table 18 provides a detailed breakdown of FOVEA's per-video inference time.

The dominant cost is network round-trip latency to commercial VLM APIs. In a direct-API benchmark (without proxy routing), median end-to-end time drops to approximately 12–15 s per video, with per-call VLM latency typically in the 1–3 s range and local computation totaling ∼5 s. This demonstrates that the architecture's computational workload is modest; the practical bottleneck in our default deployment (53.4 s) is API-proxy queueing overhead rather than inherent algorithmic cost.



Figure 5: **Case 3 (sora_45): Challenging detection.** A visually complex scene with inter-agent disagreement (σ_w=0.222). The system barely classifies this as Fake via dynamic weighted fusion ($\bar{s}_w$ = 0.564, s_{final} = 0.531).

Table 18: **Inference latency breakdown per video.** Measured over the full test set with Claude Opus 4.6. The first block reports *local CPU* compute; the second block reports the *wall-clock* cost of the same VLM-issued steps including API round-trip and proxy queueing. The two blocks are alternative views of the same pipeline rather than additive components.

Component	**Time (s)**
Local CPU compute (no network)	
Evidence extraction (FFmpeg + frame caching)	∼1.0
CV sub-skill execution (all analysts, parallel)	∼2.0
LangGraph orchestration + JSON parsing	∼1.0
Sub-skill router pre-pass ($\mathbf{V}_{\text{prev}}$)	∼1.0
Local computation total	∼5.0
Wall-clock incl. API round-trip and proxy queueing	
Human Eyes agent (VLM call)	∼3.5
Planner agent (VLM call)	∼4.0
Analyst VLM calls (3–5 in parallel, slowest path)	∼28
Judge agent (VLM call)	∼3.5
API proxy/network overhead (across all calls above)	∼14.4
End-to-end average	53.4
End-to-end median	52.5

G.3 Analyst Activation Frequency

Table 19 reports the activation frequency of each analyst agent, determined by the Planner's dynamic routing on the full test set. The Planner activates the majority of analysts for most videos, with Style (95–97%), Spatial (92–95%), and Temporal (98–99%) near-universally selected across most configurations. Physics is activated on 62–74% of videos, skipped primarily on static or near-static content where motion analysis is uninformative. Watermark shows the most variation: activated on 9–30% of videos, with higher rates in configurations without Human Eyes (since many watermark-heavy videos are caught at early-exit and never reach the Planner). This selective activation pattern confirms that the Planner makes content-dependent routing decisions rather than blindly activating all analysts.

Table 19: **Analyst activation frequency (%) by the Planner.** Measured on a stratified audit subset of the GenVideo test set (475 videos sampled proportionally across the ten generators) with full per-call routing logs retained. "Planner Cases" is the number of audit videos that reached the Planner after Human Eyes early-exit; the residual gap to 475 corresponds to videos resolved at the Human Eyes stage (or, for the *w/o Human Eyes* configuration, additionally filtered by metadata-only short-circuits).

Backbone	**Planner Cases**	**Style**	**Physics**	**Spatial**	**Temporal**	**Watermark**
GPT-5.2 (full)	256	95%	62%	95%	99%	9%
GPT-5.2 (w/o Human Eyes)	261	97%	64%	92%	98%	30%
Claude Opus 4.6 (full)	230	100%	74%	93%	100%	6%

H IMPLEMENTATION DETAILS

FOVEA is implemented in Python using LangGraph (LangChain, Inc., 2024) as the multi-agent orchestration framework, which provides stateful graph execution, parallel analyst execution, and deterministic workflow composition.

Table 20: **Sub-skill inventory across analyst agents.** Each sub-skill is conditionally activated by the vision sub-skill router based on video content.

Analyst	Sub-Skill	Description
Style	FFT Spectral Analysis	Frequency-domain signature detection
Physics	Optical Flow (Farneback)	Motion coherence & background coupling
	Geometry Stability	Straight-line distortion measurement
	NSG-Lite	Lightweight physical consistency check
Spatial	Error Level Analysis (ELA)	JPEG recompression boundary artifacts
	Patch Anomaly Detection	Local texture inconsistency
	Boundary Anomaly Analysis	Edge melting and halo detection
	Edge Coherence Analysis	Unnatural soft boundary detection
	Multi-Face Collapse Detection	Per-face structural distortion check
	Blur Uniformity Analysis	Selective face degradation detection
Temporal	Local Phase Coherence (LPC)	Facial edge stability across frames
	Feature Point Tracking	Trajectory smoothness & survival rate
Watermark	(VLM-only)	Semantic text/logo recognition

VLM Backbone. All agent reasoning is powered by Claude Opus 4.6 (Anthropic, 2026) (ablation studies use GPT-5.2 (OpenAI, 2025)). Each VLM call uses structured output generation (JSON mode with Pydantic schema enforcement). Frames are passed as base64-encoded images with a configurable maximum of 40 images per call, sampled uniformly to preserve temporal coverage of the video's beginning, middle, and end.

CV Toolkit. Classical CV sub-skills are implemented using OpenCV and standard signal processing libraries, including: dense optical flow (Farneback), FFT spectral analysis, Error Level Analysis via JPEG recompression, Laplacian-based blur detection, face detection (Haar cascades and DNN-based), feature point tracking (Shi-Tomasi corners with Lucas-Kanade flow), and local phase coherence computation. All sub-skills operate on cached frames and require no GPU.

Specialized Sub-Skill Definitions. We briefly define the less-standard sub-skills. *Geometry Stability* measures straight-line distortion by detecting Hough lines across frames and computing the standard deviation of their angular displacement; a score $>2^\circ$ flags warping. *NSG-Lite* is a simplified CPU-only physical-consistency heuristic inspired by NSG-VD's normalized spatiotemporal gradient: it computes the mean-squared difference between adjacent-frame Farneback flow magnitudes and thresholds against a calibrated value; unlike the full NSG-VD (which relies on score functions of a pre-trained diffusion model), NSG-Lite operates entirely on optical-flow statistics. *Patch Anomaly Detection* divides each frame into 32×32 patches, computes local DCT energy variance, and flags patches whose variance deviates $>2\sigma$ from the frame mean. *Edge Coherence Analysis* applies Canny edge detection to consecutive frames and computes the Jaccard similarity of edge maps; AI-generated videos often show lower inter-frame edge consistency (<0.3). *Multi-Face Collapse Detection* runs DNN-based face detection on all frames, crops each face, and evaluates structural distortion via landmark symmetry ratios and aspect-ratio consistency; faces with symmetry deviation $>15\%$ are flagged.

CV-to-VLM Data Interface. To prevent VLM hallucination over high-dimensional CV outputs, every sub-skill emits two parallel artifacts via a typed JSON evidence contract: (i) *scalar summary fields* that the VLM consumes as plain text (e.g., `flow_mean_magnitude`, `flow_temporal_variance`, `fft_high_freq_energy_ratio`, `ela_boundary_score`, `face_anomaly_score`,

phase_coherence_drop), and (ii) *rendered visual aids* as base64-encoded PNG thumbnails (e.g., dense optical-flow color wheels, FFT log-magnitude spectra, ELA difference maps, per-face crop grids) attached as additional images in the same VLM call. The VLM is prompted to first read the scalar fields, then optionally cross-check with the visual aid; dense fields (e.g., the full $H \times W$ flow tensor) are never passed verbatim. Each analyst's evidence record stores both the scalars and a hash of the rendered aids for reproducibility.

Storage and Caching. Evidence extraction results are cached in a local filesystem organized by video ID. Analysis results are persisted to a SQLite database for experiment tracking and reproducibility. The system supports experiment isolation through named configurations, enabling parallel evaluation across different settings.

Frame Sampling Strategy. The sampling rate adapts to video duration: 4 fps for videos $\leq$60 s and 2 fps for longer content. For each VLM call, frames are selected using a uniform stride strategy that preserves temporal coverage of the video's beginning, middle, and end. The Human Eyes agent inspects only three key frames (first, middle, last) for rapid screening, while analyst agents receive the full cached frame set. This adaptive strategy balances comprehensiveness with API cost constraints.

Compute Resources. All experiments are run on a standard workstation with an Intel Xeon CPU (no GPU required for FOVEA). VLM inference is performed via commercial API calls. The only computational resources needed locally are CPU cycles for OpenCV-based CV sub-skills and network bandwidth for API communication. In our deployment, local CV sub-skill execution and LangGraph orchestration consume under 5 s in total per video; the remaining inference time (average 53.4 s) is dominated by network round-trips and API-proxy queueing overhead.

Reproducibility and API Versioning. We use pinned API model versions (e.g., claude-opus-4-6, gpt-5.2, gemini-3-1-pro) with temperature=0 and fixed random seeds for frame sampling. While exact replication requires access to the same commercial APIs, the multi-backbone evaluation (Table 13) demonstrates consistent results across three independent VLM families (Claude, GPT, Gemini), confirming robustness to model choice. Should any specific API version be deprecated, the architecture transfers directly to successor models.

I Detailed Agent Descriptions

This section provides detailed descriptions of each agent in the FOVEA pipeline, complementing the overview in Section 4.2.

Human Eyes Agent. The Human Eyes agent serves as a rapid early-exit gate, inspecting three key frames (first, middle, last) as a deliberate speed-vs-coverage tradeoff. It evaluates four dimensions aligned with the downstream analysts: (i) *spatial artifacts* (severe anatomical collapse, melted facial features, grossly distorted hands), (ii) *watermark presence* (recognizable generator logos or platform branding), (iii) *physics anomalies* (objects floating or bleeding into backgrounds), and (iv) *style incoherence* (plastic-like textures, broken perspective, unnatural uniformity). Temporal analysis is excluded at this stage because three frames are insufficient for reliable motion assessment. For each dimension, the agent outputs a confidence score $c_{\text{HE}}^{(d)} \in [0, 1]$ with a brief textual finding. If any dimension exceeds 0.8, the video is immediately classified as fake (early-exit). The threshold is deliberately conservative, ensuring *high precision, low recall*. When no dimension triggers early-exit, the agent forwards a structured observation record to the Planner containing all per-dimension scores and their associated findings, even sub-threshold ones. In our experiments, 37.1% of fake videos are caught at this stage, and the remaining observation records enable the Planner to make informed routing decisions without redundant VLM inference on the three key frames.

Planner Agent. The Planner receives the full set of sampled frames together with the Human Eyes observation record, and determines which of the five analyst agents to activate for thorough, full-video forensic analysis. Rather than relying on a fixed numerical formula, the Planner VLM directly reasons over the combined evidence: it weighs sub-threshold signals from Human Eyes (e.g., style $= 0.75$, below the 0.8 early-exit threshold) against its own inspection of the complete frame set.

This content-adaptive reasoning allows the Planner to activate the most relevant analysts for each video; for instance, it activates the Style analyst when Human Eyes reports elevated style concerns, or deactivates Watermark when no text or logos appear anywhere in the video. Each activation decision is accompanied by a free-text rationale. Activated analysts receive the full cached frame set, enabling substantially more thorough analysis than the three-frame Human Eyes screening. This design transforms Human Eyes from a binary gate into an active information source, eliminating redundant VLM reasoning between the screening and planning stages.

Style Analyst. Distinguishes AI-generated synthetic content from authentic camera footage and traditional CG/animation. Employs *FFT (Fast Fourier Transform) analysis* to detect frequency-domain signatures characteristic of neural rendering: AI-generated frames often exhibit spectral peaks or attenuated high-frequency components absent in both real footage and professional CG. The VLM additionally assesses surface texture quality, material uniformity, and rendering consistency.

Physics Analyst. Evaluates physical plausibility across frames. Integrates three CV sub-skills: (i) *optical flow analysis* to detect motion coherence violations, background coupling artifacts, and non-inertial motion anomalies; (ii) *geometry stability checking* to measure straight-line distortion in backgrounds, which indicates perspective warping from neural generation; and (iii) *NSG-lite*, a lightweight physical consistency analyzer. The VLM interprets these signals to distinguish genuine physics violations (e.g., objects floating, impossible trajectories) from artistic choices in CG/animation.

Spatial Analyst. Detects frame-level spatial anomalies using six CV sub-skills: *ELA (Error Level Analysis)* for boundary artifacts, *patch anomaly detection* for local texture inconsistencies, *boundary anomaly analysis* for edge melting and halos, *edge coherence analysis* for unnatural soft boundaries, *multi-face collapse detection* that examines every detected face for structural distortion, and *blur uniformity analysis* for selective face degradation. The agent applies a "one drop of blood" principle: any face exhibiting AI-characteristic collapse overrides other signals.

Temporal Analyst. Analyzes consistency of subjects across consecutive frames, targeting artifacts invisible in single-frame analysis. Employs *Local Phase Coherence (LPC)* analysis to measure facial edge stability between frames, a signal that detects subtle geometry warping characteristic of diffusion-based generation. Additionally, *feature point tracking* measures trajectory smoothness, survival rate, and jerk ratios to distinguish normal video from AI-interpolated or frame-by-frame generated content. The agent also accounts for legitimate sources of temporal discontinuity (scene cuts, photo slideshows) to avoid false positives.

Watermark Analyst. Identifies generator-specific watermarks, editor tool marks, platform branding, and trial version overlays. This agent is purely VLM-based (no CV sub-skills), as watermark detection requires semantic understanding of text and logo content rather than pixel-level analysis.

Judge Agent. The Judge receives structured results from all active analysts and produces a final verdict through a three-stage decision mechanism:

1. **Strong evidence override**: If any analyst reports a fake score $\geq$0.85 with confidence $\geq$0.6, the video is immediately classified as fake. This rule captures cases where a single modality provides overwhelming evidence (e.g., a visible generator watermark).
2. **Majority vote**: Among analysts with confidence $\geq$0.6, a majority verdict (fake vs. real) is adopted if one exists.
3. **Dynamic weighted fusion**: When no consensus emerges, the Judge computes a confidence-weighted average of fake scores, with a disagreement penalty applied when analysts diverge significantly.

An optional visual review step allows the Judge to re-examine key frames alongside analyst reports before issuing the final verdict.

Sub-Skill Routing Details. CV sub-skills within each analyst are *not* executed unconditionally. Instead, a lightweight **vision sub-skill router** (a separate VLM call with up to 4 preview frames, denoted $\mathbf{V}_{\text{prev}} \subseteq \mathbf{V}_{\text{full}}$, where $\mathbf{V}_{\text{full}}$ is the up-to-40-frame set used by the analyst itself) consults

the analyst's skill manifest (a YAML document describing each sub-skill's purpose, strengths, and applicability) and the video's metadata to select only relevant sub-skills. For example, the spatial router may disable multi-face collapse detection for a landscape video, or the physics router may skip optical flow for a static slideshow. The complete sub-skill inventory is presented in Table 20 in the main text.

J Baseline Reproduction Protocol

For fair comparison, all supervised baselines are evaluated under the identical protocol specified by Zhang et al. (2025); Chen et al. (2024).

Training data. All five baselines (DeMamba, NPR, TALL, STIL, NSG-VD) are trained on 10,000 real/fake pairs: real videos from Kinetics-400 (Kay et al., 2017) and fake videos from Pika. No GenVideo test-set data is used during training.

Statistical reporting convention. Following NSG-VD (Zhang et al., 2025) and DeMamba (Chen et al., 2024), we report point estimates without error bars, reflecting the standard practice on the GenVideo benchmark where all prior work uses single-run evaluation. The test sets are fixed and public, and our pipeline is near-deterministic (temperature=0, fixed seeds), so variance from stochastic sampling is negligible. Per-generator sample sizes (56–1,400) are reported in Table 21 to allow readers to assess statistical power.

NSG-VD details. NSG-VD uses a Velocity_Single_TALL_MMD architecture with a SingleSwinBlockDiscriminator (feature dimension 300, input resolution 224×224, 8 frames/video). The detection method is deep_MMD (maximum mean discrepancy) with threshold >1 classifying as fake. The official pre-trained weights (`standard-Pika-d.pth`, 1.7 MB) are used. The reference set for MMD computation is drawn from MSR-VTT validation split (72 real velocity features). Training hyperparameters: lr=10^{-4}, sigma=1000, sigma0=0.1, epochs=200, batch_size=24.

Duration-controlled protocol. For the duration-controlled evaluation ($\geq$4 s), we filter all test videos by computing duration as frame_count/fps using OpenCV. The same filter is applied to both FOVEA and all baselines. Baseline numbers in Table 8 are obtained by running the official model checkpoints on the filtered test set under the same inference protocol as the standard evaluation.

Other baselines. DeMamba uses an XClip-Base + Mamba architecture; NPR uses ResNet-50; TALL uses Swin Transformer-B; STIL uses SCNet-50. All are trained with the Pika+Kinetics-400 protocol. Baseline numbers in the standard evaluation (Table 1) are taken from Zhang et al. (2025) to ensure consistency.

K Full GenVideo Dataset Statistics

The full GenVideo benchmark contains $\sim$2.3M videos: 1.08M AI-generated videos from 12 generators and 1.21M real videos for training. The training set generators include ZeroScope (133k), I2VGen-XL (62k), SVD (149k), AnimateDiff (26k), VideoCrafter (39k), Pika (98k), DynamiCrafter (46k), Stable Diffusion (201k), SEINE (25k), Latte (150k), OpenSora (177k), and CogVideo-5B. Real training videos are sourced from Kinetics-400 (Kay et al., 2017) (260k) and Youku-mPLUG (953k).

Duration Distribution Analysis. A key finding of our work is the severe duration distribution bias in the GenVideo test set. Table 22 details the filtering results when applying a 4-second minimum threshold. Six of ten generators lose *all* test videos (MorphStudio, MoonValley, HotShot, Show-1, Crafter, Lavie), retaining only four generators (ModelScope, Gen-2, Sora, WildScrape) with a total of 2,391 fake videos out of the original 8,588 (27.8%). The real test set (MSR-VTT) retains 92.0% of videos, confirming the imbalance. A simple duration-only baseline, classifying any video below 5 s as fake, would exploit this gap and achieve approximately 80–85% accuracy on the standard protocol; this further motivates the need for the duration-controlled evaluation in Section 5.

Table 21: **GenVideo test set statistics and real-world video distributions.** AI-generated test videos are predominantly ultra-short ($\leq$4 s) and low-resolution compared to real-world content. Platform statistics from YouTube (2024a;b); TikTok (2024); Meta (2025).

Video Source	Paradigm	#Videos	Typ. Duration	Resolution
GenVideo Test Set — AI-Generated (10 unseen generators)				
ModelScope (Wang et al., 2023)	Diffusion-UNet, T2V	700	4 s	256×256
MorphStudio	Proprietary, T2V	700	2 s	1280×720
MoonValley	Proprietary, T2V	626	$\leq$3 s	1024×576
HotShot	Diffusion-UNet, T2V	700	1 s	672×384
Show-1	Diffusion-UNet, T2V	700	$\leq$4 s	576×320
Gen-2 (Runway Research, 2023)	Proprietary, T2V+I2V	1,380	4 s	896×512
Crafter	Diffusion-UNet, T2V	1,400	$\leq$4 s*	256×256
Lavie	Diffusion-UNet, T2V	1,400	2 s	1280×2048
Sora (OpenAI, 2024)	Diffusion-Trans., T2V	56	$\leq$60 s	Varied
WildScrape	Mixed sources	926	2–6 s	512–1024
Subtotal (Fake Test)	—	*8,588*	*1–4 s (8/10)*	*256–1280p*
GenVideo Test Set — Real				
MSR-VTT (Xu et al., 2016)	Web video clips	10,000	10–30 s	Varied (320–720p)
Real-World Video Platforms (for reference)				
YouTube (YouTube, 2024a)	User-generated	20M+ uploads/day	Shorts $\leq$3 min; long-form varied	720p–4K
TikTok (TikTok, 2024)	Short-form UGC	—	15 s–10 min	1080×1920
Instagram Reels (Meta, 2025)	Short-form UGC	—	up to 3 min	1080×1920

*Measured actual duration $\sim$1.6 s, shorter than the generator's specified capability.

Table 22: **Duration filtering results ($\geq$4 s).** Number of videos retained per generator after applying the minimum duration threshold.

Generator	Total	Typ. Duration	Retained ($\geq$4 s)	Retention Rate
ModelScope	700	4 s	700	100.0%
MorphStudio	700	2 s	0	0.0%
MoonValley	626	$\leq$3 s	0	0.0%
HotShot	700	1 s	0	0.0%
Show-1	700	$\leq$4 s	0	0.0%
Gen-2	1,380	4 s	1,348	97.7%
Crafter	1,400	$\leq$4 s	0	0.0%
Lavie	1,400	2 s	0	0.0%
Sora	56	$\leq$60 s	53	94.6%
WildScrape	926	2–6 s	290	31.3%
Total Fake	8,588	—	2,391	27.8%
MSR-VTT (Real)	10,000	10–30 s	9,200	92.0%

Why 4 seconds? We choose 4 s as a deliberately conservative lower bound for three reasons. First, 4 s exceeds the hard length limits of the earliest short-form platforms that still reflect a human-recognizable action unit: Vine capped clips at 6 s and Snapchat at 10 s. Second, modern short-form platforms such as YouTube Shorts (YouTube, 2024b) (up to 3 min), Instagram Reels (Meta, 2025) (up to 3 min), and TikTok (TikTok, 2024) (15 s to 10 min) all distribute videos well above 4 s, so any clip below this threshold is effectively unusable as standalone disinformation. Third, a 4 s window contains enough frames (typically 96–120 at 24–30 fps) for meaningful temporal reasoning such as optical-flow analysis and motion-coherence checks; below this, any method degenerates toward single-image classification.

This duration gap is problematic because (i) ultra-short AI-generated videos (<4 s) are rarely encountered in real-world misinformation scenarios; (ii) duration itself can serve as a trivial confounding signal; and (iii) low frame counts in ultra-short clips limit the effectiveness of temporal analysis methods. We acknowledge that the $\geq$4 s filter retains only 4 of 10 generators, an inherent limitation of GenVideo's duration distribution rather than a design choice to favor our method. A duration-matched control experiment (trimming real videos to 2–4 s) and complete re-evaluation are planned for the camera-ready version.