CAN AI-GENERATED VIDEO DETECTORS KEEP UP?

Anonymous authors
Paper under double-blind review



ABSTRACT

Detecting AI-generated videos (AIGVs) has become increasingly important as powerful video generation models continue to emerge. However, existing benchmarks provide limited indications of how detectors will perform in practical deployment for two key reasons: (1) strong performance on internal test sets often fails to transfer to broader data distributions; and (2) commonly used one-to-many protocols and non-temporal splits do not conform to actual application. We introduce **AIGV-Annual**, a million-scale, temporally partitioned and extensible benchmark for AI-generated video detection, covering videos from 100 fake generators and 13 real sources. AIGV-Annual establishes a cumulative many-to-many protocol, in which detectors are trained on generators released before each temporal cutoff and evaluated on those released afterward, enabling the measurement of temporal decay and update gain under continual generator evolution. Using AIGV-Annual, we systematically evaluate 18 supervised baselines and 16 zero-shot configurations across multiple temporal cutoffs. Our experiments reveal substantial out-of-distribution degradation, temporal decay and varying update gains from newer training data. We further introduce a spatial-temporal detector. Extensive experiments demonstrate that our method outperforms state-of-the-art methods on existing benchmarks and in-the-wild data.

1 INTRODUCTION

AI-generated video (AIGV) is evolving rapidly, driven by the continual release of new generative models. Their increasingly realistic outputs pose serious threats to media forensics, content moderation and misinformation control, making reliable detection imperative. Since new generators always emerge after detector training (Figure 1(a)), a detector must go beyond recognizing earlier generators and generalize to future ones.

To evaluate AIGV detection, recent benchmarks and evaluation protocols have been proposed (Bai et al., 2024; Ma et al., 2025b; Chen et al., 2024c; Ni et al., 2026; Ma et al., 2026). As shown in Figure 1, although these works have steadily improved dataset scale and generator coverage, they have two key limitations: i) **Limited external validity**: detectors that perform well on internal test sets often degrade substantially on external AIGV datasets, indicating that within-benchmark scores may not reliably reflect performance across broader data distributions; ii) **Deployment misalignment**: existing one-to-many and non-temporal protocols do not reflect practical deployment, where detectors should be trained on all available generators and evaluated on later releases under a chronological,

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

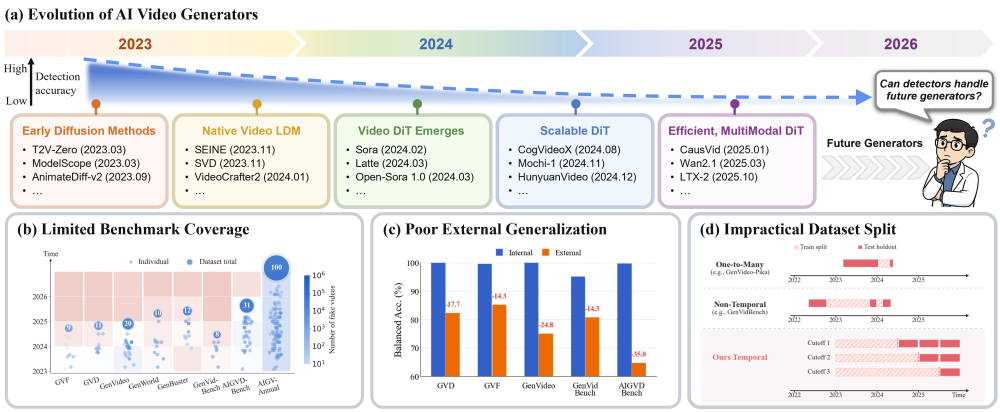


Figure 1: **Limitations of existing AI-generated video benchmarks.** (a) Rapid evolution of video generators. (b) Limited and outdated generator coverage. (c) Strong internal performance can mask weak external generalization. (d) Conventional splits overlook generator release order.

cumulative many-to-many setting. As a result, *existing benchmarks provide static snapshots of detector performance rather than reliable assessments under broader and evolving data distributions.*

We address these limitations with **AIGV-Annual**, a million-scale, temporally partitioned, and extensible benchmark for AI-generated video detection. Its design comprises three features: (i) *diversity*: 1.29M videos, including 743K real and 551K fake videos from 13 real sources and 100 generators, spanning diverse generator families, architectures, release periods, and semantic content; (ii) *temporality*: generators are grouped by public release time under a cumulative many-to-many protocol, where detectors train on all covered generators available by each cutoff and evaluate on each later group; and (iii) *extensibility*: new generators can be appended without redefining earlier splits. This design supports three complementary evaluations: cross-dataset transfer on external datasets, temporal decay across future groups, and update gain across training cutoffs.

Using AIGV-Annual, we examine *cross-dataset generalization* with representative detectors, evaluate 18 supervised baselines for *temporal decay* and *update gain* under the cumulative temporal protocol, and assess 16 zero-shot configurations across temporal release groups. Our results reveal substantial cross-dataset degradation and show that detection performance often declines on later-released generators. For some detectors, incorporating newer training data yields uneven gains across detectors and can even reduce performance. Our analysis suggests a detector-generator misalignment: *detectors favor different forensic cues, while the relative strengths of spatial and temporal cues vary across generated videos*. Combining both types of evidence may therefore reduce the blind spots of relying on either alone. Extensive evaluations show that our detector outperforms the compared baselines on future generators, existing benchmarks, and in-the-wild datasets.

Overall, our contributions are: (1) **Benchmark**: AIGV-Annual contains 1.29M videos from 100 fake video generators and 13 real video sources, providing broad real and fake source coverage. (2) **Protocol**: We organize generated videos by public generator release time and establish a cumulative many-to-many protocol for future-generator generalization. (3) **Evaluation**: We benchmark 18 supervised baselines and 16 zero-shot configurations under cross-dataset and temporal evaluation settings, revealing substantial generalization gaps across datasets and later-released generators. (4) **Method**: Guided by our findings, we propose a detector that achieves strong performance and remains robust under generator evolution.

2 Related Work

2.1 Benchmarks for AI-generated Video Detection

Many datasets have been proposed to evaluate detectors for AI-generated videos, mostly by expanding dataset scale, generator coverage, and deployment scenario coverage (see Table 1). Early benchmarks such as GVD (Bai et al., 2024) and GVF (Ma et al., 2025b) are relatively small and cover only early

Table 1: Comparison with related AI-generated video detection benchmarks.

Benchmark	Year	Venue	Scale	Real	Fake	Fake Coverage					Evaluation Protocol	
						≤2023	2024	2025	2026	Unk.	Train–Test	Temporal
GVF	2025	ICME	2.8k	2	9	4	5	–	–	–	Many-to-Many	No
GVD	2024	PRCV	11.6k	2	11	6	5	–	–	–	One-to-Many	No
GenVideo	2024	arXiv	2.3M	3	20	14	5	–	–	1	One-to-Many	No
GenWorld	2025	arXiv	100k	4	10	4	5	1	–	–	Many-to-Many	No
GenBuster	2025	arXiv	200k	1	12	–	4	8	–	–	ID vs. OOD	Partial
GenVidBench	2026	AAAI	218k	2	8	–	–	–	–	–	Pair1 vs. Pair2	No
AIGVDBench	2026	CVPR	844k	1	31	6	16	9	–	–	One-to-Many	No
AIGV-Annual	**2026**	**–**	**1.29M**	**13**	**100**	**16**	**32**	**29**	**13**	**10**	**Many-to-Many**	**Yes**

Table 2: **Cross-dataset generalization of AIGV detectors trained on different benchmarks.** For existing benchmarks, we report the better score from available official checkpoints and reproductions under official settings. Acc. denotes balanced accuracy; all scores are reported as percentages. Avg., Worst, and Δ_{cross} denote the mean, minimum, and range across cross-dataset evaluation splits.

Training Set	Detector	Evaluation Set (AUC)										Evaluation Set (Acc.)									
		GVF	Gen Video	GenVidBench P1	GenVidBench P2	AIGVD Bench	AIGV-Annual	Mint Vid	Avg.	Worst	Δ_{cross}	GVD	GVF	Gen Video	GenVidBench P1	GenVidBench P2	AIGVD Bench	AIGV-Annual	Mint Vid	Avg.	Worst	Δ_{cross}
GVD	AIGVDet	63.73	76.04	61.19	68.08	72.77	64.62	74.83	68.75	61.19	14.85	–	49.88	49.98	50.50	50.50	50.87	51.00	51.24	50.57	49.88	1.36
GVF	DeCoF	–	98.73	94.85	79.63	89.56	71.15	71.62	84.26	71.15	27.58	97.62	–	92.47	80.38	73.50	65.84	68.50	61.32	77.09	61.32	36.30
GenVideo	DeMamba	77.48	–	75.37	68.46	78.09	66.30	63.97	71.61	63.97	14.12	85.62	69.79	–	67.31	63.69	65.16	61.82	58.51	67.41	58.51	27.11
GenVidBench	MViTv2	89.62	95.46	–	–	88.22	71.98	69.54	82.96	69.54	25.92	92.92	79.13	89.10	–	–	74.95	65.50	64.40	77.67	64.40	28.52
AIGVDBench	DeCoF	35.51	15.73	80.69	67.33	–	46.32	66.19	51.96	15.73	64.96	46.18	40.81	26.35	76.88	62.56	–	45.44	60.83	51.29	26.35	50.53
AIGV-Annual	DeCoF	92.90	93.85	98.76	81.05	98.47	–	77.27	90.38	77.27	21.49	78.54	83.73	86.85	93.84	72.08	93.60	–	69.13	82.54	69.13	24.71
AIGV-Annual	DeMamba	90.90	89.92	96.01	89.18	94.72	–	78.40	89.86	78.40	17.61	85.14	78.67	73.52	84.71	78.55	90.50	–	73.87	80.71	73.52	16.98
AIGV-Annual	**Ours**	99.71	99.50	99.90	98.95	99.86	–	91.96	98.31	91.96	7.94	84.70	97.66	97.80	97.94	93.44	98.57	–	82.48	93.23	82.48	16.09

text-to-video generators, resulting in limited scale and generator diversity. To broaden generator coverage, GenVideo (Chen et al., 2024c) collects fake videos from around 20 generators. GenVidBench (Ni et al., 2026) further uses cross-source pairs over 8 state-of-the-art generators. AIGVDBench (Ma et al., 2026) expands generator coverage to over 30 generation methods. These benchmarks partially address the generalization problem, as they only expand coverage without changing the evaluation assumption. Additionally, MintVid (Tan et al., 2026) focuses on recent generators, GenWorld (Chen et al., 2025b) studies real-world simulation scenarios, and GenBuster++ (Wen et al., 2025) examines detection across multiple settings. However, none of these benchmarks directly test the practical question faced by deployed detectors: whether models trained on currently known generators can remain reliable on generators released later.

2.2 Methods for AI-generated Video Detection

Existing detectors mainly exploit spatial artifacts, temporal inconsistencies, or pretrained representations. Early video-level approaches rely on explicit temporal cues. STIL (Gu et al., 2021) models adjacent-frame differences, TALL (Xu et al., 2023) arranges consecutive frames into thumbnail layouts, and AIGVDet (Bai et al., 2024) extracts motion cues from optical flow. Recent methods explore more structured representations. DeMamba (Chen et al., 2024c) employs state-space modeling, DeCoF (Ma et al., 2025b) examines frame consistency, NSG-VD (Zhang et al., 2025c) introduces physics-driven modeling, and ReStraV (Internò et al., 2025) analyzes representation geometry. STALL (Ben Hayun et al., 2026) develops training-free spatial-temporal likelihoods, while VideoVeritas (Tan et al., 2026) combines fine-grained perception with fact-based reasoning. Despite these advances, reliable generalization to unseen generators remains challenging.

3 Motivation and Analysis

A Practical Review of Existing Benchmarks Existing benchmarks primarily report detector performance on their internal test sets, making it unclear whether these results remain reliable across broader data distributions encountered in deployment. We examine their cross-benchmark reliability, generator coverage and evaluation protocols. Our analysis reveals two key limitations: limited external validity and deployment misalignment.

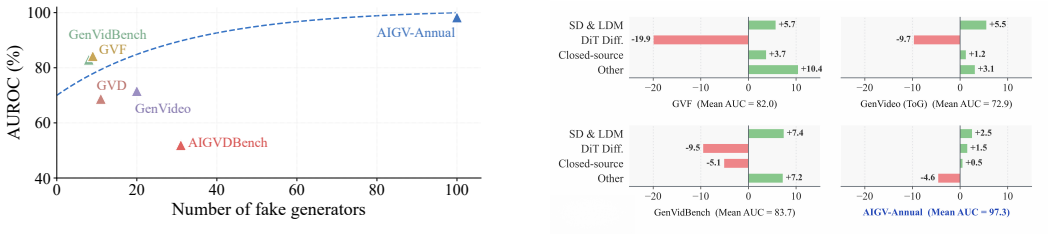


Figure 2: **Performance variation across data distributions.** We use detectors trained on their respective benchmarks. (Left) Generator coverage and cross-benchmark detection performance. (Right) Generator-family performance deviations from the corresponding benchmark average.

Limitation 1: Limited External Validity. We identify two issues: (i) *limited cross-benchmark reliability* and (ii) *insufficient generator coverage*. **(i) Limited cross-benchmark reliability.** We use Δ_{cross}, the difference between the highest and lowest scores across evaluation splits, to quantify cross-benchmark variation. Table 2 shows that strong within-benchmark performance may drop to near or below chance on other benchmarks. **(ii) Insufficient generator coverage.** Existing benchmarks cover at most 31 fake generators and a limited range of generator families (Table 1). Figure 2 shows substantial performance variation across generator families. Detectors trained on AIGV-Annual exhibit smaller cross-dataset and cross-family variations, motivating broader generator coverage for reliable evaluation.

Limitation 2: Deployment Misalignment. Real-world deployment is both *chronological* and *cumulative*: at any cutoff, the detector can be trained on all previously released generators, but it needs to be able to generalize to later-released generators (Epstein et al., 2023; Pellegrini et al., 2025). GVD, GenVideo, and AIGVDBench adopt one-to-many evaluation, GVF uses a fixed many-to-many split, and GenVidBench evaluates non-temporal source pairs. As illustrated in Figure 1(d), these protocols ignore generator release order and fail to preserve the past-to-future boundary, motivating a temporally cumulative evaluation protocol.

4 AIGV-Annual Benchmark

We propose AIGV-Annual, a million-scale, temporally partitioned, and extensible benchmark designed to address the limitations identified above. It combines broad generator coverage with a deployment-aligned cumulative many-to-many protocol, enabling systematic evaluation under continual generator evolution.

4.1 Temporal Partition and Protocol

We formulate AIGV detection as binary classification under a **cumulative many-to-many** protocol. Generators are grouped into half-year periods by release time; at each cutoff, detectors train on all earlier groups and test on each later group. The temporal partition is applied only to fake generators, while real videos are sampled from fixed source pools. We further collect videos from commercial platforms with undisclosed generators and versions as an **Unknown** group to assess external validity beyond known generators.

4.2 Dataset Construction

Fake Video Sources. AIGV-Annual contains 551K fake videos from 100 generators spanning diverse release periods, architecture families, and generation pipelines. We aggregate public model-generated datasets and community sources from Hugging Face (Hugging Face, 2026) and ModelScope (Wang et al., 2023), using only official training splits when available and capping each generator at 10K videos. The *Unknown* group contains 35.2K videos from 10 commercial platforms with undisclosed generators and versions. Figure 3 summarizes the generator coverage, source composition, and video metadata ranges.

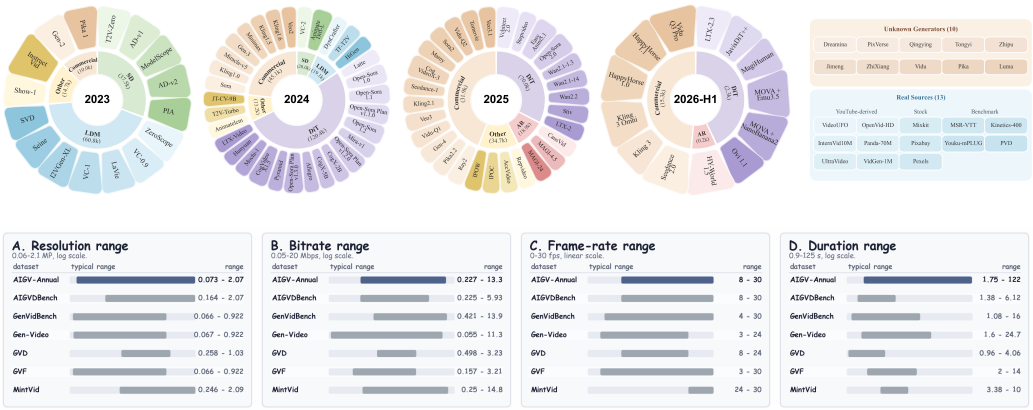


Figure 3: Overview of the generators in AIGV-Annual.

Real Video Sources. For training, we sample real videos from VideoUFO (Wang & Yang, 2025), OpenVid-HD (Nan et al., 2024), InternVid-10M-FLT (Wang et al., 2024c), Panda-70M (Chen et al., 2024d), UltraVideo (Xue et al., 2025), and VidGen-1M (Tan et al., 2024b). For testing, we collect real videos from stock-footage platforms, including MixKit (Mixkit, 2023), Pixabay (Pixabay, 2024), and Pexels (Pexels, 2024), as well as benchmark datasets including MSR-VTT (Xu et al., 2016), Kinetics-400 (Kay et al., 2017), Youku-mPLUG (Li et al., 2022a), and PVD (Bolya et al., 2026).

5 Experiments

5.1 Experimental Setup

Data split. We instantiate the temporal protocol with three chronological training cutoffs: 2024-H1, 2024-H2, and 2025-H1. For each cutoff, a detector is trained on all generators released up to that point and evaluated *separately* on each later half-year group. The unknown group is additionally held out as an extra test set under every cutoff.

Evaluation Metrics and Compared Methods. Following recent AIGV detection studies (Chen et al., 2024c; Internò et al., 2025), we report balanced accuracy (Acc.), defined as the mean of real and AI-generated video classification accuracies, and the area under the receiver operating characteristic curve (AUC). All results are reported as percentages. We compare 18 supervised baselines across three categories: eight *video-based detectors*: STIL, TALL, AIGVDet, DeMamba, DeCoF, Plug-and-Play, NSG-VD, and ReStraV; four *general video backbones*: VideoMAE, VideoSwin, MViTv2, and UniFormer; and six *image-based detectors*: CNNSpot, UnivFD, NPR, Effort, Community-Forensics, and SAFE. We further evaluate 16 zero-shot configurations across three categories: eight *image-based detectors*: UnivFD, NPR, DRCT, B-Free, DDA, MIRROR, DGS-Net, and PROBE; six *video-based detectors*: DeCoF, D3, NVIDIA's AIGV detector, CoCoDetect, STALL, and Native-Scale; and two *LLM-based video detectors*: Skyra and VideoVeritas.

5.2 Benchmarking Existing Methods

We evaluate existing detectors along three dimensions. **(i) Temporal generalization.** Across training cutoffs, detector performance varies substantially across future release groups and often decreases as the temporal gap increases (Table 3). The magnitude and direction of these changes differ across methods, indicating that evaluation on a single group cannot fully characterize temporal generalization. **(ii) Zero-shot generalization.** To assess their general applicability to emerging generators, we evaluate existing detectors on AIGV-Annual without additional training. Most detectors perform poorly in this setting. In contrast, the frequency-aware methods DDA (Chen et al., 2025a) and NVIDIA's detector (NVIDIA, 2026) remain strong performers (Table 4). Their results suggest that frequency-domain evidence may transfer more effectively across generator releases. **(iii) Update gain.** Adding newer training data yields different effects across detector architectures (Figure 4). General

Table 3: **Temporal generalization under cumulative training cutoffs.** Detectors are trained on generators released no later than each cutoff and evaluated on later release groups. Values report balanced accuracy (%). (↓ drop, ↑ gain).

Method	Venue (Year)	Train ≤ 2024-H1					Train ≤ 2024-H2				Train ≤ 2025-H1		
		2024-H2	2025-H1	2025-H2	2026-H1	Unk.	2025-H1	2025-H2	2026-H1	Unk.	2025-H2	2026-H1	Unk.
Image Detectors													
Single-model Detectors													
CNNSpot	CVPR (2020)	79.7	79.7 0.0	74.5 ↓5.2	72.0 ↓7.7	82.3	82.8	79.0 ↓3.8	73.7 ↓9.1	83.8	75.0	72.3 ↓2.7	79.1
UnivFD	CVPR (2023)	65.9	65.5 ↓0.4	67.5 ↑1.6	64.9 ↓1.0	71.9	66.9	66.5 ↓0.4	65.8 ↓1.1	73.0	66.4	66.9 ↑0.5	74.3
NPR	CVPR (2024a)	70.8	67.0 ↓3.8	69.2 ↓1.6	57.6 ↓13.2	72.0	70.7	71.5 ↑0.8	67.4 ↓3.3	77.5	69.9	68.9 ↓1.0	72.4
Effort	ICML (2024)	85.2	**85.0** ↓0.2	**82.4** ↓2.8	83.7 ↓1.5	**89.5**	**90.3**	**86.2** ↓4.1	83.9 ↓6.4	**90.8**	**87.2**	84.0 ↓3.2	**92.3**
Ensemble Detectors													
SAFE	KDD (2025)	59.1	56.5 ↓2.6	56.3 ↓2.8	51.8 ↓7.3	55.8	53.3	54.7 ↑1.4	49.6 ↓3.7	51.5	55.6	50.6 ↓5.0	52.7
Community-Forensics	CVPR (2025)	80.3	78.5 ↓1.8	72.7 ↓7.6	73.1 ↓7.2	80.1	83.6	80.3 ↓3.3	78.0 ↓5.6	83.6	81.4	79.1 ↓2.3	85.2
General Video Backbones													
MViTv2	CVPR (2022c)	75.8	74.5 ↓1.3	70.3 ↓5.5	72.7 ↓3.1	77.0	81.5	74.4 ↓7.1	74.0 ↓7.5	83.3	76.3	75.3 ↓1.0	84.6
VideoMAE	NeurIPS (2022)	76.8	72.5 ↓4.3	69.9 ↓6.9	70.9 ↓5.9	74.0	79.9	72.9 ↓7.0	73.8 ↓6.1	79.9	78.4	74.6 ↓3.8	82.8
VideoSwin	CVPR (2022)	72.5	69.2 ↓3.3	62.4 ↓10.1	57.6 ↓14.9	74.5	71.5	63.0 ↓8.5	52.1 ↓19.4	75.9	67.2	63.1 ↓4.1	80.0
UniFormer	ICLR (2022b)	59.6	56.6 ↓3.0	49.9 ↓9.7	67.6 ↑8.0	59.2	58.1	51.5 ↓6.6	61.3 ↑3.2	62.2	55.1	61.4 ↑6.3	66.3
Video Detectors													
Generated Video Detectors													
STIL	ACM MM (2021)	75.7	74.3 ↓1.4	67.5 ↓8.2	59.1 ↓16.6	75.7	75.1	69.8 ↓5.3	57.5 ↓17.6	76.4	71.0	63.9 ↓7.1	76.9
TALL	ICCV (2023)	68.3	67.5 ↓0.8	60.7 ↓7.6	49.0 ↓19.3	69.9	69.4	62.3 ↓7.1	51.9 ↓17.5	71.2	63.9	54.5 ↓9.4	72.2
DeMamba	arXiv (2024c)	80.2	76.6 ↓3.6	65.8 ↓14.4	70.2 ↓10.0	75.4	76.5	67.1 ↓9.4	71.7 ↓4.8	76.2	75.0	69.9 ↓5.1	82.9
AIGVDet	PRCV (2024)	75.7	73.2 ↓2.5	70.8 ↓4.9	60.6 ↓15.1	78.3	81.1	69.1 ↓12.0	66.7 ↓14.4	82.9	74.7	69.3 ↓5.4	84.3
DeCoF	ICME (2025b)	72.1	71.4 ↓0.7	68.4 ↓3.7	66.5 ↓5.6	76.4	75.5	70.4 ↓5.1	68.6 ↓6.9	79.7	69.7	70.5 ↑0.8	77.8
Plug-and-Play	CVPR (2025)	**85.5**	84.8 ↓0.7	82.3 ↓3.2	71.0 ↓14.5	88.6	85.1	81.5 ↓3.6	70.3 ↓14.8	85.9	82.4	70.0 ↓12.4	87.5
NSG-VD	NeurIPS (2025c)	54.1	52.3 ↓1.8	48.2 ↓5.9	48.4 ↓5.7	58.2	59.1	50.8 ↓8.3	50.0 ↓9.1	62.8	50.5	50.0 ↓0.5	64.8
ReStraV	NeurIPS (2025)	67.9	64.9 ↓3.0	63.1 ↓4.8	46.6 ↓21.3	66.3	69.9	67.9 ↓2.0	39.3 ↓30.6	71.3	69.1	33.3 ↓35.8	73.5

Table 4: **Zero-shot generalization across temporal release groups.** Official checkpoints are evaluated without retraining. Values report balanced accuracy (%). (↓ drop, ↑ gain).

Test Set	UnivFD 2023	NPR 2024a	DRCT 2024a	DeCoF 2025b	D3 2025a	B-Free 2025	DDA 2025a	NVIDIA's 2026	CoCo Detect 2026	STALL 2026	Skyra 2026a	MIRROR 2026b	DGS-Net 2026	Video Veritas 2026	Native-Scale 2026b	PROBE 2026
2024-H2	51.4	50.4	67.8	44.9	62.4	71.8	**84.0**	83.7	43.8	54.6	55.7	61.3	64.4	67.6	70.6	76.2
2025-H1	50.2 ↓1.2	51.8 ↑1.4	60.2 ↓7.6	44.0 ↓0.9	62.4 0.0	64.0 ↓7.8	**83.1** ↓0.9	82.6 ↓1.2	45.6 ↑1.8	55.0 ↑0.4	52.7 ↓3.0	59.2 ↓2.1	66.5 ↑2.1	68.3 ↑0.7	71.0 ↑0.4	66.3 ↓9.9
2025-H2	50.4 ↓1.0	50.0 ↓0.5	61.6 ↓6.2	39.9 ↓5.0	59.1 ↓3.3	57.4 ↓14.4	**81.7** ↓2.4	78.1 ↓5.6	46.4 ↑2.6	55.3 ↑0.7	52.0 ↓3.7	52.9 ↓8.4	56.5 ↓7.9	62.7 ↓4.9	69.4 ↓1.2	62.5 ↓13.7
2026-H1	49.8 ↓1.6	52.6 ↑2.2	60.6 ↓7.2	39.6 ↓5.3	40.3 ↓22.1	66.9 ↓4.9	74.4 ↓9.6	**80.0** ↓3.7	50.4 ↑6.6	54.6 0.0	68.2 ↑12.5	58.4 ↓2.9	61.6 ↓2.8	70.0 ↑2.4	67.2 ↓3.4	66.9 ↓9.3

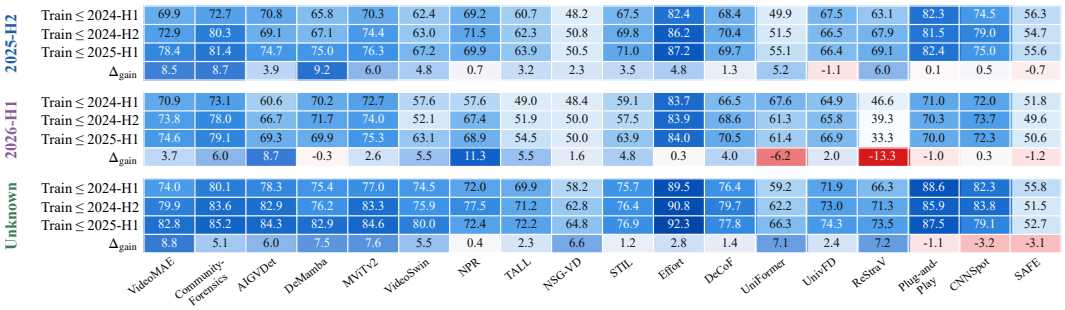
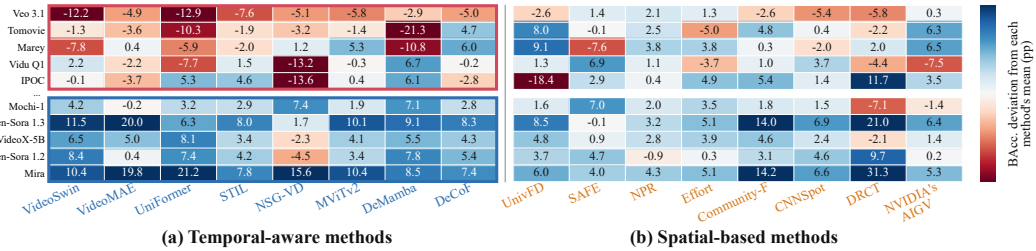


Figure 4: **Temporal update gains.** Values report balanced accuracy (%). Δ_{gain} denotes the change from Train ≤ 2024-H1 to Train ≤ 2025-H1.

video backbones, including VideoMAE, MViTv2 and VideoSwin, achieve relatively consistent positive gains, suggesting that general-purpose video representations can effectively incorporate evidence from newly added generators. In contrast, SAFE (Li et al., 2025) consistently degrades, suggesting that additional data provide limited benefit when the learned detection evidence does not transfer.

5.3 A Spatial-Temporal Detector

Motivation. **Detector-Generator Misalignment.** Spatial and temporal detectors focus on different cues, whose strengths vary across generated content. *Detector side:* spatial and temporal detectors



(a) Temporal-aware methods (b) Spatial-based methods

Figure 5: **Detector performance on generators in AIGV-Annual.** Rows list generators from newer to older. Columns show temporal (blue) and spatial (orange) methods.

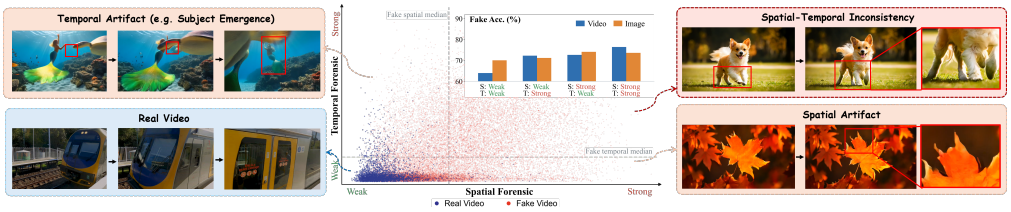


Figure 6: **Spatial and temporal forensic degrees.** We measure the spatial forensic based on (Betser et al., 2025) and temporal forensic based on the optical-flow magnitude and directional-consistency analysis of (Yang et al., 2025b). Blue: real (YouTube videos in AIGV-Annual); red: fake (fake videos from AIGV-Annual). The bar chart shows the average detection performance of methods in Figure 5 on fake videos in each region.

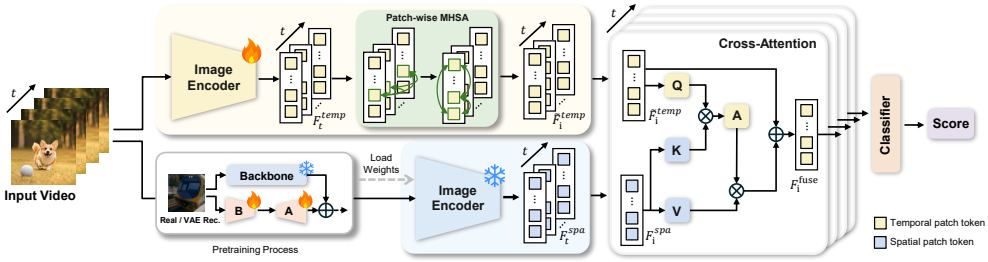


Figure 7: **Framework of our spatial-temporal detector.** We illustrate the overall pipeline and each module, including LoRA-based pretraining process of the spatial encoder, patch-wise temporal modeling, and cross-attention between the two branches.

exhibit different performance patterns across generators (Figure 5). *Generator side:* generated videos exhibit spatial artifacts and temporal inconsistencies with varying relative strengths (Figure 6). These observations suggest a mismatch between detector reliance and available forensic evidence, which may hinder generalization across diverse generators. **Spatial-Temporal Complementarity.** The relative strengths of spatial and temporal detectors vary with the available cues (Figure 6). Relying on either type alone may therefore miss evidence better captured by the other. This motivates jointly using spatial and temporal evidence to reduce such blind spots and support cross-generator generalization.

Overview. As illustrated in Figure 7, given an input video, the temporal branch extracts patch tokens and applies separate inter-frame and intra-frame self-attention to model relations across frames and among patches within each frame. The spatial branch uses a frozen pretrained encoder to extract complementary spatial features. Cross-attention integrates these spatial features into the enhanced temporal tokens through a residual connection, followed by the classification head for video-level prediction.

Details of Network Design. Our detector consists of three main components. **(i) Temporal branch.** A backbone encoder extracts patch tokens from each sampled frame. Inter-frame self-attention

Table 5: Comparison of balanced accuracy between our detector and representative supervised methods under cumulative temporal evaluation.

	Train ≤ 2024-H1						Train ≤ 2024-H2					Train ≤ 2025-H1			
Method	**2024-H2**	**2025-H1**	**2025-H2**	**2026-H1**	**Unk.**	**Avg.**	**2025-H1**	**2025-H2**	**2026-H1**	**Unk.**	**Avg.**	**2025-H2**	**2026-H1**	**Unk.**	**Avg.**
AIGVDet	75.7	73.2	70.8	60.6	78.3	71.7	81.1	69.1	66.7	82.9	75.0	74.7	69.3	84.3	76.1
DeCoF	72.1	71.4	68.4	66.5	76.4	71.0	75.5	70.4	68.6	79.7	73.6	69.7	70.5	77.8	72.7
DeMamba	80.2	76.6	65.8	70.2	75.4	73.6	76.5	67.1	71.7	76.2	72.9	75.0	69.9	82.9	75.9
Community-Forensics	80.3	78.5	72.7	73.1	80.1	76.9	83.6	80.3	78.0	83.6	81.4	81.4	79.1	85.2	81.9
CNNSpot	79.7	79.7	74.5	72.0	82.3	77.6	82.8	79.0	73.7	83.8	79.8	75.0	72.3	79.1	75.5
Plug-and-Play	85.5	84.8	82.3	71.0	88.6	82.4	85.1	81.5	70.3	85.9	80.7	82.4	70.0	87.5	80.0
Effort	85.2	85.0	82.4	83.7	89.5	85.2	90.3	86.2	83.9	90.8	87.8	87.2	84.0	92.3	87.8
Ours	**93.8**	**93.2**	**85.3**	**87.7**	**97.2**	**91.4** (+6.2)	**95.0**	**91.1**	**88.4**	**96.5**	**92.8** (+5.0)	**94.5**	**90.6**	**97.1**	**94.1** (+6.3)

Table 6: Comparison of balanced accuracy on existing benchmarks.

Method	**GVD**	**GVF**	**GenVideo**	**GenVidBench**		**Avg.**
				Pair 1	**Pair 2**	
DeCoF	91.6	69.1	64.7	71.4	86.6	76.7
DeMamba	70.7	55.4	76.8	54.6	51.1	61.7
MViTv2	70.1	62.5	93.6	75.0	82.8	76.8
STIL	71.2	65.0	91.3	72.6	66.9	73.4
Plug-and-Play	90.5	62.9	91.8	72.5	93.0	82.1
Effort	92.8	70.9	94.6	70.3	97.5	85.2
Ours	**95.2**	**97.9**	**97.0**	**95.5**	**99.1**	**96.9** (+11.7)

Table 7: Comparison of balanced accuracy across in-the-wild data without training limitation.

Method	**Training set**	**MintVid**	**AIGVDBench Closed**	**AIGV-Annual Unknown**	**Avg.**
STALL	Training-free	50.9	54.3	55.0	53.4
ForgeLens	ProGAN	54.6	66.2	66.7	62.5
Effort	GenImage	56.8	76.8	67.8	67.1
MViTv2	GenVidBench	62.8	78.1	70.0	70.3
DeMamba	GenVideo	56.7	64.9	62.3	61.3
DeCoF	GVF	61.3	65.8	68.9	65.3
DeCoF	AIGVDBench	60.9	67.5	45.4	57.9
DeCoF	AIGV-Annual	69.1	93.6	77.8	80.2
DeMamba	AIGV-Annual	73.9	90.5	82.9	82.4
Ours	**AIGV-Annual**	**82.5**	**97.7**	**97.1**	**92.4** (+10.0)

operates across frames at corresponding spatial grid positions, followed by intra-frame self-attention among patches within each frame (Vaswani et al., 2017), yielding enhanced temporal representations. **(ii) Spatial branch.** A frozen image encoder pretrained using dual data alignment (Chen et al., 2025a) extracts spatial patch tokens from each sampled frame to provide complementary artifact cues. **(iii) Fusion and prediction.** Cross-attention uses the enhanced temporal tokens to form queries and the same-frame spatial tokens to form keys and values (Vaswani et al., 2017). The projected attention outputs are added to the temporal tokens through a learnable scaled residual connection, yielding fused tokens. These tokens are aggregated and passed to the classification head for video-level prediction.

5.4 Comparative Evaluation

Comparison on AIGV-Annual. Table 5 compares supervised detectors under the cumulative temporal protocol. Our method achieves the best balanced accuracy on every future group under all three training cutoffs. Adding newer training data generally improves its performance on the shared future groups, while its performance on the unknown source group remains consistently high.

Comparison on Existing Benchmarks. Table 6 reports performance on GVD (Bai et al., 2024), GVF (Ma et al., 2025b), GenVideo (Chen et al., 2024c) and the two partitions of GenVidBench (Ni et al., 2026), with detectors trained separately on the official training split of each benchmark. These benchmarks differ in their data sources and generator coverage, providing complementary evaluations under their established protocols. Our detector ranks first in every reported setting and achieves an average balanced accuracy of 96.9%, demonstrating its consistent performance across benchmarks.

Comparison on In-the-Wild Datasets. Table 7 reports performance on three in-the-wild evaluation sets without imposing a common training set. For detector architectures evaluated with multiple training sets, training on AIGV-Annual generally produces stronger external performance than training on existing benchmarks. Our detector achieves an average balanced accuracy of 92.4%, demonstrating its strong generalization to in-the-wild data.

5.5 Ablation and Further Analysis

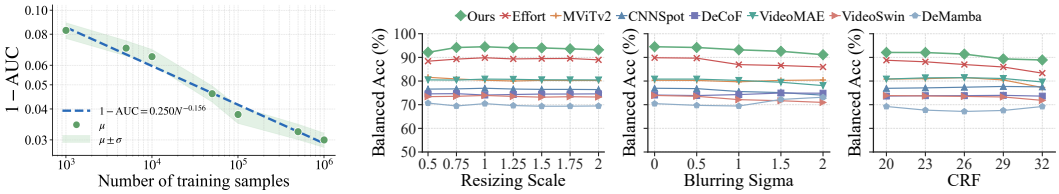
We study backbone choice and data augmentation, component contributions, and scaling and robustness. Unless otherwise stated, we use the training cutoff of 2025-H1.

Table 8: Component ablation studies.

Spatial	Temporal	Patch-wise MHSA	Cross-Attn.	2025-H2 Acc.	2025-H2 AUC	2026-H1 Acc.	2026-H1 AUC	Unknown Acc.	Unknown AUC
✓				81.4	92.9	81.8	90.1	85.6	97.7
	✓	✓		91.8	93.6	84.8	89.2	90.8	96.2
✓	✓			92.5	96.9	88.4	93.6	95.5	98.4
✓	✓	✓		93.0	97.3	89.0	94.1	95.8	98.6
✓	✓	✓	✓	**94.5**	**98.7**	**90.6**	**95.6**	**97.1**	**99.8**

Table 9: Comparison of backbones.

Backbone	2025-H2 Acc.	2025-H2 AUC	2026-H1 Acc.	2026-H1 AUC	Unknown Acc.	Unknown AUC
CLIP-ViT (2021)	86.9	95.8	83.5	89.0	91.7	97.1
MetaCLIP2 (2025)	86.9	96.1	81.3	88.3	92.7	98.1
DINOv2 (2023)	92.3	97.5	88.1	94.6	95.9	98.4
PE (2026)	94.1	97.5	**91.2**	**95.9**	96.4	98.5
DINOv3 (2025)	**94.5**	**98.7**	90.6	95.6	**97.1**	**99.8**



(a) Scaling training samples.

(b) Robustness evaluation.

Figure 8: Scaling and robustness evaluation. **(a)** Detection error as the number of training samples increases. **(b)** Robustness of different detectors to resizing, Gaussian blur, and video compression.

Component Ablation. We conduct component ablations to assess the contributions of the two branches and the attention modules (Table 8). The full model achieves the best Acc. and AUC on test sets. Specifically, combining the spatial and temporal branches improves performance over either branch alone. Adding patch-wise MHSA to the temporal branch improves the two-branch model. Adding cross-attention improves both metrics on all three test sets.

Backbone and Data Augmentation. We evaluate backbone choices and data augmentation strategies for our detector. Backbone affects the performance ceiling, but the gaps among strong pretrained backbones are relatively limited (Table 9). We study quality compression, photometric transformation, random resized cropping and dynamic resolution in Table 10. Results show that data augmentation improves detection performance under compression, appearance, and resolution variations.

Table 10: Ablation of data augmentation.

Augmentation	2025-H2 Acc.	2025-H2 AUC	2026-H1 Acc.	2026-H1 AUC	Unknown Acc.	Unknown AUC
None	93.5	98.2	89.8	94.8	96.2	99.1
QC	93.9	98.5	90.2	95.1	96.6	99.2
PT	93.7	98.4	90.1	95.0	96.3	99.4
RRC	92.6	97.1	88.5	93.9	95.6	98.9
Dyn.	94.0	98.5	90.2	95.4	96.7	99.5
QC + PT + Dyn.	**94.5**	**98.7**	**90.6**	**95.6**	**97.1**	**99.8**

Scaling and Robustness. Increasing the number of training samples consistently improves detection performance, while the gains gradually diminish as the dataset grows (Figure 8a). Robustness to common video degradation remains an important challenge for detection. To simulate degradation during post-processing and transmission, as shown in Figure 8b, we apply resizing, Gaussian blur and H.264 compression at different levels.

6 Conclusion

In this paper, we introduce AIGV-Annual, a million-scale and temporally partitioned benchmark with a cumulative many-to-many protocol for evaluating AIGV detectors under continual generator evolution. Extensive evaluations of supervised and zero-shot methods reveal substantial cross-dataset generalization gaps, temporal degradation, and varying update gains from newer training data. We further present a spatial-temporal detector that models inter-frame relations and integrates spatial evidence through cross-attention. Extensive experiments show that it outperforms existing methods. We believe that AIGV-Annual provides a solid foundation for advancing robust AIGV detection, while the proposed detector offers a strong baseline for future research.

Limitations and Future Work. The current benchmark focuses on binary detection of fully generated videos. Videos containing localized synthetic edits or mixtures of real and generated content remain outside its scope. Future work will extend the dataset to partial manipulations.

AI USE STATEMENT

In preparing this manuscript, generative AI tools were used to assist with language editing, LaTeX formatting and to generate some character elements used in figures. All AI-assisted changes were reviewed and verified by the authors, who take full responsibility for the final content.

REFERENCES

Zhipu ai official website. https://www.zhipuai.cn/en, 2025a.

Zhixiang ai video generation platform. https://www.zhixiang.ai/, 2025b.

Alibaba Cloud. Wanx. https://tongyi.aliyun.com/wan/, 2024.

Alibaba Cloud. HappyHorse video generation models. Alibaba Cloud Model Studio documentation, 2026. URL https://www.alibabacloud.com/help/en/model-studio/video-generate-edit-model.

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

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

Roy Betser, Meir Yossef Levi, and Guy Gilboa. Whitened clip as a likelihood surrogate of images and captions. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*. PMLR, 2025.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Dollar, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. 38:60884–60937, 2026.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. OpenAI Technical Report, February 2024. URL https://openai.com/index/video-generation-models-as-world-simulators/.

ByteDance. Dreamina ai. https://dreamina.capcut.com, 2024a.

ByteDance. Jimeng ai. https://jimeng.jianying.com, 2024b.

Zijie Cao, Weijie Tu, Yao Xiao, Weijian Deng, Liang Lin, and Pengxu Wei. Where detectors fail: Probing generative space for generalizable ai-generated image detection. *arXiv preprint arXiv:2605.24906*, 2026.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 7621–7639, 2024a.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023a.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7310–7320. IEEE, 2024b.

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

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, and Shouhong Ding. Dual data alignment makes AI-generated image detector easier generalizable. In *Advances in Neural Information Processing Systems*, 2025a.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-Wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13320–13331. IEEE, 2024d.

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

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *ICLR*, 2023b.

Ethan Chern, Hansi Teng, Hanwen Sun, Hao Wang, Hong Pan, Hongyu Jia, Jiadi Su, Jin Li, Junjie Yu, Lijie Liu, et al. Speed by simplicity: A single-stream architecture for fast audio-video generative foundation model. *arXiv preprint arXiv:2603.21986*, 2026.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, Jim Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, Xinlei Chen, Zhuang Liu, Saining Xie, Scott Yih, Shang-Wen Li, and Hu Xu. Meta clip 2: A worldwide scaling recipe. In *Advances in Neural Information Processing Systems*, 2025.

Zijun Cui, Xiulong Liu, Hao Fang, Mingwei Xu, Jiageng Liu, Zexin Xu, Weiguo Pian, Shijian Deng, Feiyu Du, Chenming Ge, and Yapeng Tian. Do joint audio-video generation models understand physics? *arXiv preprint arXiv:2605.07061*, 2026.

Google DeepMind. Veo. `https://deepmind.google/models/veo/`, 2024.

David C Epstein, Ishan Jain, Oliver Wang, and Richard Zhang. Online detection of ai-generated images. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 382–392. IEEE, 2023.

Weichen Fan, Chenyang Si, Junhao Song, Zhenyu Yang, Yinan He, Long Zhuo, Ziqi Huang, Ziyue Dong, Jingwen He, Dongwei Pan, Yi Wang, Yuming Jiang, Yaohui Wang, Peng Gao, Xinyuan Chen, Hengjie Li, Dahua Lin, Yu Qiao, and Ziwei Liu. Vchitect-2.0: Parallel transformer for scaling up video diffusion models. *arXiv preprint arXiv:2501.08453*, 2025.

Huidong Feng, Wentao Chen, Jie Chen, Xinqi Cai, Ruolong Ma, Yinglin Zheng, Yuxin Lin, and Ming Zeng. CoCoVideo: The high-quality commercial-model-based contrastive benchmark for AI-generated video detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11346–11356, 2026.

Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, Xunsong Li, Yifu Li, Shanchuan Lin, Zhijie Lin, Jiawei Liu, Shu Liu, Xiaonan Nie, Zhiwu Qing, Yuxi Ren, Li Sun, Zhi Tian, Rui Wang, Sen Wang, Guoqiang Wei, Guohong Wu, Jie Wu, Ruiqi Xia, Fei Xiao, Xuefeng Xiao, Jiangqiao Yan, Ceyuan Yang, Jianchao Yang, Runkai Yang, Tao Yang, Yihang Yang, Zilyu Ye, Xuejiao Zeng, Yan Zeng, Heng Zhang, Yang Zhao, Xiaozheng Zheng, Peihao Zhu, Jiaxin Zou, and Feilong Zuo. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025.

Genmo. Mochi 1. `https://www.genmo.ai/blog`, 2024.

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

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general ai-generated image detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18685–18694. IEEE, 2025.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2025.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3268–3285, 2026. doi: 10.1109/TPAMI.2025.3633890.

Hugging Face. Hugging face hub documentation. `https://huggingface.co/docs/hub/index`, 2026.

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

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *International Conference on Learning Representations*, volume 2025, pp. 23378–23402, 2025.

JiutianCV. Jt-cv-9b. `https://github.com/jiutiancv`, 2024.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *Advances in Neural Information Processing Systems*, 37:48955–48970, 2024.

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

Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15908–15918. IEEE, 2023.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Tengchuan Kou, Xiaohong Liu, Zicheng Zhang, Chunyi Li, Haoning Wu, Xiongkuo Min, Guangtao Zhai, and Ning Liu. Subjective-aligned dataset and metric for text-to-video quality assessment. *arXiv preprint arXiv:2403.11956*, 2024.

Kuaishou Technology. Kling. `https://klingai.kuaishou.com`, 2024.

Kuaishou Technology. Kling AI 3.0 model series. Official company announcement, February 2026. URL `https://ir.kuaishou.com/news-releases/news-release-details/kling-ai-launches-30-model-ushering-era-where-everyone-can-be/`.

Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, He Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 7241–7259, 2022a.

Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatiotemporal representation learning. *arXiv preprint arXiv:2201.04676*, 2022b.

Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, 2025.

Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4794–4804. IEEE, 2022c.

Yifei Li, Wenzhao Zheng, Yanran Zhang, Runze Sun, Yu Zheng, Lei Chen, Jie Zhou, and Jiwen Lu. Skyra: AI-generated video detection via grounded artifact reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4482–4493, 2026a.

Zhengcen Li, Chenyang Jiang, Hang Zhao, Shiyang Zhou, Yunyang Mo, Feng Gao, Fan Yang, Qiben Shan, Shaocong Wu, and Jingyong Su. Preserving forgery artifacts: Ai-generated video detection at native scale. *arXiv preprint arXiv:2604.04634*, 2026b.

Lightricks. LTX-2.3. Official model repository, 2026. URL `https://github.com/Lightricks/LTX-2`.

Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, Tanghui Jia, Junwu Zhang, Zhenyu Tang, Yatian Pang, Bin She, Cen Yan, Zhiheng Hu, Xiaoyi Dong, Lin Chen, Zhang Pan, Xing Zhou, Shaoling Dong, Yonghong Tian, and Li Yuan. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

Shanchuan Lin and Xiao Yang. Animatediff-lightning: Cross-model diffusion distillation. *arXiv preprint arXiv:2403.12706*, 2024.

Zongyu Lin, Wei Liu, Chen Chen, Jiasen Lu, Wenze Hu, Tsu-Jui Fu, Jesse Allardice, Zhengfeng Lai, Liangchen Song, Bowen Zhang, Cha Chen, Yiran Fei, Lezhi Li, Yinfei Yang, Yizhou Sun, and Kai-Wei Chang. Stiv: Scalable text and image conditioned video generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16249–16259, 2025.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di ZHANG, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback. In *Advances in Neural Information Processing Systems*, 2025a.

Kai Liu, Yanhao Zheng, Kai Wang, Shengqiong Wu, Rongjunchen Zhang, Jiebo Luo, Dimitrios Hatzinakos, Ziwei Liu, Hao Fei, and Tat-Seng Chua. JavisDiT++: Unified modeling and optimization for joint audio-video generation. *arXiv preprint arXiv:2602.19163*, 2026a.

Ruiqi Liu, Manni Cui, Ziheng Qin, Zhiyuan Yan, Ruoxin Chen, Yi Han, Zhiheng Li, Junkai Chen, Zhijin Chen, Kaiqing Lin, Jialiang Shen, Lubin Weng, Jing Dong, Yan Wang, and Shu Wu. MIRROR: Manifold ideal reference reconstructor for generalizable AI-generated image detection. *arXiv preprint arXiv:2602.02222*, 2026b.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation, 2025.

Luma Labs. Luma. `https://dream-machine.lumalabs.ai/`, 2023.

Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, et al. Step-video-t2v technical report: The practice, challenges, and future of video foundation model. *arXiv preprint arXiv:2502.10248*, 2025a.

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

Long Ma, Zihao Xue, Yan Wang, Zhiyuan Yan, Jin Xu, Xiaorui Jiang, Haiyang Yu, Yong Liao, and Zhen Bi. Your one-stop solution for ai-generated video detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4458–4470, June 2026.

Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.

Meitu AI. Miraclevision. `https://www.miraclevision.com/`, 2024.

MiniMax. Minimax video-01. `https://www.minimax.io/news/video-01`, 2024.

Mixkit. Mixkit free stock video. `https://mixkit.co/`, 2023.

Moonvalley. Marey. `https://www.moonvalley.com`, 2025.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024.

Zhenliang Ni, Qiangyu Yan, Mouxiao Huang, Tianning Yuan, Yehui Tang, Hailin Hu, Xinghao Chen, and Yunhe Wang. Genvidbench: A 6-million benchmark for ai-generated video detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-Bench: A Comprehensive Benchmark and Toolkit for Evaluating Video-based Large Language Models. *arXiv preprint arXiv:2311.16103*, 2023.

NVIDIA. NVIDIA Synthetic Video Detector: Model card. NVIDIA NIM, 2026. URL `https://build.nvidia.com/nvidia/synthetic-video-detector/modelcard`. Model version 1.0.0, released April 16, 2026.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24480–24489, 2023.

OpenAI. Sora-2. https://openai.com/zh-Hans-CN/index/sora-2/, 2025.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8245–8257. IEEE, 2025.

Lorenzo Pellegrini, Davide Cozzolino, Serafino Pandolfini, Davide Maltoni, Matteo Ferrara, Luisa Verdoliva, Marco Prati, and Marco Ramilli. Ai-genbench: A new ongoing benchmark for ai-generated image detection. In *2025 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2025.

Pexels. Pexels. https://www.pexels.com, 2024.

Picwand AI. Pixverse prompts guide. https://www.picwand.ai/ai-generation/pixverse-prompts/, 2025.

Pika Labs. Pika art. https://pika.art, 2025.

Pixabay. Pixabay free stock media platform. https://pixabay.com, 2024.

Zhiwu Qing, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yujie Wei, Yingya Zhang, Changxin Gao, and Nong Sang. Hierarchical spatio-temporal decoupling for text-to-video generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6635–6645. IEEE, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Runway Research. Runway. https://runwayml.com, 2023.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

ShengShu Technology. Vidu Q3: Native audio-video generation. Official product documentation, January 2026. URL https://www.vidu.com/vidu-q3.

Chenyang Si, Weichen Fan, Zhengyao Lv, Ziqi Huang, Yu Qiao, and Ziwei Liu. Repvideo: Rethinking cross-layer representation for video generation. *arXiv preprint arXiv:2501.08994*, 2025.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025.

Shangkun Sun, Xiaoyu Liang, Bowen Qu, and Wei Gao. Content-rich aigc video quality assessment via intricate text alignment and motion-aware consistency. *arXiv preprint arXiv:2502.04076*, 2025.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 28130–28139, 2024a.

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

Zhiyu Tan, Xiaomeng Yang, Luozheng Qin, and Hao Li. Vidgen-1m: A large-scale dataset for text-to-video generation. *arXiv preprint arXiv:2408.02629*, 2024b.

Team HunyuanWorld. HY-World 1.5: A systematic framework for interactive world modeling with real-time latency and geometric consistency. Technical report and official repository, 2025. URL https://3d-models.hunyuan.tencent.com/world/world1_5/HYWorld_1.5_Tech_Report.pdf.

ShengShu Technology. Vidu. https://www.vidu.com, 2025.

Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

VideoGenBias. VideoGenBias Data: Official dataset repository, 2026.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Fu-Yun Wang, Zhaoyang Huang, Weikang Bian, Xiaoyu Shi, Keqiang Sun, Guanglu Song, Yu Liu, and Hongsheng Li. Animatelcm: Computation-efficient personalized style video generation without personalized video data. In *SIGGRAPH Asia 2024 Technical Communications*, pp. 1–5. 2024a.

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

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8692–8701, 2020.

Shengyi Wang, Niantong Li, Guangzheng Hu, Hong Qi, Fei Ding, Weixu Qiao, Jinlin Wang, Xiaotong Lv, Peng Han, Zimeng Li, et al. FilmBench: A film-grade benchmark for cinematic video generation. *arXiv preprint arXiv:2607.24241*, 2026.

Wenhao Wang and Yi Yang. VidProM: A Million-scale Real Prompt-Gallery Dataset for Text-to-Video Diffusion Models. *Advances in Neural Information Processing Systems*, 37:65618–65642, 2024.

Wenhao Wang and Yi Yang. Videoufo: A million-scale user-focused dataset for text-to-video generation. *Advances in Neural Information Processing Systems*, 38, 2025.

Xiang Wang, Shiwei Zhang, Hangjie Yuan, Zhiwu Qing, Biao Gong, Yingya Zhang, Yujun Shen, Changxin Gao, and Nong Sang. A recipe for scaling up text-to-video generation with text-free videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 6572–6582, 2024b.

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, 133(5):3059–3078, 2025.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2024, pp. 42055–42079, 2024c.

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

Wondershare. Tomovie. `https://www.tomoviee.ai/`, 2025.

Keming Wu, Yijing Cui, Wenhan Xue, Qijie Wang, Xuan Luo, Zhiyuan Feng, Zuhao Yang, Sudong Wang, Sicong Jiang, Haowei Zhu, et al. WorldReasonBench: Human-aligned stress testing of video generators as future world-state predictors. *arXiv preprint arXiv:2605.10434*, 2026.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pp. 399–417. Springer, 2024.

Jiaqi Xu, Kunzhe Huang, Xinyi Zou, Yunkuo Chen, Bo Liu, Mengli Cheng, Jun Huang, and Xing Shi. Easyanimate: High-performance video generation framework with hybrid windows attention and reward backpropagation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10925–10934, 2025.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.

Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22601–22611, 2023.

Zhucun Xue, Jiangning Zhang, Teng Hu, Haoyang He, Yinan Chen, Yuxuan Cai, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, and Dacheng Tao. Ultravideo: High-quality uhd video dataset with comprehensive captions. *Advances in Neural Information Processing Systems*, 38, 2025.

Jiazhen Yan, Ziqiang Li, Fan Wang, Boyu Wang, Ziwen He, and Zhangjie Fu. DGS-Net: Distillation-guided gradient surgery for CLIP fine-tuning in AI-generated image detection. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable ai-generated image detection. *arXiv preprint arXiv:2411.15633*, 2024.

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

Xiaomeng Yang, Zhiyu Tan, and Hao Li. Ipo: Iterative preference optimization for text-to-video generation. *arXiv preprint arXiv:2502.02088*, 2025a.

Yuhang Yang, Ke Fan, Shangkun Sun, Hongxiang Li, Ailing Zeng, FeiLin Han, Wei Zhai, Wei Liu, Yang Cao, and Zheng-Jun Zha. Videogen-eval: Agent-based system for video generation evaluation. *arXiv preprint arXiv:2503.23452*, 2025b.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025c.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974, 2025.

Kaining Ying, Hengrui Hu, Siyu Ren, Jiamu Li, Fengjiao Chen, Ziwen Wang, Xuezhi Cao, Xunliang Cai, and Henghui Ding. WBench: A comprehensive multi-turn benchmark for interactive video world model evaluation. *arXiv preprint arXiv:2605.25874*, 2026.

Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: Instructing video diffusion models with human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6463–6474, 2024.

Zeroscope AI. Zeroscope ai. `https://zeroscopeai.com/`, 2024.

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

Haiyu Zhang, Xinyuan Chen, Yaohui Wang, Xihui Liu, Yunhong Wang, and Yu Qiao. Accvideo: Accelerating video diffusion model with synthetic dataset. *arXiv preprint arXiv:2503.19462*, 2025b.

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

Shuhai Zhang, Zihao Lian, Jiahao Yang, Daiyuan Li, Guoxuan Pang, Feng Liu, Bo Han, Shutao Li, and Mingkui Tan. Physics-driven spatiotemporal modeling for ai-generated video detection. *Advances in neural information processing systems*, 38:174683–174730, 2025c.

Yiming Zhang, Zhening Xing, Yanhong Zeng, Youqing Fang, and Kai Chen. Pia: Your personalized image animator via plug-and-play modules in text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7747–7756, 2024.

Zicheng Zhang, Tengchuan Kou, Shushi Wang, Chunyi Li, Wei Sun, Wei Wang, Xiaoyu Li, Zongyu Wang, Xuezhi Cao, Xiongkuo Min, et al. Q-eval-100k: Evaluating visual quality and alignment level for text-to-vision content. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10621–10631. IEEE, 2025d.

Chende Zheng, Ruiqi Suo, Chenhao Lin, Zhengyu Zhao, Le Yang, Shuai Liu, Minghui Yang, Cong Wang, and Chao Shen. D3: Training-free AI-generated video detection using second-order features. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1979–1989, 2025a.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. VBench-2.0: Advancing Video Generation Benchmark Suite for Intrinsic Faithfulness. *arXiv preprint arXiv:2503.21755*, 2025b.

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

Zangwei Zheng, Xiangyu Peng, Yuxuan Lou, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, et al. Open-sora 2.0: Training a commercial-level video generation model in $200k$. *arXiv preprint arXiv:2503.09642*, $2025c$.

ZhipuAI. Qingying. `https://chatglm.cn/video?lang=zh`, 2024.

Yuan Zhou, Qiuyue Wang, Yuxuan Cai, and Huan Yang. Allegro: Open the black box of commercial-level video generation model. *arXiv preprint arXiv:2410.15458*, 2024.

Ziwei Zhou, Zeyuan Lai, Rui Wang, Yifan Yang, Zhen Xing, Yuqing Yang, Qi Dai, Lili Qiu, and Chong Luo. AVGen-Bench: A task-driven benchmark for multi-granular evaluation of text-to-audio-video generation. *arXiv preprint arXiv:2604.08540*, 2026.

APPENDIX

A EXPERIMENTAL DETAILS

Implementation Details. For each video, we uniformly sample frames following (Chen et al., 2024c; Zhang et al., 2025c) and randomly select eight consecutive frames from the sampled sequence. The temporal branch resizes the full frames, while the spatial branch uses 224×224 center crops at the original resolution, with zero-padding when necessary. Both branches use ImageNet normalization. We train our detector for 2 epochs using AdamW (Loshchilov & Hutter, 2017) with a learning rate of 1×10^{-5} and a weight decay of 0.1.

B MORE DETAILS OF AIGV-ANNUAL

B.1 DATASET SOURCES AND STATISTICS

Fake Video Sources. *Public benchmark sources.* We collect fake videos from VBench (Huang et al., 2024), VBench++ (Huang et al., 2026), VBench-2.0 (Zheng et al., 2025b), AIGVDBench (Ma et al., 2026), GenVideo-100K (Chen et al., 2024c), GenBuster++ (Wen et al., 2025), VidProM (Wang & Yang, 2024), Q-Eval-100K (Zhang et al., 2025d), Video-Bench (Ning et al., 2023), VideoGen-RewardBench (Liu et al., 2025a), FakePartsBench (Liu et al., 2025b), and CRAVE-DB (Sun et al., 2025). For the 2026-H1 extension, we further incorporate publicly released videos from Film-Bench (Wang et al., 2026), AV-Phys-Bench (Cui et al., 2026), WorldRewardBench (Wu et al., 2026), AVGen-Bench (Zhou et al., 2026), WBench (Ying et al., 2026), and VideoGenBias (VideoGenBias, 2026). Collectively, these datasets provide diverse content, scenes, motion patterns, generation tasks, and generation pipelines.

Community-curated sources. We further collect fake videos from Hugging Face (Hugging Face, 2026) and ModelScope (Wang et al., 2023) to cover recent generators and model versions. To reduce source imbalance, we retain at most 10k videos for each generator–source combination.

Fake videos are grouped into half-year periods according to the release metadata available at collection time. The **Unknown** group contains 35k videos from 10 commercial platforms with undisclosed generator versions or release dates.

Figure 3 summarizes the generator coverage across release periods and architecture families. Tables 11, 12, 13, and 14 provide detailed statistics of the fake video sources.

Real Video Sources. *Training sources.* We collect 581k videos from YouTube, dated from 2006 to 2024-H1, through five curated collections: VideoUFO (Wang & Yang, 2025), OpenVid-HD (Nan et al., 2024), InternVid-10M-FLT (Wang et al., 2024c), Panda-70M (Chen et al., 2024d), and VidGen-1M (Tan et al., 2024b). We additionally include 20k videos from UltraVideo (Xue et al., 2025). Table 15 reports the pooled statistics of these training sources. *Testing sources.* Our testing real pool contains 142k videos from stock-footage platforms, including Mixkit (Mixkit, 2023), Pixabay (Pixabay, 2024), and Pexels (Pexels, 2024), as well as established video benchmarks, including MSR-VTT (Xu et al., 2016), Kinetics-400 (Kay et al., 2017), Youku-mPLUG (Li et al., 2022a), and PVD (Bolya et al., 2026). Table 15 reports the statistics of these testing sources.

Table 11: This table presents detailed statistics of fake video sources released in 2023.

Video source	Model version	Release	Codec	FPS	Duration (s)	Resolution	#Videos
2023-H1: 2 generator entries, 15,317 videos							
Runway (Runway Research, 2023)	Gen-2	2023/02	H.264	24	4	1408×768–2816×1536	5,317
Text2Video-Zero (Khachatryan et al., 2023)	–	2023/03	H.264	4	2	512×512	10,000
2023-H2: 14 generator entries, 107,792 videos							
AnimateDiff (Guo et al., 2023)	AnimateDiff-v1	2023/07	H.264	8	2	512×512	4,650
ZeroScope (Zeroscope AI, 2024)	–	2023/08	H.264, MPEG-4	8–30	1–40	16×576–576×320	10,000
ModelScope-T2V (Wang et al., 2023)	–	2023/08	MPEG-4	8	2	256×256	10,000
VideoCrafter09 (Chen et al., 2023a)	–	2023/08	H.264	8	2	256×256	5,215
AnimateDiff (Guo et al., 2023)	AnimateDiff-v2	2023/09	H.264	8	2	512×512	4,730
LaVie (Wang et al., 2025)	–	2023/09	H.264	8–24	2–3	512×512	10,000
Show-1 (Zhang et al., 2025a)	–	2023/09	H.264	8	4	256×160–576×320	10,000
VideoCrafter1 (Chen et al., 2023a)	–	2023/10	H.264	10	2	512×320	10,000
I2VGen-XL (Zhang et al., 2023)	–	2023/11	H.264	8	2	1280×704	5,590
SEINE (Chen et al., 2023b)	–	2023/11	H.264	8	2–4	512×320–1024×576	10,000
SVD (Blattmann et al., 2023)	–	2023/11	H.264, MPEG-4	7–8	2–4	512×320–1024×576	10,000
InstructVideo (Yuan et al., 2024)	–	2023/12	H.264	8	2	256×256	4,723
PIA (Zhang et al., 2024)	–	2023/12	H.264	4	2–6	512×512	8,169
Pika (Pika Labs, 2025)	Pika 1	2023/12	H.264	24	3	1280×720	4,715

Table 12: This table presents detailed statistics of fake video sources released in 2024.

Video source	Model version	Release	Codec	FPS	Duration (s)	Resolution	#Videos
2024-H1: 16 generator entries, 119,831 videos							
VideoCrafter2 (Chen et al., 2024b)	–	2024/01	H.264	8–10	2–4	512×320–1024×576	10,000
AnimateLCM (Wang et al., 2024a)	–	2024/02	H.264	8	2	512×512	4,730
Sora (Brooks et al., 2024)	–	2024/02	H.264	30	5–60	512×512–1920×1088	9,139
AnimateDiff-L (Lin & Yang, 2024)	–	2024/03	H.264	10	2	512×512	10,000
DynamiCrafter (Xing et al., 2024)	–	2024/03	H.264	8	2	1024×576	10,000
Latte (Ma et al., 2024)	–	2024/03	H.264	8	2	256×256–512×512	10,000
Open-Sora (Zheng et al., 2024)	Open-Sora 1.0	2024/03	H.264	8	2	256×256–512×512	10,000
Open-Sora (Zheng et al., 2024)	Open-Sora 1.1	2024/04	H.264	8	2–8	424×240	5,259
TF-T2V (Wang et al., 2024b)	–	2024/04	H.264	8	4	448×256	4,725
Open-Sora Plan (Lin et al., 2024)	v1.1.0	2024/05	H.264	24	2–9	512×512	5,193
HiGen (Qing et al., 2024)	–	2024/06	H.264	8	4	448×256	4,360
Open-Sora (Zheng et al., 2024)	Open-Sora 1.2	2024/06	H.264	24	4–9	1280×720	10,000
T2V-Turbo (Kou et al., 2024)	–	2024/06	H.264	8	4	512×320	4,685
Kling (Kuaishou Technology, 2024)	Kling 1.0	2024/06	H.264	30	5–50	1088×800–1088×960	3,028
MiracleVision (Meitu AI, 2024)	MiracleVision-v5	2024/06	H.264	24	0–5	720×480	10,000
Runway (Runway Research, 2023)	Gen-3	2024/06	H.264	24–30	5–11	1280×720–1280×768	8,712
2024-H2: 16 generator entries, 97,881 videos							
Mira (Ju et al., 2024)	Mira-v1	2024/07	H.264	6	2–10	384×240	5,028
Open-Sora Plan (Lin et al., 2024)	v1.2.0	2024/07	H.264	24	4	1280×720	4,727
CogVideoX (Yang et al., 2025c)	CogVideoX-2B	2024/08	H.264	8	6	720×480	10,000
CogVideoX (Yang et al., 2025c)	CogVideoX-5B	2024/08	H.264	8	2–6	720×480	10,000
MiniMax (MiniMax, 2024)	–	2024/08	H.264	25	2–6	1280×720	5,608
Kling (Kuaishou Technology, 2024)	Kling 1.5	2024/09	H.264	30	5–10	1088×800–1920×1088	4,521
Allegro (Zhou et al., 2024)	–	2024/10	H.264	15	6	1280×720	1,000
Open-Sora Plan (Lin et al., 2024)	v1.3.0	2024/10	H.264	18	2–5	640×352	5,230
Pyramid-Flow (Jin et al., 2025)	–	2024/10	H.264	24	2–4	640×384	10,000
CogVideoX (Yang et al., 2025c)	CogVideoX1.5-5B	2024/11	H.264	16	10	1360×768	8,580
Mochi-1 (Genmo, 2024)	–	2024/11	H.264	24–30	2–5	848×480	5,338
HunyuanVideo (Kong et al., 2024)	–	2024/12	H.264	24	1–2	960×544	10,000
LTX-Video (HaCohen et al., 2025)	–	2024/12	H.264	24	2–4	768×512	10,000
JT-CV (JiutianCV, 2024)	JT-CV-9B	2024/12	H.264	24	2	2158×1214	3,776
Kling (Kuaishou Technology, 2024)	Kling 1.6	2024/12	H.264	24–30	5–10	1088×800–1920×1088	3,871
Veo (DeepMind, 2024)	Veo 2	2024/12	H.264	24	5–8	1280×720	202

Table 13: This table presents detailed statistics of fake video sources released in 2025.

Video source	Model version	Release	Codec	FPS	Duration (s)	Resolution	#Videos
2025-H1: 20 generator entries, 125,068 videos							
CausVid (Yin et al., 2025)	–	2025/01	H.264	12–24	5–11	640×352	9,440
RepVideo (Si et al., 2025)	–	2025/01	H.264	8	6	720×480	10,000
Vchitect-2.0 (Fan et al., 2025)	Vchitect-2.0-2B	2025/01	H.264	8–16	5	768×432–1920×1080	9,440
Luma (Luma Labs, 2023)	Ray 2	2025/01	H.264	24	5	960×544–1280×720	316
Step-Video (Ma et al., 2025a)	–	2025/02	H.264	25	8	992×544	8,580
Pika (Pika Labs, 2025)	Pika 2.2	2025/02	H.264	24	5	1280×720	316
AccVideo (Zhang et al., 2025b)	–	2025/03	H.264	24	1	960×544	10,000
EasyAnimate (Xu et al., 2025)	EasyAnimate 5.1	2025/03	H.264	8	6	672×384	10,000
Open-Sora (Zheng et al., 2025c)	Open-Sora 2.0	2025/03	H.264	24	5	1024×576	4,720
Wan2.1 (Wan et al., 2025)	T2V-1.3B	2025/03	H.264	16	5	832×480	10,000
Wan2.1 (Wan et al., 2025)	T2V-14B	2025/03	H.264	16	5	1280×720	3,776
Runway (Runway Research, 2023)	Gen-4	2025/03	H.264	24	10	1280×720	5,590
Vidu (Technology, 2025)	Vidu-Q1	2025/04	H.264	24	5	1280×720	8,580
MAGI-1 (Teng et al., 2025)	MAGI-1-4.5B	2025/05	H.264	24	4	720×720	4,720
MAGI-1 (Teng et al., 2025)	MAGI-1-24B	2025/05	H.264	24	4	1280×720	4,720
IPOC (Yang et al., 2025a)	–	2025/05	H.264	8	6	720×480	10,000
IPOW (Yang et al., 2025a)	–	2025/05	H.264	16	5	832×480	4,720
Veo (DeepMind, 2024)	Veo 3	2025/05	H.264	24	8	1280×720	10,000
Kling (Kuaishou Technology, 2024)	Kling 2.1	2025/06	H.264	24	5	1920×1080	77
Seedance (Gao et al., 2025)	Seedance-1-Pro	2025/06	H.264	24	5	1920×1088	73
2025-H2: 9 generator entries, 30,437 videos							
Wan2.2 (Wan et al., 2025)	–	2025/07	H.264	16	5	832×464–768×512	10,000
CogVideoX (Yang et al., 2025c)	CogVideoX-3	2025/07	H.264	8	6	720×480	992
Moonvalley (Moonvalley, 2025)	Marey	2025/07	H.264	24	5	1920×1080	78
STIV (Lin et al., 2025)	–	2025/09	H.264, MPEG-4	10–60	1–2	512×512	7,925
Sora2 (OpenAI, 2025)	–	2025/09	H.264	30	10–15	1280×704–1792×1024	83
Vidu (Technology, 2025)	Vidu-Q2	2025/09	H.264	24	10	1920×1080	50
LTX-Video (HaCohen et al., 2026)	LTX-2	2025/10	H.264	24	5	1536×1024	5,590
ToMovie (Wondershare, 2025)	–	2025/10	H.264	24	5	720×1280	5,590
Veo (DeepMind, 2024)	Veo 3.1	2025/10	H.264	24	6–8	1280×720	129

Table 14: This table presents detailed statistics of fake video sources released in 2026 and those assigned to the Unknown group.

Video source	Model version / dataset	Release	Codec	FPS	Duration (s)	Resolution	#Videos
2026-H1: 18,928 videos							
Vidu (ShengShu Technology, 2026)	Vidu-Q3 Pro	2026/01	H.264	24	15	1920×1080	515
JavisDiT (Liu et al., 2026a)	JavisDiT++	2026/02	H.264	16	5–8	864×480	321
Kling (Kuaishou Technology, 2026)	Kling 3.0	2026/02	H.264, MPEG-4	24	5–15	1224×748–1764×1176	930
Kling (Kuaishou Technology, 2026)	Kling 3.0 Omni	2026/02	H.264	24	8–15	1920×1080	1,490
MOVA (Zhou et al., 2026)	Emu3.5 + MOVA	2026/02	H.264	24	8	640×352	235
MOVA (Zhou et al., 2026)	Nano Banana 2 + MOVA	2026/02	H.264	24	8	640×352	235
Seedance (Seedance et al., 2026)	Seedance-2.0	2026/02	H.264	24–30	4–25	164×124–3840×3840	11,166
LTX-Video (Lightricks, 2026)	LTX-2.3	2026/03	H.264	24	5–10	768×512–1920×1088	984
Sand.ai (Chern et al., 2026)	daVinci-MagiHuman	2026/03	H.264	25	8	1920×1088	321
Ovi (Low et al., 2025)	Ovi 1.1	2026/03	H.264	24	10	1280×704	235
HappyHorse (Alibaba Cloud, 2026)	HappyHorse 1.0	2026/04	H.264	24	15	1920×1080	1,169
HY-World (Team HunyuanWorld, 2025)	HY-World 1.5	2026/05	H.264	24	8–24	832×480	158
HappyHorse (Alibaba Cloud, 2026)	HappyHorse 1.1	2026/06	H.264	24	15	1920×1080	1,169
Unknown: 35,249 videos							
Vidu (Technology, 2025)	–	2024–2025/–	H.264	16	4–8	688×384–1024×576	9,488
Pika (Pika Labs, 2025)	–	2024–2025/–	H.264	8–25	2–5	480×480–1024×1024	10,000
Luma (Luma Labs, 2023)	–	2024–2025/–	H.264	24	5	1360×752	6,058
Dreamina (ByteDance, 2024a)	–	2025/–	H.264	8	3	1280×720	572
PixVerse (Picwand AI, 2025)	–	2025/–	H.264	18	4	1408×768	589
Qingying (ZhipuAI, 2024)	–	2025/–	H.264	16	6	1440×960	979
Tongyi (Alibaba Cloud, 2024)	–	2025/–	H.264	30	5	1028×720–1280×720	364
Zhipu (zhi, 2025a)	–	2025/–	H.264	16	6	1440×960	406
Jimeng (ByteDance, 2024b)	–	2025/–	H.264	8	12	1280×720	6,219
ZhiXiang (zhi, 2025b)	–	2025/–	H.264	24	5	1024×576	574

Table 15: Detailed statistics of the real video sources used in AIGV-Annual. The YouTube pool aggregates videos from VideoUFO (Wang & Yang, 2025), OpenVid-HD (Nan et al., 2024), InternVid-10M-FLT (Wang et al., 2024c), Panda-70M (Chen et al., 2024d), and VidGen-1M (Tan et al., 2024b).

Source	Subset	Year / release	Codec	FPS	Duration (s)	Resolution	#Videos
Training real sources							
YouTube	–	2006–2024-H1	H.264, MPEG-4	6–30	1–582	128×128–1920×1080	581,185
UltraVideo (Xue et al., 2025)	–	2025/06	H.264	12–60	3–10	960×400–7680×4320	20,000
Testing real sources							
Mixkit (Mixkit, 2023)	–	2025/06	H.264, MPEG-4	15–60	4–199	1280×720–1920×1080	10,922
Pixabay (Pixabay, 2024)	–	2025/06	MPEG-4	15–60	6–608	1280×720–4096×2160	2,746
Pexels (Pexels, 2024)	–	2025/04	H.264, MPEG-4	15–60	1–608	568×320–4096×2160	28,588
MSR-VTT (Xu et al., 2016)	–	2016/–	H.264	3	10–31	298×224	10,000
Kinetics-400 (Kay et al., 2017)	–	2017/–	H.264	6–30	1–10	128×88–1280×720	20000
Youku-mPLUG (Li et al., 2022a)	–	2023/–	H.264	24	10	208×160–1088×1920	20,000
PVD (Bolya et al., 2026)	–	2025/04	H.264	24–60	5–61	608×172–608×1152	50,000
						Total	**743,441**

B.2 Dataset Licenses

Fake-video sources.

- **VBench, VBench++, and VBench-2.0**
  - License: Apache License 2.0
  - `https://github.com/Vchitect/VBench`
- **AIGVDBench**
  - License: CC BY 4.0
  - `https://huggingface.co/datasets/AIGVDBench/AIGVDBench`
- **GenVideo-100K**
  - License: CC BY-NC 4.0
  - `https://modelscope.cn/datasets/cccnju/GenVideo-100K`
- **GenBuster++**
  - License: MIT License
  - `https://huggingface.co/datasets/l8cv/GenBuster-Bench-plusplus`
- **VidProM**
  - License: CC BY-NC 4.0
  - `https://huggingface.co/datasets/WenhaoWang/VidProM`
- **Q-Eval-100K**
  - License: CC BY-NC 4.0
  - `https://huggingface.co/datasets/meituan-longcat/Q-Eval-100K`
- **Video-Bench**
  - License: License information is not publicly specified
  - `https://huggingface.co/datasets/Video-Bench/Video-Bench_videos`
- **VideoGen-RewardBench**
  - License: Apache License 2.0
  - `https://huggingface.co/datasets/KlingTeam/VideoGen-RewardBench`
- **FakePartsBench**
  - License: Research and defensive use only
  - `https://huggingface.co/datasets/hi-paris/FakeParts`
- **CRAVE-DB**
  - License: License information is not publicly specified

  - https://huggingface.co/datasets/sunshk/CRAVE-DB
- **Film-Bench**
  - License: Apache License 2.0
  - https://huggingface.co/datasets/skylenage/FilmBench
- **AV-Phys-Bench**
  - License: CC BY 4.0
  - Generated videos: Non-commercial research only
  - https://huggingface.co/datasets/ZijunCui/AV-Phys-Bench
- **WorldRewardBench**
  - License: Non-commercial academic research only
  - https://huggingface.co/datasets/WorldReasonBench/WorldRewardBench
- **AVGen-Bench**
  - License: MIT License
  - https://huggingface.co/datasets/microsoft/AVGen-Bench
- **WBench**
  - License: MIT License
  - https://huggingface.co/datasets/meituan-longcat/WBench
- **VideoGenBias**
  - License: CC BY 4.0
  - Availability: Metadata and evaluation splits will be released upon publication.

Real-video sources.

- **VideoUFO**
  - License: CC BY 4.0
  - https://huggingface.co/datasets/WenhaoWang/VideoUFO
- **OpenVid-HD**
  - License: CC BY 4.0; research and non-commercial purposes only; source licenses also apply
  - https://huggingface.co/datasets/nkp37/OpenVid-1M
- **InternVid-10M-FLT**
  - License: CC BY-NC-SA 4.0
  - https://huggingface.co/datasets/OpenGVLab/InternVid
- **Panda-70M**
  - License: Non-commercial research purposes only; source licenses also apply
  - https://github.com/snap-research/Panda-70M
- **VidGen-1M**
  - License: Apache License 2.0
  - https://huggingface.co/datasets/Fudan-FUXI/VIDGEN-1M
- **UltraVideo**
  - License: CC BY 4.0 with additional restrictions; non-commercial research purposes only
  - https://huggingface.co/datasets/APRIL-AIGC/UltraVideo
- **Mixkit**
  - License: Mixkit Stock Video Free License
  - https://mixkit.co/license/
- **Pixabay**
  - License: Pixabay Content License
  - https://pixabay.com/service/license-summary/
- **Pexels**

  - License: Pexels License
  - https://www.pexels.com/license/
- **MSR-VTT**
  - License: License information is not publicly specified
  - https://www.microsoft.com/en-us/research/publication/msr-vtt-a-large-video-description-dataset-for-bridging-video-and-language/
- **Kinetics-400**
  - License: CC BY 4.0 for annotations; video-specific for videos
  - https://github.com/cvdfoundation/kinetics-dataset
- **Youku-mPLUG**
  - License: Academic research only
  - https://modelscope.cn/datasets/modelscope/Youku-AliceMind/summary
- **PE Video Dataset (PVD)**
  - License: CC BY-NC 4.0
  - https://ai.meta.com/datasets/pe-video/