Check for updates

ACM DIGITAL LIBRARY | Association for Computing Machinery | acm open

Latest updates: https://dl.acm.org/doi/10.1145/3746027.3758295

RESEARCH-ARTICLE

AEGIS: Authenticity Evaluation Benchmark for AI-Generated Video Sequences

JIEYU LI, National University of Singapore, Singapore City, Singapore

XIN ZHANG, Agency for Science, Technology and Research, Singapore, Singapore City, Singapore

JOEY TIANYI ZHOU, Agency for Science, Technology and Research, Singapore, Singapore City, Singapore

Open Access Support provided by:

National University of Singapore

Agency for Science, Technology and Research, Singapore

PDF Download
3746027.3758295.pdf
02 April 2026
Total Citations: 0
Total Downloads: 237

Published: 27 October 2025

Citation in BibTeX format

MM '25: The 33rd ACM International Conference on Multimedia
October 27 - 31, 2025
Dublin, Ireland

Conference Sponsors:
SIGMM

https://doi.org/10.1145/3746027.3758295
ISBN: 9798400720352

AEGIS: Authenticity Evaluation Benchmark for AI-Generated Video Sequences

Jieyu Li
National University of Singapore
Singapore
jieyuli@u.nus.edu

Xin Zhang
Centre for Frontier AI Research, Institute of High Performance Computing
Agency for Science, Technology and Research
Singapore
zhangx7@cfar.a-star.edu.sg

Joey Tianyi Zhou✉
Centre for Frontier AI Research, Institute of High Performance Computing
Agency for Science, Technology and Research
Singapore
joey_zhou@cfar.a-star.edu.sg

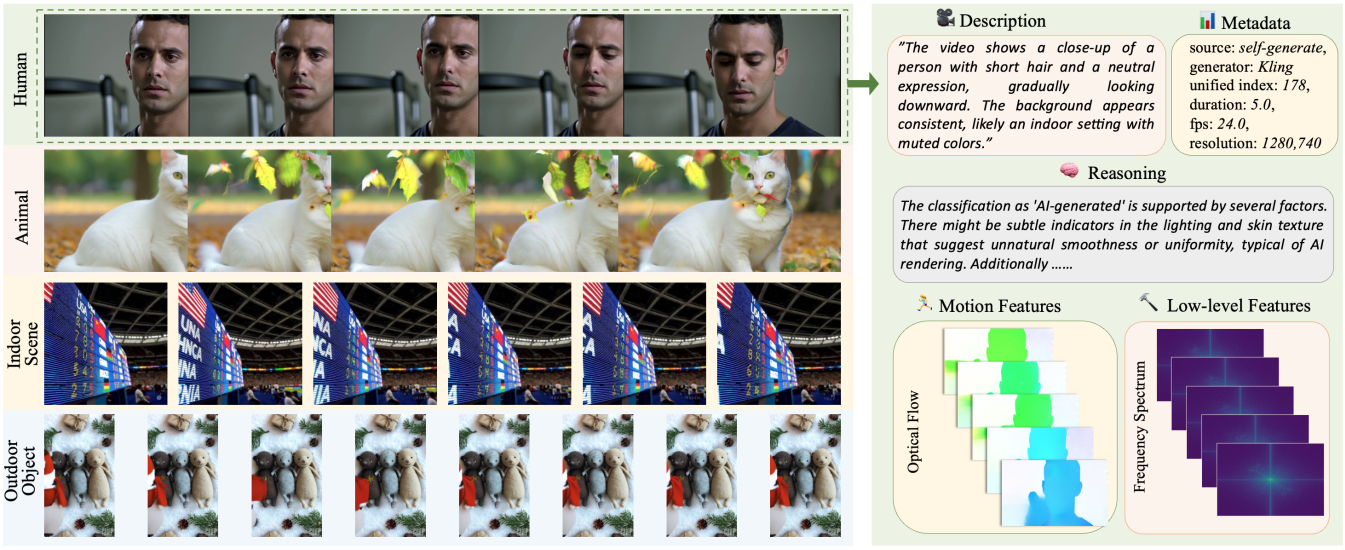

Figure 1: Overview of the AEGIS dataset. AEGIS comprises a diverse collection of synthetic videos covering a broad range of realistic scenarios, including detailed human facial expressions, natural outdoor animal behaviors, indoor public environments, and intricate static object arrangements. Each video is accompanied by rich multimodal annotations, including *Semantic-Authenticity Descriptions* (semantic summaries, generation metadata, and reasoning explanations), *Motion Features*, and *Low-level Visual Features*. These annotations enable robust and explainable analysis, supporting research in video authenticity detection as well as a variety of downstream multimodal understanding tasks.

Abstract

Recent advances in AI-generated content have fueled the rise of highly realistic synthetic videos, posing severe risks to societal trust and digital integrity. Existing benchmarks for video authenticity detection typically suffer from limited realism, insufficient scale, and inadequate complexity, failing to effectively evaluate modern vision-language models against sophisticated forgeries. To address this critical gap, we introduce **AEGIS**, a novel large-scale benchmark explicitly targeting the detection of ***hyper-realistic*** and ***semantically nuanced*** AI-generated videos. AEGIS comprises over 10,000 rigorously curated real and synthetic videos generated by diverse, state-of-the-art generative models, including Stable Video Diffusion, CogVideoX-5B, KLing, and Sora, encompassing open-source and proprietary architectures. In particular, AEGIS features specially constructed challenging subsets enhanced with GPT-4o-refined prompts, creating unprecedentedly realistic scenarios for rigorous robustness evaluation. Furthermore, we provide multimodal annotations spanning ***Semantic-Authenticity Descriptions***, ***Motion Features***, and ***Low-level Visual Features***, facilitating authenticity detection and supporting downstream tasks such as multimodal fusion and forgery localization. Extensive experiments using advanced vision-language models demonstrate limited detection capabilities on the most challenging subsets of AEGIS, highlighting

This work is completed during Jieyu Li's internship at A*STAR.
✉Corresponding author: Joey Tianyi Zhou.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

MM '25, Dublin, Ireland

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3758295

the dataset's unique complexity and realism beyond the current generalization capabilities of existing models. In essence, AEGIS establishes an indispensable evaluation benchmark, fundamentally advancing research toward developing ***genuinely robust***, ***reliable***, and ***broadly generalizable*** video authenticity detection methodologies capable of addressing real-world forgery threats. Our dataset is avaliable on https://huggingface.co/datasets/Clarifiedfish/AEGIS.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; **Computer vision**; **Machine learning**.

Keywords

Authenticity Evaluation Benchmark, AI-Generated Video Sequences

ACM Reference Format:

Jieyu Li, Xin Zhang, and Joey Tianyi Zhou✉. 2025. AEGIS: Authenticity Evaluation Benchmark for AI-Generated Video Sequences . In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3746027.3758295

1 Introduction

Recent advances in AI-generated content (AIGC) technologies have substantially simplified the creation of highly realistic video content [24, 25, 27]. These sophisticated generative models have significantly reduced production costs and fostered novel applications in various fields, such as education and entertainment [46]. However, this rapid proliferation also poses substantial social risks [15, 16]. Compared to synthetic images, the higher perceptual realism and temporal consistency inherent in generated videos exacerbate their potential for misinformation dissemination [38], erosion of public trust, and threats to information security across social and professional platforms. Consequently, there is an urgent need for robust and reliable detection methods capable of effectively distinguishing synthetic from authentic videos.

However, the advancement of robust video forgery detection methods critically depends on suitable benchmarks. Although several AIGC datasets have recently been proposed for forgery detection, they primarily target static images [20, 21, 39], and thus inherently fail to capture video-specific challenges such as temporal coherence, realistic motion dynamics, and semantic consistency across frames. Recent video benchmarks, such as VBench [14], EvalCrafter [26], and AIGCBench [8], primarily focus on evaluating generation quality or perceptual fidelity rather than explicitly targeting authenticity detection tasks. Additionally, dedicated video forgery datasets like GenVidBench [29] and DeMamba [6] also possess certain constraints, including simple animation-oriented content, narrow generative diversity, and insufficient emphasis on realism and detection complexity, thereby restricting their capability to comprehensively assess advanced detection algorithms.

To effectively overcome these critical limitations, we propose **AEGIS**, **A**uthenticity **E**valuation Benchmark for AI-**G**enerated v**I**deo **S**equences, a novel video authenticity benchmark meticulously designed to challenge and advance current detection capabilities against highly deceptive AI-generated content. AEGIS sets itself apart by exclusively assembling 5,199 synthetic videos derived from seven cutting-edge generative techniques, including prominent open-source methods such as Stable Video Diffusion [3], CogVideoX-5B [44], and I2VGen-XL [47], as well as proprietary commercial systems represented by KLing [19], Sora [32], and Pika [33]. The distinctive integration of these diverse generative techniques ensures unmatched realism, complexity, and representation of current generation paradigms, positioning AEGIS as an indispensable resource for rigorously evaluating and substantially improving model robustness in the face of emerging, highly realistic forgery threats.

To rigorously benchmark and substantially advance video forgery detection, AEGIS introduces several critical innovations that clearly distinguish it from existing benchmarks. First, it includes carefully crafted challenging subsets refined via GPT-4o [31]-generated prompts, explicitly designed to intensively evaluate model robustness against highly sophisticated and semantically nuanced forgeries. These synthetic videos are complemented by systematically curated authentic videos characterized by significant visual complexity and diversity, creating realistic evaluation conditions. Second, AEGIS provides extensive multimodal annotations, including optical flow, frequency-domain analysis, and rich semantic descriptions, to facilitate rigorous evaluation and support diverse downstream forensic tasks. Crucially, our extensive experiments using SOTA vision-language models, such as Qwen-VL [1] and Video-LLaVA [23], across zero-shot, prompt-guided, and fine-tuned scenarios, reveal notable performance gaps, especially within challenging subsets. These findings underscore both the substantial difficulty presented by AEGIS and its effectiveness to expose critical generalization limitations inherent in current detection methodologies. Consequently, AEGIS emerges as an indispensable benchmark, uniquely positioned to drive forward the development of more robust and widely generalizable video authenticity detection models.

The key contributions of this work include:

- We propose **AEGIS**, a novel large-scale benchmark for video authenticity detection, comprising 5,199 synthetic videos generated by six diverse SOTA techniques, covering both open-source and proprietary models. AEGIS significantly improves upon existing benchmarks in diversity, realism, and semantic complexity.
- We design challenging evaluation subsets using GPT-4o-refined prompts to simulate highly realistic, semantically nuanced scenarios. The resulting *Hard Test Sets* effectively expose generalization gaps in current detection models.
- We curate authentic videos with rich multimodal annotations, including semantic descriptions, optical flow, frequency-domain features, and temporal coherence metrics. Experiments with advanced vision-language models reveal clear performance limitations, underscoring AEGIS's value for robust and generalizable forgery detection.

2 Related Work

2.1 Image-level AIGC Benchmark

Synthetic image generation has significantly advanced due to Generative Adversarial Networks (GANs) [10], diffusion models [7],

Table 1: Overview of the Collected Dataset

Split	Total number	Category	Number	Source	Duration (s)	Number	Frame rate (fps)	Resolution
Traing Set	7304	Synthetic	3161	TIP-I2V (Pika)	3-7	756	24	380p-1080p
				TIP-I2V (CogVideoX-5B)		886	8	
				TIP-I2V (Stable Video Diffusion)		621	7	
				TIP-I2V (I2VGen-XL)		898	7	
		Real	4143	Vript (YouTube)	5-10	4060	24-30	
				Vript (TikTok)		283		
Validation Set	2730	Synthetic	1820	TIP-I2V (Pika)	3-7	455	24	
				TIP-I2V (CogVideoX-5B)		455	8	
				TIP-I2V (Stable Video Diffusion)		455	7	
				TIP-I2V (I2VGen-XL)		455	7	
		Real	910	Vript (YouTube)	5-10	455	24-30	
				Vript (TikTok)		455		
Hard Test Set	436	Synthetic	218	Sora	5s	107	30	270p-4K
				KLing		111	24	
		Real	218	DVF	2.4-10	109	24-30	
				self-collected (YouTube)		109		

and flow-matching techniques [9, 18]. Early models such as StyleGAN [17] notably enhanced facial realism, while recent diffusion-based and transformer-based approaches, including Stable Diffusion and DALLE-2, significantly expanded general-purpose image synthesis [22, 50]. Correspondingly, several benchmarks, including AIGCIQA2023 [39], AGIQA-20K [20], PKU-AIGIQA-4K [45], and FragFake [36], emerged to rigorously evaluate image generation quality, focusing primarily on perceptual fidelity, semantic alignment, and fine-grained detection tasks. Despite these advancements, these image-level datasets inherently lack consideration of video-specific challenges such as temporal coherence and realistic motion patterns. Our proposed AEGIS explicitly addresses these critical video-centric issues by integrating temporal and multimodal analysis, significantly extending beyond static imagery benchmarks.

2.2 Video-level AIGC Benchmarks

Recent video-level AI-generated content (AIGC) benchmarks such as VBench [14], T2VSafetyBench [28], EvalCrafter [26], and VIDEOPHY [2], primarily focus on assessing video generation quality rather than explicitly addressing authenticity detection tasks. Meanwhile, existing datasets specifically targeting video forgery detection, such as DF40 [42], Deepfake-Eval-2024 [4], and ExDDV [12], predominantly emphasize facial manipulation and in-the-wild deepfakes, limiting their broader generalization. Recent large-scale benchmarks like GenVidBench [29] and DeMamba [6], while including diverse generative sources, often incorporate less challenging animation-style videos or emphasize dataset scale over detection complexity. In contrast, our proposed AEGIS benchmark explicitly prioritizes video authenticity detection by focusing solely on hyper-realistic, semantically complex AI-generated videos, deliberately excluding easily identifiable animation-oriented content. Unlike prior datasets, AEGIS provides detailed multimodal annotations and includes specially constructed challenging subsets enhanced by GPT-4o-refined prompts, explicitly designed to rigorously evaluate the robustness and generalization capabilities of detection methods.

3 Dataset Construction

To advance research in video authenticity detection, we construct AEGIS, a comprehensive dataset consisting of both AI-generated and real-world videos. This section details our systematic construction pipeline, as illustrated in Figure 2, which comprises three main stages: data collection, data filtering, and data splitting. An overview of the final structure of the AEGIS dataset is provided in Table 1.

3.1 Data Collection

3.1.1 Real Video Collection. To ensure high realism, diversity, and visual complexity, we collect authentic videos from three sources: (1) **Vript Dataset** [43]: We utilize approximately 12,000 annotated videos sourced from YouTube (horizontal, long-form) and TikTok (vertical, short-form). This cross-platform selection captures inherent content and stylistic biases associated with different formats, thus preserving the diversity characteristic of real-world videos. (2) **DVF Dataset** [35]: Specifically selected for its high visual complexity and realism, DVF consists of diverse human-recorded video clips that capture subtle details and closely mimic the complexities of genuine human-generated content. (3) **Supplemental YouTube Collection**: To further augment the realism and practical applicability of our dataset, we independently collected minimally edited videos from YouTube, including raw street interviews, wildlife documentaries, and authentic vlogs. We systematically applied 30 predefined search queries designed to maximize diversity and ensure authenticity. Each collected video was rigorously verified through manual inspection and standardized by trimming clips to durations ranging from 2.4 to 10 seconds, removing audio tracks, and maintaining a resolution diversity spanning from 360p to 4K. This careful standardization process significantly improves the dataset's realism, representativeness, and complexity, effectively facilitating robust model evaluation.

3.1.2 Synthetic Video Collection. Our synthetic subset rigorously integrates publicly available advanced datasets and independently generated synthetic content, employing SOTA generative models to ensure both diversity and realistic complexity. (1) **TIP-I2V Dataset** [41] [2]: This dataset provides 500,000 synthetic video clips, generated from 100,000 prompts using five SOTA video generation models: Stable Video Diffusion [3], CogVideoX-5B [44], I2VGen-XL [47], Open-Sora [49], and Pika [33]. (2) **Proprietary Model Generation via KLing and Sora**: To further evaluate and enhance model robustness under challenging scenarios, we independently generated synthetic videos utilizing proprietary, SOTA generative

[1]Parts of the video frame examples in Figure 1 and Figure 2 are sourced from [41] and [43].
[2]We used the official curated subset of the original TIP-I2V dataset, available at https://huggingface.co/datasets/WenhaoWang/TIP-I2V/tree/main/subset_videos_tar

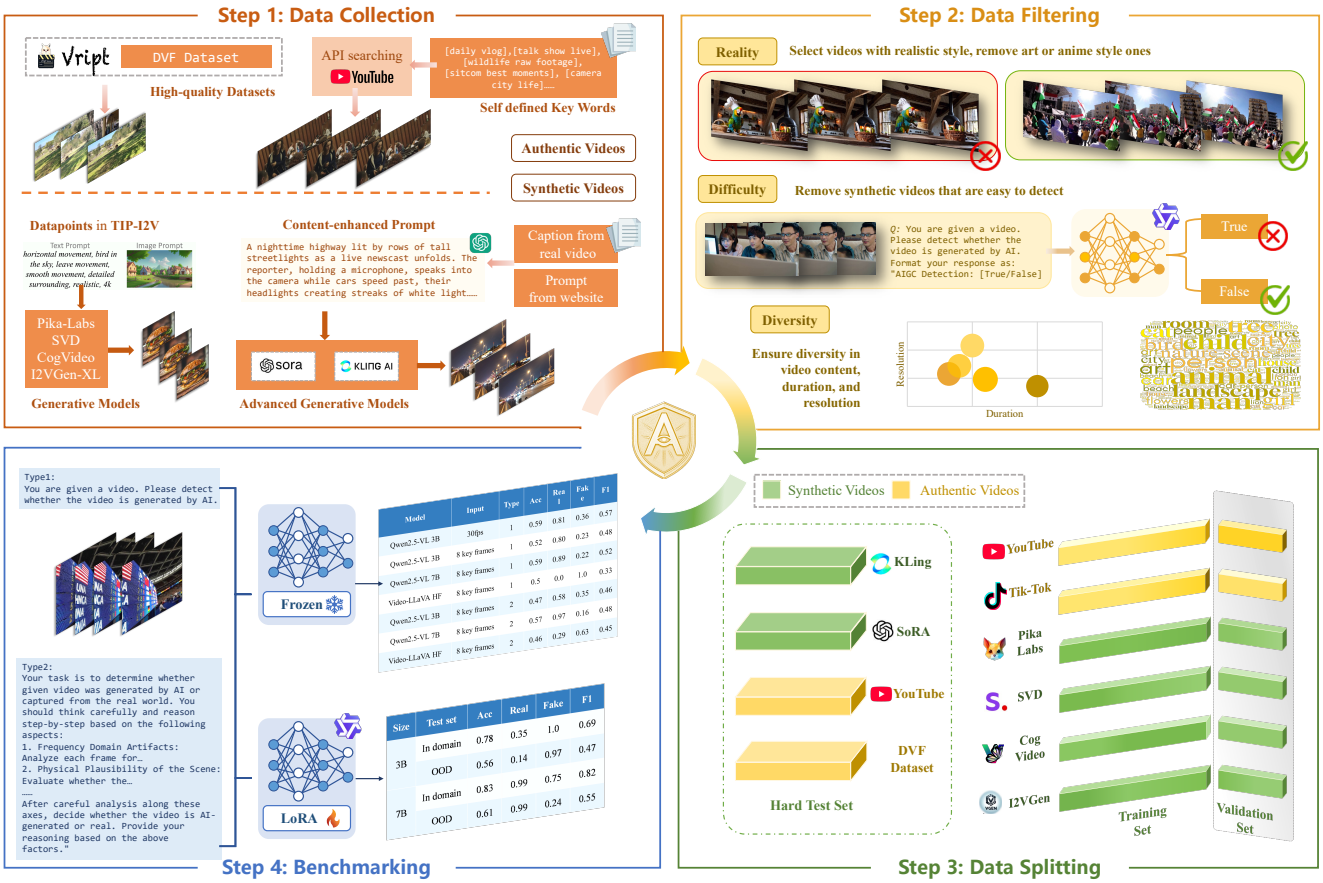

Figure 2: The AEGIS Dataset Construction Pipeline. Step 1: Data Collection – Collecting real and synthetic videos from diverse sources. Step 2: Data Filtering – Applying three key principles: reality (removing non-photorealistic content), difficulty (discarding easily detectable fakes), and diversity (ensuring variation in content, resolution, and duration). Step 3: Data Splitting – Creating balanced training, validation, and hard test sets. Step 4: Benchmarking – Evaluating vision-language models under different settings including zero-shot inference, structured reasoning prompt, and low-rank adaptation (LoRA) fine-tuning.[1]

models, namely KLing [19] and Sora [32]. High-quality textual prompts were meticulously sourced from the extensive HD-VG-130M dataset [40], and carefully crafted with reference to exemplary demonstrations provided by the official KLing [19] and Sora [32] showcases. Each prompt underwent an additional refinement process leveraging GPT-4o [31], ensuring enhanced detail, semantic precision, and realism. We systematically generated a balanced collection of 218 synthetic videos, carefully controlling for diversity in visual content, resolutions ranging from 360p to 1080p, and durations varying from 5 to 10 seconds.

3.2 Data Filtering

To support robust training and evaluation of authenticity detection models, we propose a unified filtering framework grounded in three principles, ***Reality***, ***Difficulty***, and ***Diversity***, applied to both real and synthetic subsets. The framework ensures that real-world samples exhibit high-fidelity, unaltered content, while synthetic samples are photorealistic yet non-trivial to detect. It further promotes content diversity critical for generalization across visually complex scenarios.

(1) ***Reality.*** We ensured that all included videos—whether real or synthetic—exhibit photorealistic styles by removing art-style, low-quality, or heavily edited content. For the real videos, since datasets like Vript [43] and the Supplemental YouTube Collection contain web-sourced material, many clips were either AI-generated or excessively edited. To address this, we employed Qwen2.5-VL [1] to classify approximately 9,000 clips into camera-shot (authentic), heavily edited, and AI-generated categories, discarding around 4,000 non-authentic samples. For the DVF dataset, we performed additional manual reviews to ensure visual quality and authenticity. For the synthetic videos from the TIP-I2V dataset [41], we first selected high-quality, photorealistic prompts from the 100,000 officially provided. Prompts with abstract, implausible, or stylistically inconsistent content were excluded. We then verified that the described actions were visually feasible and semantically coherent, filtering out physically unrealistic or static scenarios. After this

two-stage filtering process, 17,000 prompts were retained. For each prompt, we randomly selected one generated video from among five aforementioned candidate models. Due to the relatively low visual quality of Open-Sora's outputs [49], its generated videos were further removed from our subset. In addition, we applied manual filtering to ensure the visual realism of videos generated by KLing [19] and Sora [32].

(2) ***Difficulty.*** After the Reality filtering, the remaining synthetic videos from the TIP-I2V dataset underwent an additional round of screening using zero-shot classification with the Qwen2.5-VL model [1]. This step was designed to further eliminate samples that exhibit obvious signs of synthetic generation. Specifically, we used Qwen2.5-VL to classify each video as either "AI-generated" (True) or "not AI-generated" (False). Videos that were confidently predicted as "AI-generated" were discarded. In contrast, videos labeled as "not AI-generated" with lower model confidence were retained. This procedure resulted in a curated subset of approximately 5,000 synthetic clips, which exhibit greater visual realism and are less likely to be trivially detected as generated. This additional refinement step ensures that the synthetic data used in our evaluation is both challenging and of high perceptual quality.

(3) ***Diversity.*** We ensured diversity in both real and synthetic videos across scene types, subjects, and visual conditions. Real clips span indoor and outdoor environments, human, animal, and object subjects, as well as urban and rural scenes, with durations ranging from 2.4 to 10 seconds and resolutions from 360p to 4K. Synthetic videos were generated from diverse prompts using four SOTA generators, introducing variations in motion, actors, backgrounds, and lighting conditions. Scene tag distributions (visualized via a word cloud) and resolution-duration scatter plots confirm broad coverage across both semantic and visual dimensions.

Overall, our comprehensive filtering pipeline ensures that the resulting dataset is both high-quality and representative, enabling robust and realistic evaluation of video authenticity detection methods. After applying all filtering procedures, the finalized AEGIS dataset comprises approximately 5,199 synthetic and 5,271 authentic videos, systematically curated to support reliable benchmarking across diverse and challenging scenarios.

3.3 Data Splitting

To effectively benchmark models under realistic deployment scenarios and rigorously evaluate their generalization capabilities, we systematically divided the filtered AEGIS dataset into three subsets: Training Set, Validation Set, and Hard Test Set. The Training and Validation Sets primarily include filtered authentic videos from Vript dataset [43] and high-quality synthetic videos from TIP-I2V dataset [41]. The Training Set facilitates the learning of discriminative features that distinguish authentic from synthetic videos, while the Validation Set supports hyperparameter tuning and preliminary model evaluation. The Hard Test Set specifically evaluates model robustness and generalization under more challenging conditions. It comprises diverse authentic videos from DVF dataset [35] and Supplemental YouTube Collection, and advanced synthetic videos generated by proprietary models KLing [19] and Sora [32]. Selected for complexity and subtlety, these samples provide a critical benchmark for assessing models' capabilities in realistic scenarios involving sophisticated forgeries and nuanced visual details.

3.4 Multimodal Annotations

Effectively distinguishing AI-generated videos from authentic ones requires capturing and representing complementary visual cues across multiple dimensions, as emphasized in recent studies [5, 11]. To support this goal, AEGIS provides rich multimodal annotations for each video, covering, ***Semantic-Authenticity Descriptions***, ***Motion Features***, and ***Low-level Visual Features***.

(1) ***Semantic-Authenticity Descriptions.*** To capture both high-level semantics and authenticity-related cues, we provide two types of textual descriptions for every video: semantic descriptions and authenticity reasoning descriptions. For synthetic videos, we directly use the original prompts from the TIP-I2V dataset as semantic descriptions, which specify the intended scene, objects, and actions. For real videos, where no prompts are available, we extract frame-level embeddings using CLIP [34] and apply k-means clustering ($k = 8$) to identify representative key frames. We then query GPT-4V [30] to generate semantic descriptions summarizing the content of these key frames. In addition, for both real and synthetic videos, we provide authenticity reasoning descriptions. For each video, we inform GPT-4V of its ground-truth label (real or AI-generated), and prompt it to explain the reasoning behind the label based solely on the visual content. These explanations may highlight temporal smoothness, lighting consistency, or the presence of visual artifacts, offering human-interpretable insights into authenticity cues.

(2) ***Motion Features.*** Realistic motion tends to be temporally smooth and physically coherent, whereas synthetic videos often exhibit subtle artifacts or violations of natural dynamics. To capture such motion inconsistencies, we extract dense optical flow fields using the RAFT algorithm [37], enabling fine-grained characterization of frame-to-frame motion patterns.

(3) ***Low-level Visual Features.*** Low-level vision features address subtle yet revealing pixel-level and frequency-domain discrepancies, such as edge sharpness, compression artifacts, overly smooth textures or repetitive patterns and dynamic range variations. we compute the 2D Fast Fourier Transform (FFT) of each grayscale key frame and apply Radial Integral Operations (RIO) to summarize frequency energy across orientations.

3.5 Distinctive Contributions of AEGIS

The proposed **AEGIS** dataset advances video authenticity detection by explicitly addressing the challenges posed by hyper-realistic AI-generated videos that closely resemble authentic human-created content. Unlike existing benchmarks that often include stylized animations or trivially detectable scenarios, AEGIS focuses exclusively on visually nuanced and contextually rich videos, deliberately curated to reflect the complexities of real-world detection tasks.

Furthermore, AEGIS leverages GPT-4o-refined prompts and state-of-the-art proprietary generative models—such as KLing and Sora—to synthesize highly realistic and deceptive forgeries. These, combined with carefully selected authentic samples, form the *Hard Test Set*, a rigorous benchmark designed to evaluate model robustness and generalization under challenging, real-world conditions.

In addition, AEGIS provides ready-to-use multimodal visual cues to support downstream tasks in synthetic video detection and interpretable reasoning, facilitating deeper insight into model behavior and failure cases.

4 Benchmarking on AEGIS

In this section, we design evaluation strategies to benchmark the authenticity detection performance of SOTA vision-language models on our AEGIS dataset.

4.1 Benchmarking Setup

We evaluate two subsets. (i) An in-domain test set: a subset randomly sampled from the validation set, sharing the same distribution as the training data; (ii) The Hard Test Set (see Sec. 3.3): an out-of-domain split explicitly designed to assess model robustness and generalizability on challenging synthetic videos.

Baseline Models. We evaluate two SOTA vision-language models on the AEGIS dataset: Qwen2.5-VL [1] and Video-LLaVA [23]. Qwen2.5-VL is a strong general-purpose model with robust multimodal comprehension and competitive performance on video-centric tasks; we consider both its 3B and 7B variants. Video-LLaVA is a representative auto-regressive transformer model designed to unify image and video understanding within a single framework.

4.2 Benchmarking Strategies

To systematically examine model performance under different levels of task conditioning, we implement three evaluation strategies: (i) ***Zero-shot Inference***, (ii) ***Structured Reasoning Prompt***, and (iii) ***Low-Rank Adaptation (LoRA) [13] fine-tuning***. To leverage extracted multimodal cues during inference, pre-extracted key frames are fed to the vision-language models using the <image> token format supported by the model interface.

(1) ***Zero-shot Inference.*** We employ a minimal prompt to solicit a binary decision from the model: *"You are an expert in AI-generated-content (AIGC) detection. Given a video, determine whether it is real or AI-generated."* This setup evaluates the model's default capability to perform authenticity detection given only a task description.

(2) ***Structured Reasoning Prompt.*** We construct a multi-step prompt that guides the model through a detailed reasoning process over several visual dimensions, including frequency artifacts, lighting consistency, compression noise, and physical plausibility. Our reasoning-enhanced prompt template is provided in the supplementary link.

(3) ***LoRA Fine-tuning.*** To explore task-specific adaptation, we fine-tune Qwen2.5-VL [1] on the training set using LoRA [13] (learning rate $1e^{-4}$, rank 8, 3 epochs). We utilize the widely-used framework llama-factory [48] for effective training. This setting serves to quantify potential gains from lightweight supervision and evaluate model generalization beyond training distribution.

For each setting, we report four metrics: Acc_{all}: overall classification accuracy (from 0 to 1). Acc_{real}: accuracy on authentic videos;. Acc_{ai}: ccuracy on synthetic videos. Macro-$F1$: Unweighted average F1 score across the two classes.

4.3 Benchmarking Results

Experiments conducted on the AEGIS dataset serve two purposes: (i) to show that current VLMs struggle with video authenticity evaluation on AEGIS, and (ii) to test whether additional training on AEGIS improves their performance.

AEGIS Reveals Gaps in VLMs Zero-Shot Detection. As shown in Table 2a, SOTA models like Qwen2.5-VL [1] achieve low synthetic video detection accuracy (Acc_{ai} from 0.22 to 0.23) on the AEGIS Hard Test Set under zero-shot settings. This highlights the substantial gap between existing model capabilities and the high visual fidelity of AEGIS samples. Furthermore, prompt-based reasoning offers little improvement. As illustrated in Table 2b, accuracy further drops from 0.22 to 0.16 when applying direct textual prompts to Qwen2.5-VL 7B. This unexpected performance indicates that conventional prompting strategies fail to capture the nuanced visual and semantic cues characteristic of high-quality forgeries in AEGIS.

Table 2: Detection Accuracy on Hard Test Set

(a) Zero-shot Inference

Model	Acc_{all}	Acc_{real}	Acc_{ai}	Macro F1
Qwen2.5-VL 3B	0.52	0.80	0.23	0.48
Qwen2.5-VL 7B	0.59	0.89	0.22	0.52
Video-LLaVA-HF 7B	0.5	0.0	1.0	0.33

(b) Structured Reasoning Prompt

Model	Acc_{all}	Acc_{real}	Acc_{ai}	Macro F1
Qwen2.5-VL 3B	0.47	0.58	0.35	0.46
Qwen2.5-VL 7B	0.57	0.97	0.16	0.48
Video-LLaVA-HF 7B	0.46	0.29	0.63	0.45

Training on AEGIS boosts authenticity detection. Fine-tuning with LoRA yields substantial performance gains on the in-domain test set—for instance, the macro-F1 of Qwen2.5-VL 7B increases from 0.43 to 0.82. However, as shown in Table 3, improvements on the Hard Test Set remain marginal, with macro-F1 rising only slightly from 0.52 to 0.55. This underscores the persistent challenge of generalization to realistic, high-fidelity forgeries.

The stark contrast between performance on in-domain data and the Hard Test Set underscores the critical generalization challenges posed uniquely by AEGIS. Despite targeted fine-tuning, current models still struggle to generalize learned authenticity cues effectively when confronted with subtle and realistic videos deliberately included in the Hard Test Set.

Table 3: Detection Accuracy on Two Test Set after Fine-tuning

M	T	Eval	Acc_{all}	Acc_{real}	Acc_{ai}	Macro-$F1$
3B	ID	ZS	0.65	0.87	0.55	0.65
7B	ID	ZS	0.45	0.50	0.20	0.43
3B	ID	LoRA	0.78	0.35	1.00	0.69 (+0.04)
7B	ID	LoRA	0.83	0.99	0.75	0.82 (+0.41)
3B	HT	ZS	0.52	0.80	0.23	0.48
7B	HT	ZS	0.59	0.89	0.22	0.52
3B	HT	LoRA	0.56	0.14	0.97	0.47 (-0.01)
7B	HT	LoRA	0.61	0.99	0.24	0.55 (+0.03)

M: Model size (3B = Qwen2.5-VL-3B, 7B = Qwen2.5-VL-7B); **T**: Test set (ID = In-domain, HT = Hard test set); **Eval**: Evaluation Type (ZS = Zero-shot, LoRA = After LoRA fine-tuning).

This limitation strongly indicates the need for future research to explore more advanced and robust fine-tuning or domain adaptation strategies explicitly tailored toward enhancing model generalization to AEGIS-level forgery complexities. These insights collectively underline the unique value and significant challenge presented by AEGIS, clearly establishing it as a critical resource for advancing robust, realistic, and highly generalizable AI-generated video detection research.

5 Conclusion

In this work, we presented AEGIS, a novel large-scale video authenticity benchmark explicitly targeting sophisticated AI-generated videos. Unlike existing datasets, AEGIS prioritizes hyper-realistic scenarios and excludes simplistic or easily detectable samples, significantly enhancing detection complexity and realism. Through rigorous data filtering, strategic dataset

partitioning, and the inclusion of deceptive samples from advancing generative models (e.g., Sora, KLing), AEGIS raises the bar for synthetic video detection. Experimental evaluations reveal that even SOTA vision-language models struggle to generalize in zero-shot settings, particularly on the Hard Test Set. Furthermore, AEGIS offers multi-dimensional visual cues with rich multimodal annotations. These not only support downstream detection tasks but also facilitate interpretable reasoning, enabling finer-grained analysis of model failures and decision boundaries. We believe AEGIS establishes a foundational shift in synthetic video detection research by providing a challenging, diverse, and explainability-oriented benchmark, essential for the development of robust and trustworthy multimodal AI systems.

Acknowledgements

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-RP-2024-033).

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv:2502.13923* (2025).

[2] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. 2024. Videophy: Evaluating physical commonsense for video generation. *arXiv:2406.03520* (2024).

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv:2311.15127* (2023).

[4] Nuria Alina Chandra, Ryan Murtfeldt, Lin Qiu, Arnab Karmakar, Hannah Lee, Emmanuel Tanumihardja, Kevin Farhat, Ben Caffee, Sejin Paik, Changyeon Lee, et al. 2025. Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024. *arXiv:2503.02857* (2025).

[5] Chirui Chang, Zhengzhe Liu, Xiaoyang Lyu, and Xiaojuan Qi. 2024. What Matters in Detecting AI-Generated Videos like Sora? *arXiv:2406.19568* (2024).

[6] Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, et al. 2024. Demamba: Ai-generated video detection on million-scale genvideo benchmark. *arXiv:2405.19707* (2024).

[7] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion models in vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* (2023).

[8] Fanda Fan, Chunjie Luo, Wanling Gao, and Jianfeng Zhan. 2023. Aigcbench: Comprehensive evaluation of image-to-video content generated by ai. *BenchCouncil Trans. Benchmarks Stand. Eval.* (2023).

[9] Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. 2024. Discrete flow matching. In *Proc. NeurIPS*.

[10] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proc. NeurIPS*.

[11] Xinan He, Yue Zhou, Bing Fan, Bin Li, Guopu Zhu, and Feng Ding. 2025. VLForgery Face Triad: Detection, Localization and Attribution via Multimodal Large Language Models. *arXiv:2503.06142* (2025).

[12] Vlad Hondru, Eduard Hogea, Darian Onchis, and Radu Tudor Ionescu. 2025. ExDDV: A New Dataset for Explainable Deepfake Detection in Video. *arXiv:2503.14421* (2025).

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models.. In *Proc. ICLR*.

[14] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proc. IEEE/CVF CVPR*.

[15] Yoori Hwang, Ji Youn Ryu, and Se-Hoon Jeong. 2021. Effects of disinformation using deepfake: The protective effect of media literacy education. *Cyberpsychology, Behavior, and Social Networking* (2021).

[16] Xinyi Jin, Zhuoyue Zhang, Bowen Gao, Shuqing Gao, Wenbo Zhou, Nenghai Yu, and Guoyan Wang. 2025. Assessing the perceived credibility of deepfakes: The impact of system-generated cues and video characteristics. *New Media & Society* (2025).

[17] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proc. IEEE/CVF CVPR*.

[18] Leon Klein, Andreas Krämer, and Frank Noé. 2023. Equivariant flow matching. In *Proc. NeurIPS*.

[19] Kuaishou. 2024. Kling: AI Video Generation Model. https://https://klingai.kuaishou.com//.

[20] Chunyi Li, Tengchuan Kou, Yixuan Gao, Yuqin Cao, Wei Sun, Zicheng Zhang, Yingjie Zhou, Zhichao Zhang, Weixia Zhang, Haoning Wu, et al. 2024. Aigiqa-20k: A large database for ai-generated image quality assessment. In *Proc. IEEE/CVF CVPR*.

[21] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. 2023. Agiqa-3k: An open database for ai-generated image quality assessment. *IEEE Trans. Circuits Syst. Video Technol.* (2023).

[22] Xiaoming Li, Xinyu Hou, and Chen Change Loy. 2024. When stylegan meets stable diffusion: a w+ adapter for personalized image generation. In *Proc. IEEE/CVF CVPR*.

[23] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv:2311.10122* (2023).

[24] Li Lin, Neeraj Gupta, Yue Zhang, Hainan Ren, Chun-Hao Liu, Feng Ding, Xin Wang, Xin Li, Luisa Verdoliva, and Shu Hu. 2024. Detecting multimedia generated by large ai models: A survey. *arXiv:2402.00045* (2024).

[25] Xiao Liu, Xinhao Xiang, Zizhong Li, Yongheng Wang, Zhuoheng Li, Zhuosheng Liu, Weidi Zhang, Weiqi Ye, and Jiawei Zhang. 2024. A survey of ai-generated video evaluation. *arXiv:2410.19884* (2024).

[26] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. 2024. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proc. IEEE/CVF CVPR*.

[27] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. 2024. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv:2402.17177* (2024).

[28] Yibo Miao, Yifan Zhu, Lijia Yu, Jun Zhu, Xiao-Shan Gao, and Yinpeng Dong. 2024. T2vsafetybench: Evaluating the safety of text-to-video generative models. In *Proc. NeurIPS*.

[29] Zhenliang Ni, Qiangyu Yan, Mouxiao Huang, Tianning Yuan, Yehui Tang, Hailin Hu, Xinghao Chen, and Yunhe Wang. 2025. GenVidBench: A Challenging Benchmark for Detecting AI-Generated Video. *arXiv:2501.11340* (2025).

[30] OpenAI. 2023. GPT-4V. https://openai.com/index/gpt-4v-system-card/.

[31] OpenAI. 2024. GPT-4o. https://platform.openai.com/docs/models/gpt-4o.

[32] OpenAI. 2024. Sora: AI Video Generation Model. https://openai.com/sora.

[33] Pika Labs. 2024. Pika: AI Video Generation Platform. https://www.pika.art/.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proc. ICML*.

[35] Xiufeng Song, Xiao Guo, Jiache Zhang, Qirui Li, Lei Bai, Xiaoming Liu, Guangtao Zhai, and Xiaohong Liu. 2024. On learning multi-modal forgery representation for diffusion generated video detection. *arXiv:2410.23623* (2024).

[36] Zhen Sun, Ziyi Zhang, Zeren Luo, Zeyang Sha, Tianshuo Cong, Zheng Li, Shiwen Cui, Weiqiang Wang, Jiaheng Wei, Xinlei He, Qi Li, and Qian Wang. 2025. FragFake: A Dataset for Fine-Grained Detection of Edited Images with Vision Language Models. *arXiv:2505.15644* (2025).

[37] Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Proc. ECCV*.

[38] Cristian Vaccari and Andrew Chadwick. 2020. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society* (2020).

[39] Jiarui Wang, Huiyu Duan, Jing Liu, Shi Chen, Xiongkuo Min, and Guangtao Zhai. 2023. Aigciqa2023: A large-scale image quality assessment database for ai generated images: from the perspectives of quality, authenticity and correspondence. In *CAAI ICAI*.

[40] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. 2023. VideoFactory: Swap Attention in Spatiotemporal Diffusions for Text-to-Video Generation. *arXiv:2305.10874* (2023).

[41] Wenhao Wang and Yi Yang. 2024. TIP-I2V: A Million-Scale Real Text and Image Prompt Dataset for Image-to-Video Generation. *arXiv:2411.04709* (2024).

[42] Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, et al. 2024. Df40: Toward next-generation deepfake detection. *arXiv:2406.13495* (2024).

[43] Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. 2024. Vript: A video is worth thousands of words. In *Proc. NeurIPS*.

[44] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. *arXiv:2408.06072* (2024).

[45] Jiquan Yuan, Fanyi Yang, Jihe Li, Xinyan Cao, Jinming Che, Jinlong Lin, and Xixin Cao. 2024. PKU-AIGIQA-4K: A Perceptual Quality Assessment Database for

Both Text-to-Image and Image-to-Image AI-Generated Images. *arXiv:2404.18409* (2024).

[46] Ruihan Zhang, Borou Yu, Jiajian Min, Yetong Xin, Zheng Wei, Juncheng Nemo Shi, Mingzhen Huang, Xianghao Kong, Nix Liu Xin, Shanshan Jiang, et al. 2025. Generative AI for Film Creation: A Survey of Recent Advances. *arXiv:2504.08296* (2025).

[47] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. 2023. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv:2311.04145* (2023).

[48] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv:2403.13372* (2024).

[49] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. 2024. Open-sora: Democratizing efficient video production for all. *arXiv:2412.20404* (2024).

[50] Yufan Zhou, Bingchen Liu, Yizhe Zhu, Xiao Yang, Changyou Chen, and Jinhui Xu. 2023. Shifted diffusion for text-to-image generation. In *Proc. IEEE/CVF CVPR*.