SafeGuard: A Multi-Agent Perception-Reasoning Framework for Social-Risk AI-Generated Video Detection

Wenlin Wu[1*], Sheng Zhou[2*], Peipei Song[1], Wenhao Wang[3], Junbin Xiao[1†], and Xun Yang[1†]

[1] University of Science and Technology of China
[2] King Abdullah University of Science and Technology
[3] Vast Intelligence Lab
wenlin097@mail.ustc.edu.cn, {xyang21,junbinxiao}@ustc.edu.cn

Abstract. As video generation paradigms evolve from localized manipulation to full-scene synthesis, AI-generated video detection becomes increasingly challenging, as forgeries exhibit coherent global structure and high perceptual realism. However, existing benchmarks are biased toward perceptual fidelity and primarily evaluate detectors based on perceptual artifacts, providing limited coverage of scenarios that require reasoning about violations of physical laws, structural coherence, or social logic. This dataset bias shapes current approaches and results in a *Perception–Reasoning Gap*: artifact-centric models capture low-level statistical irregularities yet lack semantic inference, whereas vision-language models perform semantic reasoning but remain insensitive to fine-grained forensic cues. To bridge this gap, we propose *SafeGuard*, a multi-agent framework that enables collaborative specialization between forensic perception and semantic reasoning. A hierarchical perceptual solver extracts fine-grained forensic evidence, while a self-reflective verifier enforces consistency between semantic inference and physical plausibility, forming an interpretable evidence chain. To support evaluation, we introduce *SafeVid*, a novel AI-generated video detection benchmark comprising 20K videos spanning 10 social risk categories, designed to evaluate physical plausibility, structural consistency, and the rationality of social behaviors. Extensive experiments demonstrate the generalization of *SafeGuard*, improving accuracy on SafeVid by +18.7% and consistently outperforming prior methods across four public benchmarks. The code and dataset are publicly available at https://github.com/williamw99/SafeGuard.

Keywords: AI-Generated Video Detection · Social Safety · Multi-Agent Collaboration · Benchmark

* Equal contribution; † Corresponding authors: Xun Yang and Junbin Xiao.

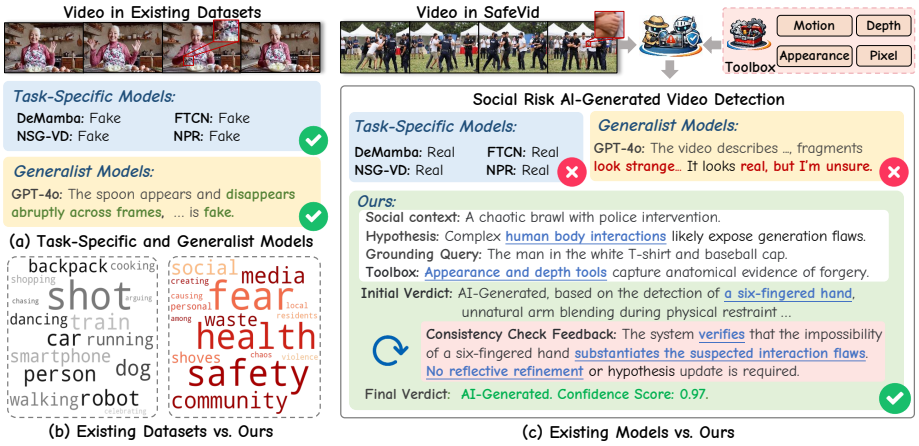

Fig. 1: The illustration of social-risk AI-generated video detection. Our framework integrates perceptual forensic evidence with semantic reasoning.

1 Introduction

Recent advances in video generation have shifted synthesis from localized manipulation (*e.g.*, facial editing [67]) to full-scene generation. Modern systems, including Sora [39], Pika [35], Wan [46], Nano Banana Pro [11], and controllable generation [8, 30], can produce globally coherent and semantically consistent videos that closely resemble authentic footage. Unlike early deepfake research [17,29,52], which focused on identity manipulation, current generative models can synthesize complete events, including misinformation, violent incidents, and politically sensitive scenarios. As synthetic videos grow in semantic complexity and social impact, detection approaches extend beyond low-level artifact analysis to evaluate event-level authenticity and contextual plausibility [53]. However, existing AI-generated video detection benchmarks [6, 38, 43, 49, 51] (*e.g.*, GenVideo [6], DVF [43], and GenVidBench [38]) emphasize visual realism in benign scenarios, such as daily activities or natural scenes, as illustrated in Fig. 1 (a), while largely underrepresenting high-risk semantic contexts. This limited coverage creates a distribution gap between benchmark data and real-world social-risk environments, limiting the generalization of detection systems in safety-critical settings. These limitations motivate benchmarks and detection frameworks tailored to social-risk AI-generated videos, enabling semantically grounded and context-aware authenticity assessment.

However, existing AI-generated video detection methods face a fundamental *Perception–Reasoning Gap*: task-specific detectors excel at extracting low-level forensic cues but struggle to assess physical feasibility or social-contextual consistency, resulting in limited robustness and interpretability. In contrast, large vision–language models, such as Qwen3-VL [3], InternVL3.5 [48], and GPT-4o [19], provide strong semantic reasoning yet remain insensitive to subtle perceptual evidence. As illustrated in Fig. 1 (c), this limitation leads to critical failures in social-risk scenarios. This tension reflects an inherent trade-off: preserving

fine-grained perceptual evidence for forensic analysis often conflicts with the abstraction required for semantic reasoning.

Motivated by this challenge, we propose **SafeGuard**, a multi-agent framework that combines the fine-grained evidence extraction capabilities of low-level visual forensic models with the high-level semantic reasoning abilities of large vision-language models. The framework formulates detection as a closed-loop, evidence-driven reasoning process. Specifically, it consists of two complementary modules: a *Hierarchical Perceptual Solver* and a *Self-Reflective Verifier*. The Solver adopts a coarse-to-fine strategy to localize critical regions and extract explicit forensic cues using lightweight specialized models. The Verifier then evaluates semantic–perceptual consistency by cross-validating the extracted evidence against contextual reasoning. As illustrated in Fig. 1 (c), when analyzing a complex social event such as a chaotic brawl, the Solver first formulates a semantic hypothesis and deploys targeted tools to gather concrete forensic evidence (*e.g.*, detecting a six-fingered hand). The verifier then leverages this physical anomaly as evidence to validate the hypothesis and reach the current verdict. Through this iterative mutual grounding process, SafeGuard reconciles low-level forensic sensitivity with high-level semantic reasoning, producing interpretable decisions supported by an explicit chain of evidence.

To validate our framework, we construct a social-risk AI-generated video detection benchmark, **SafeVid**, consisting of 20K videos with 10K authentic and 10K AI-generated samples. The videos are produced by 21 state-of-the-art generative models and cover 10 social risk categories. SafeVid is designed to evaluate detection robustness under high-risk semantic conditions and to encourage generalization beyond visually benign scenarios, as illustrated in Fig. 1. Extensive experiments demonstrate that SafeGuard consistently outperforms state-of-the-art models on SafeVid and four existing public benchmarks, including LOKI [68], GenVideo [6], GenVidBench [38], and DVF [43]. Different from prior training-based detection methods that require task-specific fine-tuning on target datasets, our overall framework requires no end-to-end fine-tuning for the video detection task. Notably, our adopted lightweight forensic tools utilize publicly available GenVideo-finetuned weights, and our ablation studies confirm these optimized cues yield non-trivial performance gains in social-risk AI-generated video detection. Beyond superior detection accuracy, our framework provides a transparent reasoning process that links semantic suspicion, forensic cue extraction, and logical deduction, enabling interpretable assessment of socially sensitive risks.

In summary, the main contributions are as follows:

- We introduce the task of social-risk AI-generated video detection, requiring models to jointly assess physical plausibility, identity consistency, and social-logic coherence beyond traditional artifact-based inspection.
- We propose SafeGuard, a multi-agent framework that integrates fine-grained forensic cue extraction via a Hierarchical Perceptual Solver with semantic–physical consistency verification through a Self-Reflective Verifier, bridging the perception-reasoning divide to jointly analyze semantic falsehoods and visual forensic traces in AI-generated video detection.

– We construct SafeVid, a benchmark of 20K videos generated by 21 state-of-the-art models across 10 socially critical risk categories, curated to evaluate detection robustness in semantically complex and socially sensitive scenarios.
– Extensive experiments on SafeVid and four public benchmarks demonstrate the effectiveness and generalization capability of SafeGuard, which achieves strong detection performance across diverse social-risk scenarios.

2 Related Works

2.1 AI-Generated Video Detection Methods and Datasets

AI-generated video detection aims to distinguish synthetic videos from authentic content, and it inherently intersects with vision-language comprehension and reasoning paradigms [9, 25, 57, 62–64, 77]. With the evolution of generative models, the task has progressed from low-level artifact inspection to jointly modeling perceptual cues and high-level semantic consistency. Existing approaches can be categorized as: **(1) Training-based methods.** These methods leverage supervised or instruction-tuned learning to capture discriminative forgery evidence. Early models focus on appearance, motion, or geometric inconsistencies, *e.g.*, DeMamba [6] and DeCoF [34] model pixel- and flow-level artifacts, while WaveRep [12], ReStraV [20], and NSG-VD [71] incorporate physical or frequency-domain cues. Traditional AI-generated image detectors [44, 75] and deepfake methods [10,37,59] follow a similar paradigm, exploiting texture, blending, or warping artifacts. Recent training-based approaches [27, 43, 54] extend this with semantic modeling via large multimodal architectures, using multimodal alignment or instruction tuning to extract forgery evidence. However, they require substantial training and may underutilize fine-grained perceptual cues. **(2) Training-free methods.** These approaches avoid task-specific optimization by leveraging external priors or foundation models. Physics-inspired methods like D3 [74] assess consistency with physical laws, whereas LLM-driven pipelines [14] such as LAVID [31] frame detection as an agent-based reasoning process. Training-free methods reduce computational cost and improve interpretability but often sacrifice fine-grained perceptual grounding or semantic reasoning, limiting performance in complex scenarios.

While these methods advance detection capabilities, their evaluation is largely based on benchmarks covering generic or everyday scenarios. Existing datasets, such as GenVidBench [38], GenVideo [6], DVF [43] and LOKI [68], primarily focus on general real-world content (*e.g.*, landscapes and daily activities). Policy-oriented datasets like Video-SafetyBench [33] emphasize harmful content moderation but lack forensic-level annotations for real–synthetic discrimination. Consequently, current benchmarks provide limited coverage of socially sensitive and high-risk scenarios. To fill this gap, SafeVid introduces an AI-generated video detection benchmark tailored to social safety, safety-critical events (*e.g.*, protests and riots), enabling evaluation of detection frameworks that jointly model fine-grained perceptual cues and high-level semantic reasoning.

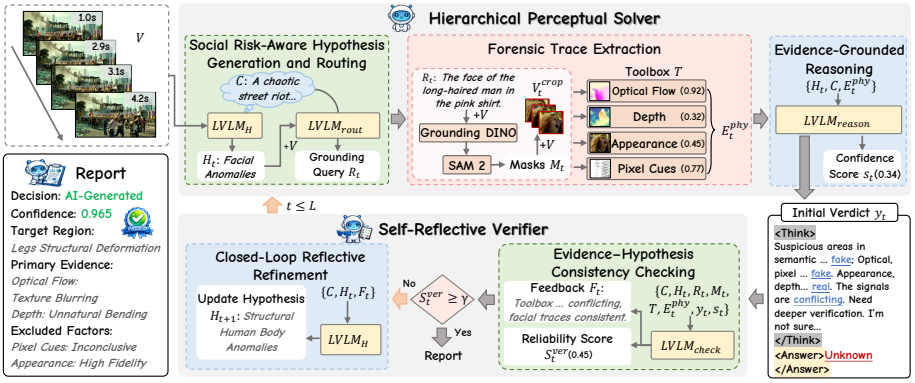

Fig. 2: The SafeGuard Framework. SafeGuard combines a **Hierarchical Perceptual Solver** for coarse-to-fine trace extraction with a **Self-Reflective Verifier** that checks evidence–hypothesis consistency and performs closed-loop refinement.

2.2 Multi-Agent System for Visual Forensics

Multi-Agent Systems (MAS) address complex tasks by decomposing problem-solving into coordinated interactions among specialized agents, enabling scalable and modular reasoning [16, 58]. In visual forensics, MAS have been applied primarily to manipulation detection [18, 24, 28], such as face forgery and deepfake identification. Methods like UniShield [18] and AIFo [28] employ role-specialized agents for adversarial analysis, while Agent4FaceForgery [24] uses agent societies to simulate forgery and evaluate detector robustness. These approaches excel at capturing fine-grained visual artifacts but typically lack semantic reasoning, limiting their interpretability in context-dependent manipulations. MAS have also been explored in AI safety and multimodal risk assessment [1], using structured debate and adversarial argumentation to evaluate semantic risks and mitigate unsafe behaviors. While effective at high-level reasoning, they generally overlook subtle visual cues, reducing their applicability to perceptually grounded detection. Overall, existing MAS approaches tend to focus on low-level visual forensics or high-level semantic reasoning, without a unified mechanism integrating both. Our framework demonstrates that orchestrating multi-agent collaboration with hierarchical perceptual and semantic reasoning can bridge this gap, enabling artifact detection and social event-level assessment.

3 The SafeGuard Framework

In this section, we propose an agent-centric framework, SafeGuard, which separates low-level perception and high-level reasoning into a *Hierarchical Perceptual Solver* and a *Self-Reflective Verifier*, respectively. The solver generates social risk-aware forgery hypotheses, extracts structured forensic evidence, and produces an initial verdict. The verifier audits this output for evidence–hypothesis consistency and performs reflective refinement when grounding is insufficient.

This two-level design structures AI-generated video detection into semantic suspicion, evidence grounding, and logically validated decision-making.

3.1 Hierarchical Perceptual Solver

Social Risk-Aware Hypothesis Generation and Routing AI-generated video forgeries often concentrate in semantically and socially salient regions (*e.g.*, human bodies or interactions), while static content like buildings is usually well-generated, as shown in Fig. 2. To focus analysis on high-risk areas, we introduce a social risk-aware module for hypothesis generation and routing. A large vision-language model (LVLM) summarizes scene dynamics into a global description C (*e.g.*, a chaotic street riot with intense interactions) and generates a forgery hypothesis H_t specifying plausible artifact types (*e.g.*, identity swapping, limb distortion, temporal tearing):

$$H_t = \mathrm{LVLM}_{\mathrm{H}}(C). \tag{1}$$

This hypothesis then guides coarse spatial–temporal routing, identifying candidate regions R_t for analysis while filtering out irrelevant background:

$$R_t = \mathrm{LVLM}_{\mathrm{rout}}(V, C, H_t). \tag{2}$$

By coupling contextual abstraction with hypothesis-driven region selection, this module concentrates forensic scrutiny on semantically high-risk areas, improving detection accuracy and efficiency without sacrificing reliability.

Forensic Trace Extraction High-level semantic analysis alone is insufficient for video authentication, as generation artifacts are often subtle, localized, and physically grounded. LVLMs may overlook fine-grained inconsistencies such as motion jitter, geometric distortions, or boundary noise. Hence, the solver incorporates a localized and physically sensitive mechanism to verify suspected manipulations. Specifically, this module performs fine-grained forensic analysis on candidate regions identified by the preceding module. We adopt a language-guided segmentation pipeline for precise region localization. Given the grounding query R_t, frame-wise spatial grounding is conducted over the N input frames of V. Grounding DINO [32] is employed to predict a set of localized bounding boxes $\{B_t^i\}_{i=1}^N$. Conditioned on these bounding boxes, SAM 2 [42] generates the corresponding pixel-level masks $M_t = \{M_t^i\}_{i=1}^N$. The masks M_t are subsequently applied to the original video V to isolate the target content, yielding a set of focused regions $V_t^{\mathrm{crop}} = V \odot M_t$. Structured forensic cues E_t^{phy} are then derived via a task-specific toolbox T, where each tool assesses the authenticity of the cropped region from complementary physical perspectives.

$$E_t^{\mathrm{phy}} = \Big\{ f_{\mathrm{o}}(V_t^{\mathrm{crop}}), f_{\mathrm{d}}(V_t^{\mathrm{crop}}), f_{\mathrm{a}}(V_t^{\mathrm{crop}}), f_{\mathrm{p}}(V_t^{\mathrm{crop}}) \Big\}, \tag{3}$$

where $f_{(\cdot)}(V_t^{\mathrm{crop}}) \in (0, 1)$ represents the estimated probability that the region is real. Specifically, $f_{\mathrm{o}}(\cdot)$ captures low-level motion information through optical

flow to detect temporal inconsistencies, $f_{\mathrm{d}}(\cdot)$ focuses on geometric cues derived from depth to identify physically implausible structures, $f_{\mathrm{a}}(\cdot)$ encodes discriminative appearance features to ensure object-level identity consistency, and $f_{\mathrm{p}}(\cdot)$ examines high-frequency pixel-level signals, such as boundary noise and subtle manipulation traces. By integrating these confidence scores, the system establishes an evidential representation that captures physical inconsistencies across motion, depth, appearance, and pixel-level cues.

Evidence-Grounded Reasoning The presence of localized forensic cues E_t^{phy} alone is insufficient for reliable video integrity assessment, as low-level irregularities may admit multiple semantic interpretations. In complex social scenarios, identical physical patterns could arise from either natural dynamics (*e.g.*, rapid motion or occlusion) or generative manipulation. Consequently, decision-making requires reasoning that evaluates evidence under semantic and hypothesis constraints. To this end, this stage performs hypothesis-conditioned reasoning by jointly modeling the extracted evidence E_t^{phy}, the global semantic context C, and the forgery hypothesis H_t. Rather than directly mapping cues to a categorical outcome, the model constructs a structured inference process that examines whether the observed physical signals are causally and logically consistent with the anomaly mechanism postulated in H_t. Formally, the reasoning module produces an initial verdict y_t and a confidence score s_t:

$$(y_t, s_t) = \mathrm{LVLM}_{\mathrm{reason}}\left(H_t, C, E_t^{\mathrm{phy}}\right), \tag{4}$$

By integrating context, hypothesis, and localized forensic cues, this reasoning converts isolated low-level observations into a coherent forensic narrative (*e.g.*, detecting temporal tearing inconsistent with motion dynamics). This ensures that conclusions are both logically grounded and physically traceable, providing the transparency and reliability required for high-stakes video forensics.

3.2 Self-Reflective Verifier

Evidence–Hypothesis Consistency Checking Although the Hierarchical Perceptual Solver generates an initial verdict y_t with a confidence score s_t, this preliminary result may still contain conflicts or hallucinations [4, 65]. To ensure the reliability of the reasoning process, this stage functions as a rigorous gatekeeper. Its primary objective is to validate the coherence of the solver's entire execution trace before accepting the final verdict. Specifically, the verifier performs a multi-criteria audit that scrutinizes the plausibility of the hypothesis H_t within the social risk context C, the precision of the segmentation masks M_t in isolating target regions R_t, the appropriateness of the four task-specific tools $T = \{f_{\mathrm{o}}, f_{\mathrm{d}}, f_{\mathrm{a}}, f_{\mathrm{p}}\}$ generating E_t^{phy} regarding the scene dynamics, and the logical consistency between the extracted cues E_t^{phy} and the initial verdict y_t. This verification yields a reliability score S_t^{ver} together with feedback F_t:

$$(S_t^{\mathrm{ver}}, F_t) = \mathrm{LVLM}_{\mathrm{check}}(C, H_t, R_t, M_t, T, E_t^{\mathrm{phy}}, y_t, s_t), \tag{5}$$

where specific failure modes can be explicitly identified. When low reliability is detected, the verifier not only flags errors but also corrects reasoning results: unlike passive auditing, it acts as an active corrector by translating identified failure modes into natural-language feedback F_t (*e.g.*, "the motion evidence contradicts the static background assumption"), which provides actionable guidance for correction. If the score falls below a critical reliability threshold (*i.e.*, $S_t^{\text{ver}} < \gamma$), the system rejects the initial verdict and triggers the subsequent refinement phase.

Closed-Loop Reflective Refinement When $S_t^{\text{ver}} < \gamma$, the system leverages the feedback F_t to update the hypothesis from H_t to H_{t+1} and re-initiates the perception-reasoning cycle. This iteration continues until the verdict passes the consistency check or reaches the predefined cycle bound L.

$$H_{t+1} = \begin{cases} H_t, & \text{if } S_t^{\text{ver}} \geq \gamma \text{ or } t > L, \\ \text{LVLM}_{\text{H}}(C, H_t, F_t), & \text{otherwise.} \end{cases} \tag{6}$$

By integrating rigorous consistency verification with adaptive refinement, the proposed verifier ensures that the final reports are not only accurate but also explanatorily grounded and explicitly traceable to concrete forensic evidence, while preserving logical coherence. This reliability-aware design is especially crucial in high-stakes forensic contexts, where minimizing artifact-induced false positives is as important as ensuring precise detection.

4 The SafeVid Dataset

To facilitate research on social-safety-oriented video forensics, we construct a new benchmark named SafeVid. In this section, we first introduce the safety-driven taxonomy. We then describe the procedures for real-world video collection and diversified synthetic video generation under multiple paradigms. Finally, we present statistical analysis and comparisons with existing related benchmarks.

4.1 Datasets Construction

SafeVid Taxonomy To cover representative social safety risks, we define a taxonomy of 10 fine-grained categories aligned with safety guidelines [73] and existing benchmarks [26, 33, 50, 76]. These categories are *Violence and Aggression*, *Public Safety Incidents*, *Corruption and Ethical Violations*, *Radical Behavior and Terrorism*, *Criminal Activities*, *Misinformation and Propaganda*, *Discrimination and Exclusion*, *Public Health Risks*, *Invasion of Privacy*, and *Industrial and Environmental Hazards*. The taxonomy forms the semantic backbone of SafeVid, guiding both real-world video collection and synthesized video generation.

Real-World Video Collection To ensure diversity and realism, we collect videos from three public sources: YouTube [69], OpenVid-1M [36], and UltraVideo [60]. To align with the predefined taxonomy, we design a filtering strategy

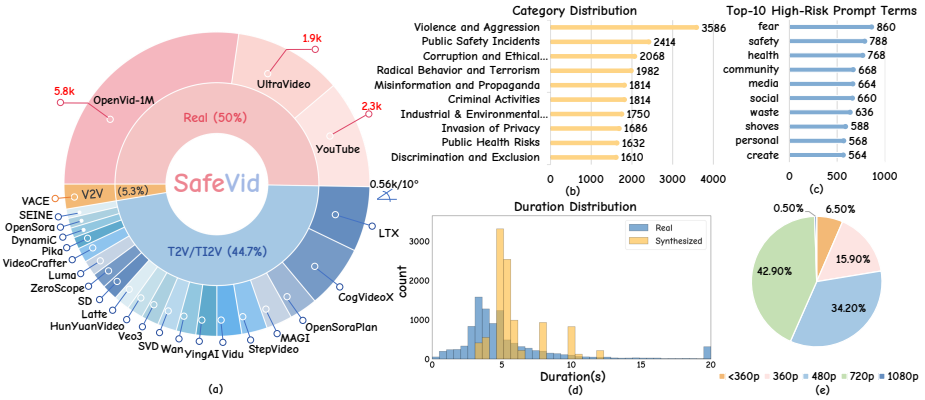

Fig. 3: SafeVid statistics. (a) Video count distribution. (b) Social risk category distribution. (c) Top-10 high-risk terms in video generation prompts (including T2V and TI2V). (d) Video duration distribution. (e) Video resolution distribution.

to filter out irrelevant videos. CLIP [41] computes semantic similarity between candidate videos and taxonomy keywords, and videos are retained if the similarity exceeds a predefined threshold of 0.75. This filtering approach mitigates noise introduced by keyword-based retrieval and ensures semantic consistency with the predefined public safety categories. Through this process, we obtain 10,178 authentic videos.

Synthesized Video Generation To simulate an evolving social risk landscape, we construct synthetic data under two settings: **(1) Text-Driven Video Synthesis.** Given a safety keyword (*e.g.*, traffic accident) and category (*e.g.*, public safety incidents), we build scenario prompts and enrich them with scene attributes (*e.g.*, lighting, motion, environment) using GPT-4o [19], forming a diverse prompt pool. Videos are then generated via two routes: ① *Text-to-Video (T2V)*, rendered by multiple state-of-the-art models (*e.g.*, Veo3 [15], Wan [46], CogVideoX [66], HunYuanVideo [23]); ② *Text+Image-to-Video (TI2V)*, where anchor images produced by a text-to-image model (*e.g.*, Qwen-Image [56]) are animated by an image-to-video model (*e.g.*, SVD [5]). This process yields 6,473 T2V and 3,166 TI2V videos. **(2) Real-Video Editing.** To diversify manipulation patterns, we perform video-to-video editing on authentic videos. Semantic targets (*e.g.*, identity or object replacement) are segmented using SAM 2 [42], and edited via a V2V model (*e.g.*, VACE [22]) conditioned on the source video, mask, and instruction. This results in 539 edited videos with varied artifacts.

4.2 Dataset Analysis and Comparison

Statistics and Analysis SafeVid contains 20,356 videos, evenly divided into 10,178 real and 10,178 synthetic videos. Specifically, T2V/TI2V videos constitute 44.7% and V2V samples 5.3%, complementing the 50% real portion. This

Table 1: Comparison of AI-generated video detection datasets. **Scenario**: *General* denotes all open-domain scenes, such as cartoons and animations; *Real-World* consists of videos resembling realistic, physically plausible environments; and *Social Safety* focuses on social safety-oriented scenarios curated for high-risk or sensitive content. **Tax.** indicates the fine-grained social safety taxonomy.

Dataset	Scenario	Gen. Paradigm			Tax.	Video Source	#Mod.	#Vid.
		T2V	TI2V	V2V				
GVD [2]	General	✓	✓	✗	✗	VOS2, Got-10k	8	22k
GVF [34]	General	✓	✗	✗	✗	MSCD, MSR-VTT	8	5k
GenVidDet [21]	General	✓	✓	✗	✗	InternVid, HD-VG-130M	8	**73k**
DVF [43]	General	✓	✓	✗	✗	Youtube-8M, InternVid-10M	8	4k
GenVideo [6]	General	✓	✓	✗	✗	Kinetics-400, Youku-mPLUG, MSR-VTT	20	19k
GenVidBench [38]	General	✓	✓	✗	✗	Vript, HD-VG-130M	8	69k
LOKI [68]	General	✓	✗	✗	✗	-	7	1k
VidForensic [31]	General	✓	✗	✗	✗	PANDA-70M, Youtube	8	2k
GenWorld [7]	Real-World	✓	✓	✗	✗	Kinetics-400, Nuscenes, RT-1, DL3DV-10K	10	20k
GenBuster [55]	Real-World	✓	✗	✗	✗	OpenVid-1M	12	4k
SafeVid (Ours)	**Social Safety**	✓	✓	✓	✓	**YouTube, OpenVid-1M, UltraVideo**	**21**	20k

balanced composition enables evaluation across both fully generative and real-video editing scenarios, enhancing artifact diversity and forensic difficulty. The dataset is explicitly designed for social safety assessment. As shown in Fig. 3 (b), *Violence and Aggression* accounts for the largest share (∼18%), with the remaining categories evenly distributed, ensuring broad coverage of socially sensitive contexts. The prompt statistics for video generation (including T2V and TI2V) in Fig. 3 (c) further demonstrate the dataset's safety-oriented design. High-frequency terms such as *fear*, *safety*, and *health* indicate an alignment between the semantic distribution of generated content and real-world risk scenarios.

Dataset Comparison As shown in Table 1, SafeVid is the first AI-generated dataset dedicated to the social safety domain, focusing on safety-critical events rather than generic content. Unlike prior datasets that predominantly rely on T2V synthesis (with limited TI2V synthesis and no V2V editing), SafeVid encompasses T2V, TI2V, and V2V, enabling evaluation across both fully generative and real-video editing scenarios and increasing artifact diversity. In addition, SafeVid introduces a fine-grained safety taxonomy and incorporates 21 generative models to enhance source heterogeneity. The combination of multi-paradigm synthesis, safety-aware categorization, and diverse generation models establishes a challenging benchmark for social safety assessment.

5 Experiments

5.1 Experimental Setup

Model Evaluation To assess the effectiveness and generalization of SafeGuard, we benchmark it against state-of-the-art approaches that reflect two paradigms in AI-generated video detection: **(1) Task-specific models.** These approaches are explicitly developed for AI-generated video detection and rely on supervision tailored to forensic discrimination. We include *trainable* approaches spanning

Table 2: Dataset splits of SafeVid. The dataset comprises **Real** and **Synthesized** videos (*i.e.*, videos generated or edited by T2V/TI2V/V2V models).

Type	In-Domain (ID)	Out-of-Domain (OOD)
Real	YouTube (2.3k)	OpenVid-1M, UltraVideo (7.7k)
Synthesized	SVD, OpenSora, DynamiC, Latte, ZeroScope, Pika, SD, SEINE, VideoCrafter (2.3k)	LTX, YingAI, StepVideo, Vidu, CogVideoX, OpenSoraPlan, HunyuanVideo, MAGI, VACE, Veo3, Wan, Luma (7.7k)

both image-level (NPR [44]) and video-level (FTCN [75], TALL [59], XCLIP [37], MINTIME [10], DeMamba [6], ReStraV [20], NSG-VD [71], BusterX++ [55], and Skyra-RL [27]) modeling. We further evaluate the *training-free* approach D3 [74] to assess artifact-level generalization without task-specific retraining. **(2) Generalist models.** To evaluate the performance of general-purpose large vision-language foundation models on AI-generated video detection, we benchmark these models under a zero-shot prompting setting. We consider advanced open-source models, including LLaVA-Video [72],VideoLLaMA3 [70], InternVL3.5 [48], Qwen3-VL [3], and reasoning-oriented Video-R1 [13], Video-RFT [47], VideoChat-R1 [25], as well as the proprietary model GPT-4o [19], and Gemini-2.5-Pro [11]. All generalist models are evaluated using a unified prompt template to ensure a fair comparison. Detailed implementation settings, including hyperparameters and prompt templates, are provided in the Appendix.

Datasets We evaluate all methods on our SafeVid dataset and four widely used public benchmarks: `GenVideo` [6], a large-scale benchmark for AI-generated video detection comprising 2,262k training videos and 18.5k testing videos; `DVF` [43], which includes 6.7k diffusion-generated videos produced by eight distinct diffusion-based generators; `LOKI` [68], a benchmark featuring 1.3k synthesized videos in diverse real-world contexts; and `GenVidBench` [38], containing 143k videos and designed to assess cross-model and cross-scenario generalization. For experiments, the training set is fixed and drawn from the `GenVideo` [6] training set, *i.e.*, $D_{\text{train}} = \{\texttt{GenVideo}_{\text{train}}\}$. The test set is partitioned into two complementary scenarios: **(1) In-Domain (ID)**, containing 4.6k video samples generated using the same sources and methods as in the training set; **(2) Out-of-Domain (OOD)**, containing 15.4k videos from completely unseen sources and generative models. Formally, the overall evaluation is defined as $D_{\text{test}} = D_{\text{SafeVid}} \cup D_{\text{public}}$, where $D_{\text{SafeVid}} = \{R_{\text{ID}}, S_{\text{ID}}, R_{\text{OOD}}, S_{\text{OOD}}\}$ (R denotes real videos, S denotes synthesized videos) and $D_{\text{public}} = \{\texttt{GenVideo}_{\text{test}}, \texttt{DVF}, \texttt{LOKI}, \texttt{GenVidBench}\}$. The data sources for `SafeVid` splits are summarized in Table 2.

Evaluation Metric Following prior work [27], we adopt Accuracy (**Acc**) and F1-score (**F1**) as our primary evaluation metrics, with F1 computed using *Synthesized* videos as the positive class based on hard predictions. For task-specific trainable models, we additionally report Area Under the Curve (**AUC**), Average Precision (**AP**), and Equal Error Rate (**EER**) in the Appendix following [6, 52].

Table 3: Model comparison across SafeVid and other existing benchmarks. *T*, *F*, *O*, and *P* denote *Trainable*, *Training-free*, *Open-source*, and *Proprietary* methods. For evaluation efficiency, BusterX++ [55], Skyra-RL [27], and the generalist models are evaluated on randomly sampled 30% subsets of each benchmark, and the human evaluators are assessed on 10% randomly sampled subsets. In all cases, stratified sampling is applied to preserve the real/synth ratio. † indicates models pre-trained on `GenVideo` training set. Qwen3-VL and InternVL3.5 refer to Qwen3-VL-32B [3] and InternVL3.5-38B [48], respectively. **Best** and Second-best are highlighted. Blue highlights the best-performing baseline across both task-specific and generalist models.

Method	Set	SafeVid (Ours)						Existing Benchmarks								Overall	
		ID		OOD		Avg.		GenVideo		DVF		LOKI		GenVidBench		Avg.	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Human		77.0	85.4	80.9	80.0	79.2	82.5	-	-	-	-	-	-	-	-	79.2	82.5
Task-Specific Models																	
NPR [44]	*T*	71.6	75.5	62.8	63.3	65.0	66.6	85.3†	71.6†	54.6	54.7	57.2	63.8	53.0	60.8	63.0	63.5
FTCN [75]	*T*	82.3	82.5	41.6	44.4	51.6	53.5	95.6†	95.0†	83.4	83.7	68.1	74.3	65.3	72.4	72.8	75.8
TALL [59]	*T*	70.4	74.1	52.2	57.7	56.7	61.7	91.7†	90.5†	71.6	70.8	65.9	72.8	65.0	72.6	70.2	73.7
XCLIP [37]	*T*	73.2	70.5	39.9	35.0	48.0	43.6	85.7†	76.6†	82.1	82.2	66.0	69.7	62.7	69.6	70.0	71.1
MINTIME [10]	*T*	78.8	79.3	45.5	44.2	53.7	53.1	95.3†	94.6†	83.6	83.9	68.7	73.2	64.8	71.8	73.2	75.3
DeMamba [6]	*T*	78.8	80.3	43.0	48.5	51.8	56.1	94.4†	90.2†	90.1	90.9	69.0	73.8	71.4	78.3	75.5	78.7
RestraV [20]	*T*	36.1	49.7	46.5	61.0	43.9	58.4	73.9†	72.9†	33.3	47.4	51.9	66.5	61.8	76.1	53.0	64.3
NSG-VD [71]	*T*	52.4	16.8	51.7	13.4	51.9	14.3	94.0†	94.2†	68.0	43.4	32.7	3.2	67.0	52.6	62.7	41.5
BusterX++ [55]	*T*	69.7	68.8	64.7	55.2	66.8	61.7	76.1	85.0	70.4	66.6	70.4	75.1	51.6	59.1	67.1	69.5
Skyra-RL [27]	*T*	63.3	43.5	56.2	24.1	59.3	33.0	71.8	59.0	49.2	21.7	45.6	33.7	25.8	13.5	50.3	32.2
D3 [74]	*F*	42.9	20.9	49.4	40.1	46.6	31.9	12.7	2.5	46.5	24.3	48.0	26.4	34.2	34.2	37.6	23.9
Generalist Models																	
LLaVA-Video [72]	*O*	35.9	19.2	44.7	28.3	41.0	24.3	48.3	62.4	56.6	55.5	46.7	48.9	53.0	64.8	49.1	51.2
VideoLLaMA3 [70]	*O*	36.2	11.8	48.9	16.9	43.4	14.5	55.2	68.8	58.3	46.5	49.2	42.2	48.4	56.6	50.9	45.7
InternVL3.5 [48]	*O*	42.9	18.1	52.7	19.7	48.5	19.0	58.5	70.9	61.2	50.1	56.3	55.2	47.3	54.3	54.4	49.9
Qwen3-VL [3]	*O*	62.0	60.2	62.1	47.2	62.1	53.7	78.4	86.7	75.0	73.2	70.0	74.8	58.3	67.2	68.8	71.1
Video-R1 [13]	*O*	42.6	23.3	49.3	25.7	46.4	24.6	44.7	57.3	55.8	42.8	49.5	44.0	42.9	48.6	47.9	43.5
Video-RFT [47]	*O*	43.3	20.2	49.1	22.0	46.6	21.2	45.8	58.5	55.0	39.6	48.0	40.4	42.4	47.8	47.5	41.5
VideoChat-R1 [25]	*O*	44.0	24.8	53.4	25.9	49.4	25.4	56.1	69.0	60.2	48.6	50.9	45.3	45.3	52.2	52.4	48.1
GPT-4o [19]	*P*	65.9	70.1	63.7	52.0	64.6	60.5	77.6	86.0	74.5	75.6	55.8	66.4	61.5	68.8	66.8	71.5
Gemini-2.5-Pro [11]	*P*	64.3	59.0	58.4	39.6	61.0	48.9	83.2	89.8	76.8	78.9	59.6	70.4	54.6	60.9	67.1	69.8
Agentic Frameworks																	
SafeGuard(Qwen3-VL)		92.2	90.0	76.2	71.2	82.9	79.7	97.2	95.1	96.3	**93.2**	**84.7**	**83.8**	80.6	78.7	88.3	86.1
SafeGuard(GPT-4o)		**93.0**	**90.7**	**80.1**	**72.8**	**85.5**	**80.9**	**97.9**	**95.6**	**97.0**	**93.2**	81.7	77.7	**84.6**	**85.1**	**89.3**	**86.5**

Implementation Details We implement SafeGuard with GPT-4o [19] as the vision–language backbone for the Hierarchical Perceptual Solver. To alleviate bias and failure modes from single-model reliance, we adopt Gemini-2.5-Pro [11] as the backbone for the Self-Reflective Verifier. For video processing, we uniformly sample 8 frames as model input. In the Forensic Trace Extraction stage (Sec. 3.1), `grounding-dino-base` [32] is used for coarse localization, followed by SAM 2 initialized with the pre-trained `sam2.1_hiera_base_plus` weights [42] for pixel-wise masking. The forensic toolbox $T = \{f_o, f_d, f_a, f_p\}$ integrates RAFT [45], Depth Anything V2 [61], DINOv2 [40], and D3 [74] to extract forensic cues from motion, depth, appearance, and pixel-level cues. All the tools are optimized on the `GenVideo` training split following the default configurations, while D3 [74] and Depth Anything V2 [61] are adapted and fine-tuned for synthesized video detection. In Eq. 6, we set $\gamma = 0.5$ and $L = 3$.

Table 4: Ablation study of SafeGuard across SafeVid and other existing benchmarks. SHGR: Social Risk-Aware Hypothesis Generation and Routing, **VC**: Video Cropping, **FTE**: Forensic Trace Extraction, **SRV**: Self-Reflective Verifier.

Variant	SafeVid		Existing Benchmarks								Overall	
	Avg.		GenVideo		DVF		LOKI		GenVidBench		Avg.	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Baseline (GPT-4o)	64.6	60.5	77.6	86.0	74.5	75.6	55.8	66.4	61.5	68.8	66.8	71.5
w/ SHGR	66.0	59.1	77.7	87.9	75.0	81.8	59.9	68.4	63.8	71.1	68.5	73.7
w/ SHGR + VC	67.5	71.0	84.3	89.4	80.9	84.2	64.4	72.5	65.0	72.4	72.4	77.9
w/ SHGR + VC + SRV	74.5	71.8	92.3	95.5	91.6	87.9	77.6	74.8	73.1	71.6	81.8	80.3
w/ SHGR + VC + FTE	79.0	76.0	93.0	92.6	94.0	90.5	79.3	70.0	75.4	74.0	84.1	80.6
SafeGuard(Ours)	**85.5**	**80.9**	**97.9**	**95.6**	**97.0**	**93.2**	**81.7**	**77.7**	**84.6**	**85.1**	**89.3**	**86.5**

Table 5: Ablation study of the forensic toolbox across SafeVid and other existing benchmarks. f_o, f_d, f_a, and f_p denote the tools that extract optical flow, depth, appearance, and pixel-level information from the input video, respectively.

Variant	SafeVid		Existing Benchmarks								Overall	
	Avg.		GenVideo		DVF		LOKI		GenVidBench		Avg.	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
w/o tool use	74.5	71.8	92.3	95.5	91.6	87.9	77.6	74.8	73.1	71.6	81.8	80.3
w/ f_o	75.6	74.7	93.1	95.2	93.6	88.1	78.2	70.0	74.6	72.4	83.0	80.1
w/ $f_o + f_d$	78.5	78.4	95.1	95.0	94.0	90.0	78.7	73.9	78.7	76.0	85.0	82.7
w/ $f_o + f_d + f_a$	83.3	79.5	96.2	95.1	96.3	92.8	81.6	75.3	82.5	80.6	88.0	84.7
w/ $f_o + f_d + f_a + f_p$	**85.5**	**80.9**	**97.9**	**95.6**	**97.0**	**93.2**	**81.7**	**77.7**	**84.6**	**85.1**	**89.3**	**86.5**

5.2 Main Comparison

Model Performance on SafeVid As shown in Table 3, task-specific detectors achieve strong ID performance but exhibit severe degradation on OOD splits. Several video-specific and training-free methods perform close to random results, revealing the challenge of extracting reliable cues on SafeVid. Generalist models show modest performance on both splits, reflecting limited adaptation to AI-generated video detection. SafeGuard outperforms all baselines on ID and OOD sets, improving average Acc and F1 by +18.7% over BusterX++ [55] and +14.3% over NPR [44]. To further demonstrate the framework's versatility, we implement Qwen3-VL-32B [3] as an open-source backbone for both the Solver and the Verifier. The consistent performance improvement across proprietary and open-source models validates the general effectiveness of our approach. Notably, SafeGuard achieves performance comparable to human evaluators as reported in "Human" of Table 3; however, the remaining OOD gap underscores the forensic difficulty of OOD videos.

Model Performance on Other Existing Benchmarks To evaluate cross-dataset generalization, we further assess SafeGuard on `GenVideo`, `DVF`, `LOKI`, and `GenVidBench` in Table 3. As the task-specific models within the toolbox are pre-trained on `GenVideo`, the 97.9% accuracy on this dataset serves as an in-domain reference. Without domain-specific fine-tuning, SafeGuard achieves 97.0% on

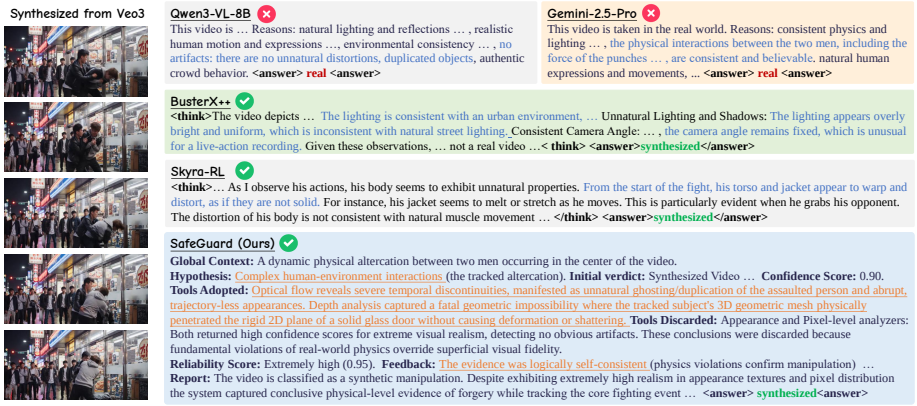

Fig. 4: A case study of SafeGuard and other existing models on SafeVid. The visual anomalies within the video manifest as repeating humanoid outlines during combat and figures that appear abruptly without plausible trajectories.

`DVF`, 81.7% on `LOKI`, and 84.6% on `GenVidBench`, outperforming state-of-the-art methods (DeMamba [6] and BusterX++ [55]) by margins of +6.9%, +11.7%, and +13.2% respectively, demonstrating strong cross-dataset robustness. The performance on `LOKI` and `GenVidBench` highlights the inherent forensic challenges posed by advanced generative models, which produce highly realistic videos devoid of stereotypical artifacts. In contrast, `DVF` comprises videos generated by earlier models, exhibiting salient and easily detectable visual inconsistencies. Under the `LOKI` benchmark, SafeGuard (Qwen3-VL [3]) outperforms SafeGuard (GPT-4o [19]), indicating effective utilization of backbone-specific visual priors under diverse artifact distributions.

5.3 Ablation Study

We conduct ablation studies to quantify the contribution of individual modules within SafeGuard. By default, the Solver utilizes GPT-4o [19] and the Verifier employs Gemini-2.5-Pro [11]. Across all variants, the evaluation protocol remains strictly constant, modifying exclusively the ablated components.

Effect of the Hierarchical Perceptual Solver Table 4 summarizes the impact of the *Hierarchical Perceptual Solver* (HPS), comprising SHGR, VC, and FTE. Progressive integration of these components increases Avg. Acc from 66.8% to 84.1%. Specifically, SHGR yields an initial improvement, VC strengthens hypothesis-guided spatial grounding, and FTE introduces a substantial performance enhancement (+11.7% Avg. Acc), underscoring the critical necessity of explicit forensic evidence. Overall, the HPS converts suspicious-region hypotheses into verifiable, localized traces of potential forgery.

Effect of the Self-Reflective Verifier Table 4 demonstrates the efficacy of the *Self-Reflective Verifier* (SRV), which enhances overall performance by +5.2% Avg. Acc compared to the SHGR+VC+FTE. This consistent improvement across benchmarks indicates that raw forensic evidence extraction remains insufficient for reliable decision-making. Notably, activating the SRV without FTE yields performance degradation (81.8% Avg. Acc), demonstrating that reflective reasoning cannot independently compensate for an absence of explicit forensic evidence. Rather, the SRV functions to enforce hypothesis–evidence consistency and mitigate reasoning discrepancies through closed-loop refinement.

Effect of the Forensic Toolbox Table 4 indicates that integrating the FTE substantially improves performance over reasoning alone (+9.4% Avg. Acc), further underscoring the necessity of grounded forensic evidence. As detailed in Table 5, progressively aggregating task-specific models within the toolbox produces consistent accuracy improvement, culminating in 89.3% Avg. Acc and 86.5% Avg. F1. Specifically, the integration of depth and appearance models contributes significantly to these observed improvements.

5.4 Case Study

As shown in Fig. 4, the generalist models Qwen3-VL [3] and Gemini-2.5-Pro [11] over-rely on the realistic appearance and incorrectly predict *real*. BusterX++ [55] outputs the correct label but provides only a coarse, high-level explanation without identifying concrete forensic traces. Skyra-RL [27] adopts a generic artifact-checklist reasoning approach, which is weakly grounded in the subtle, localized inconsistencies present in this high-fidelity video. In contrast, SafeGuard correctly predicts *synthesized* and delivers a complete, reproducible reasoning pipeline, supplying traceable evidence at each step.

6 Conclusion

To address the scarcity of datasets covering socially sensitive threats and the *Perception–Reasoning Gap* in existing detectors, we introduce SafeVid, a benchmark for high-risk scenarios, and SafeGuard, a multi-agent forensic framework. Our analysis shows that task-specific detectors often fail on out-of-domain data, while general large vision-language models overlook subtle visual artifacts. SafeGuard mitigates these issues by combining forensic tools with semantic reasoning, providing a transparent, verifiable chain of evidence and robust support for high-stakes, socially sensitive forensic scenarios.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant U22A2094 and Grant 62272435, and also supported by the Yangtze River Delta Science and Technology Innovation Community Joint Research (Basic Research) Project under Grant 2025CSJZN01600.

References

1. Asad, A., Obadinma, S., Shayanfar, R., Zhu, X.: Reddebate: Safer responses through multi-agent red teaming debates. In: International Conference on Machine Learning (2026)
2. Bai, J., et al.: Ai-generated video detection via spatial-temporal anomaly learning. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 460–470 (2024)
3. Bai, S., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bai, Z., et al.: Hallucination of multimodal large language models: A survey. arXiv preprint arXiv:2404.18930 (2024)
5. Blattmann, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
6. Chen, H., et al.: Demamba: Ai-generated video detection on million-scale genvideo benchmark. Science China Information Sciences p. 162103 (2026)
7. Chen, W., et al.: Genworld: Towards detecting ai-generated real-world simulation videos. arXiv preprint arXiv:2506.10975 (2025)
8. Cheng, Z., et al.: Moca: Modeling object consistency for 3d camera control in video generation. In: International Conference on Learning Representations (2026)
9. Chinchure, A., et al.: Spotlight: Identifying and localizing video generation errors using vlms. arXiv preprint arXiv:2511.18102 (2025)
10. Coccomini, D.A., et al.: Mintime: Multi-identity size-invariant video deepfake detection. IEEE Transactions on Information Forensics and Security pp. 6084–6096 (2024)
11. Comanici, G., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
12. Corvi, R., et al.: Seeing what matters: Generalizable ai-generated video detection with forensic-oriented augmentation. Advances in Neural Information Processing Systems pp. 26418–26446 (2025)
13. Feng, K., et al.: Video-r1: Reinforcing video reasoning in mllms. Advances in Neural Information Processing Systems pp. 99114–99137 (2025)
14. Gao, Y., et al.: David-xr1: Detecting ai-generated videos with explainable reasoning. arXiv preprint arXiv:2506.14827 (2025)
15. Google DeepMind: Veo 3 Model Card. Model card, Google DeepMind (2025), `https://storage.googleapis.com/deepmind-media/Model-Cards/Veo-3-Model-Card.pdf`
16. Hong, S., et al.: Metagpt: Meta programming for a multi-agent collaborative framework. In: International Conference on Learning Representations. pp. 23247–23275 (2024)
17. Huang, B., et al.: Implicit identity driven deepfake face swapping detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4490–4499 (2023)
18. Huang, Q., et al.: Unishield: An adaptive multi-agent framework for unified forgery image detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8121–8132 (2026)
19. Hurst, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
20. Internò, C., et al.: Ai-generated video detection via perceptual straightening. Advances in Neural Information Processing Systems pp. 20672–20705 (2025)

21. Ji, L., et al.: Distinguish any fake videos: Unleashing the power of large-scale data and motion features. arXiv preprint arXiv:2405.15343 (2024)
22. Jiang, Z., et al.: Vace: All-in-one video creation and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17191–17202 (2025)
23. Kong, W., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
24. Lai, Y., et al.: Agent4faceforgery: Multi-agent llm framework for realistic face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14073–14083 (2026)
25. Li, X., et al.: Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. arXiv preprint arXiv:2504.06958 (2025)
26. Li, Y., et al.: Unim: A unified any-to-any interleaved multimodal benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15902–15911 (2026)
27. Li, Y., et al.: Skyra: Ai-generated video detection via grounded artifact reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4482–4493 (2026)
28. Liang, M., et al.: From evidence to verdict: An agent-based forensic framework for ai-generated image detection. arXiv preprint arXiv:2511.00181 (2025)
29. Lin, L., et al.: Detecting multimedia generated by large ai models: A survey. arXiv preprint arXiv:2402.00045 (2024)
30. Ling, H., et al.: Everybodydance: Bipartite graph–based identity correspondence for multi-character animation. Advances in Neural Information Processing Systems pp. 14123–14151 (2025)
31. Liu, Q., et al.: Lavid: An agentic lvlm framework for diffusion-generated video detection. arXiv preprint arXiv:2502.14994 (2025)
32. Liu, S., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European Conference on Computer Vision. pp. 38–55 (2024)
33. Liu, X., et al.: Video-safetybench: A benchmark for safety evaluation of video lvlms. Advances in Neural Information Processing Systems (2025)
34. Ma, L., et al.: Detecting ai-generated video via frame consistency. In: 2025 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6 (2025)
35. Mellis, I.: Pika. `https://pika.art/` (2025), 2025-02-06
36. Nan, K., et al.: Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In: International Conference on Learning Representations. pp. 1045–1064 (2025)
37. Ni, B., et al.: Expanding language-image pretrained models for general video recognition. In: European Conference on Computer Vision. pp. 1–18 (2022)
38. Ni, Z., et al.: Genvidbench: A 6-million benchmark for ai-generated video detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 15582–15590 (2026)
39. OpenAI: Sora system card. `https://openai.com/index/sora-system-card/` (2024), accessed: 2025-08-14
40. Oquab, M., et al.: Dinov2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
41. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763 (2021)
42. Ravi, N., et al.: Sam 2: Segment anything in images and videos. In: International Conference on Learning Representations. pp. 28085–28128 (2025)

43. Song, X., et al.: On learning multi-modal forgery representation for diffusion generated video detection. Advances in Neural Information Processing Systems pp. 122054–122077 (2024)
44. Tan, C., et al.: 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. pp. 28130–28139 (2024)
45. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European Conference on Computer Vision. pp. 402–419 (2020)
46. Wan, T., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
47. Wang, Q., et al.: Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. Advances in Neural Information Processing Systems pp. 4350–4376 (2025)
48. Wang, W., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
49. Wang, W., Yang, Y.: Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. Advances in Neural Information Processing Systems pp. 65618–65642 (2024)
50. Wang, W., Yang, Y.: Tip-i2v: A million-scale real text and image prompt dataset for image-to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14898–14908 (2025)
51. Wang, W., Yang, Y.: Videoufo: A million-scale user-focused dataset for text-to-video generation. Advances in Neural Information Processing Systems (2025)
52. Wang, W., et al.: Scaling laws for deepfake detection. arXiv preprint arXiv:2510.16320 (2025)
53. Wang, X., et al.: Affordbot: 3d fine-grained embodied reasoning via multimodal large language models. Advances in Neural Information Processing Systems pp. 140367–140387 (2025)
54. Wen, H., et al.: Busterx: Mllm-powered ai-generated video forgery detection and explanation. arXiv preprint arXiv:2505.12620 (2025)
55. Wen, H., et al.: Busterx++: Towards unified cross-modal ai-generated content detection and explanation with mllm. arXiv preprint arXiv:2507.14632 (2025)
56. Wu, C., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
57. Xiao, J., et al.: Next-qa: Next phase of question-answering to explaining temporal actions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9777–9786 (2021)
58. Xu, Y., et al.: Self-evolving agents as dynamic graph transformation: A survey and new perspective. ResearchGate preprint (2026)
59. Xu, Y., et al.: Tall: Thumbnail layout for deepfake video detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22658–22668 (2023)
60. Xue, Z., et al.: Ultravideo: High-quality UHD video dataset with comprehensive captions. In: Advances in Neural Information Processing Systems (2025)
61. Yang, L., et al.: Depth anything v2. Advances in Neural Information Processing Systems pp. 21875–21911 (2024)
62. Yang, S., et al.: Towards robust detection of chinese toxic variants via dynamic knowledge graph–llm reasoning. In: Proceedings of the ACM Web Conference 2026. pp. 3183–3194 (2026)

63. Yang, X., et al.: Deconfounded video moment retrieval with causal intervention. In: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. pp. 1–10 (2021)
64. Yang, X., et al.: Video moment retrieval with cross-modal neural architecture search. IEEE Transactions on Image Processing pp. 1204–1216 (2022)
65. Yang, X., et al.: Fine: Fine-grained neuron-level model editing for reliable and safe llms. IEEE Transactions on Pattern Analysis and Machine Intelligence (2026)
66. Yang, Z., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: International Conference on Learning Representations. pp. 83048–83077 (2025)
67. Yao, X., et al.: A latent transformer for disentangled face editing in images and videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13789–13798 (2021)
68. Ye, J., et al.: Loki: A comprehensive synthetic data detection benchmark using large multimodal models. International Conference on Learning Representations (2025)
69. YouTube LLC: Youtube. `https://www.youtube.com/` (2025), accessed: 2025-08-14
70. Zhang, B., et al.: Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106 (2025)
71. Zhang, S., et al.: Physics-driven spatiotemporal modeling for ai-generated video detection. Advances in Neural Information Processing Systems pp. 174683–174730 (2025)
72. Zhang, Y., et al.: Llava-next: A strong zero-shot video understanding model (April 2024), `https://llava-vl.github.io/blog/2024-04-30-llava-next-video/`
73. Zhao, H., et al.: Qwen3guard technical report. arXiv preprint arXiv:2510.14276 (2025)
74. Zheng, C., et al.: D3: Training-free ai-generated video detection using second-order features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12852–12862 (2025)
75. Zheng, Y., et al.: Exploring temporal coherence for more general video face forgery detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15044–15054 (2021)
76. Zhou, S., et al.: Egotextvqa: Towards egocentric scene-text aware video question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3363–3373 (2025)
77. Zhou, S., et al.: Scene-text grounding for text-based video question answering. IEEE Transactions on Multimedia pp. 1417–1430 (2025)