Advancing Reliable Synthetic Video Detection: Insights from the SAFE Challenge

Kirill Trapeznikov[1], Gabriel Mancino-Ball[1], Jonathan Li[1], Paul Cummer[1], Jai Aslam[1], Danial Samadi Vahdati[2], Tai Nguyen[2], Matthew C. Stamm[2], Peter Bautista[3], Michael Davinroy[3], Laura Cassani[3], Jill Crisman[4]

[1]*{kirill.trapeznikov, gabriel.mancino.ball, jonathan.li, paul.cummer, jai.aslam}@str.us, STR;*
[2]*{ds3729, tdn47, mcs382}@drexel.edu, Drexel University;*
[3]*{pbautista, mdavinroy, lcassani}@aptima.com, Aptima, Inc.;*
[4]*jill.crisman@ul.org, UL Research Institutes*

Abstract

The proliferation of generative video technologies has intensified the need for reliable methods to detect and characterize synthetic media. To address this challenge, we organized the SAFE: Synthetic Video Detection Challenge, co-located with the *Authenticity and Provenance in the Age of Generative AI (APAI) Workshop* at ICCV 2025. The competition invited participants to develop and evaluate algorithms capable of distinguishing real from synthetic videos under fully blind evaluation conditions with over 600 submissions from 12 teams over a 90 day span. Hosted on the Hugging Face platform, the challenge comprised two primary tasks: (1) detection of synthetic video content generated by diverse state-of-the-art models, and (2) detection of synthetic content following common post-processing operations such as resizing, re-compression, motion blur and others. The challenge data consisted of 13 modern high quality synthetic video models with generated content matched to real videos from 21 diverse and challenge sources, all adding up to 20 hours of 6,000 video samples. This paper describes the challenge design, dataset construction, evaluation methodology, and outcomes, offering insights into the generalization and robustness of contemporary synthetic video detection methods. Our findings highlight measurable progress in cross-generator generalization but also persistent vulnerabilities to post-processing artifacts. `https://safe-video-2025.dsri.org`

1. Introduction

The emergence of large-scale generative video models spanning diffusion, transformer, and adversarial architectures, has rapidly blurred the boundary between authentic and synthetic media. Modern text-to-video and image-to-video systems can synthesize human motion, environmental dynamics, and complex scenes with heightened realism, creating both opportunities for creativity and new challenges for verifying media authenticity. As generative systems evolve, they increasingly produce content that defeats traditional semantic inconsistency markers.

Ensuring video authenticity and provenance has therefore become a central problem across journalism, digital safety, fraud prevention, and scientific integrity. Detecting synthetically generated or manipulated videos requires analytics that are robust to visual variability, resilient to post-processing, and generalizable across unseen generation techniques. However, most existing benchmarks provide fixed and openly available datasets, which makes it difficult to keep pace with the rapidly changing generative landscape. These datasets are also typically limited to facial or human-centric content, leaving open questions about robustness across broader semantic domains.

The SAFE: Synthetic Video Detection Challenge was designed to fill this gap. Executed using Hugging Face infrastructure, SAFE introduced a fully blind, model-submission-based framework where all evaluations were run against a sequestered dataset unavailable to participants. This approach ensured fairness and a closer approximation to real-world forensic workflows, where detectors must operate on unseen sources and cannot rely on metadata or file-format cues. The challenge received over 600 submissions from more than twelve teams over a 90-day period. Two tasks were released: *(Task 1) Synthetic Video Detection:* baseline evaluation of generalization across modern high quality generators and diverse real-world content sources; and *(Task 2) Post-Processed Video Detection:* assessment of robustness to realistic transformations such as resizing, compression and other changes common during online dissemination.

Together, these tasks provided a structured, multi-factor test of both core detection capability and operational robustness, offering a view into how current synthetic-video detectors perform under controlled yet realistic blind conditions. This paper makes the following key contributions:

- *Evaluation on realistic synthetic and real content* The challenge evaluated detection performance (Task 1) across 13 modern high quality text+image-to-video (TI2V) and other models with generated content matched to videos from 21 diverse and challenging real sources. The data comprised 20 hours of video across 6,000 samples of content designed to match the distribution of videos found in the wild.
- *A dual-task design measuring generalization and robustness.* While Task 1 established baseline performance, Task 2 simultaneously evaluated post-processing degradations to reveal detector vulnerabilities. Task 2 data consisted of videos with 14 post-

processing techniques commonly applied to videos when shared online (i.e. resizing, recompression, etc.)

- *Fully blind benchmark for synthetic video detection.* Unlike prior open-dataset efforts, SAFE required participants to submit their models to be evaluated on sequestered data using the same computational platform. Reproducing real world settings, participants had no knowledge of video content and sourcing. The competition also reduced the risk of over-training by restricting the number of daily submissions and providing metrics only on the public subset of the data.
- *Comprehensive analysis of state-of-the-art detectors.* This paper reports aggregated findings from top submissions, identifying several key trends such as (i) strong detection performance of unmodified synthetic video content in the fully blind competition setup (ii) non-trivial detection degradation after common post-processing, (iii) emerging good practice for detector designs such as using large pre-trained vision backbones and autoencoding augmentation (encoding input data into a latent representation and subsequently reconstructing the original input via a decoder) and (iv) surprising transferability of detectors to novel synthetic sources.

2. Background and Related Work

Early manipulated media benchmarks primarily targeted faces and controlled capture environments. FaceForensics++ introduced one of the first large-scale benchmarks for facial manipulation detection, encompassing four automated generation methods (Face2Face, FaceSwap, DeepFakes, and NeuralTextures). The study reported pronounced performance degradation under compression, underscoring the sensitivity of detectors to video quality and encoding conditions [1]. Building on this foundation, the DeepFake Detection Challenge (DFDC) dramatically expanded scale and diversity, releasing over 100,000 videos featuring 3,426 paid actors and multiple pipelines to facilitate community-wide benchmarking [2].

Subsequent efforts have focused on improving robustness and representativeness. DeeperForensics-1.0 introduced 50k real and 10k forged videos accompanied by extensive "in-the-wild" perturbations, providing a framework for evaluating models under real-world noise and transformations [3]. Broader demographic coverage was addressed by KoDF, a large-scale dataset focused on Korean subjects with 175k fake and 62k real clips generated by six synthesis models, highlighting distribution shifts relative to prior corpora [4]. Moving beyond highly curated settings, the WildDeepfake dataset collected 7k face sequences from 707 deepfake videos gathered entirely from the internet; baseline detectors trained on curated datasets degrade substantially on this challenging distribution [5].

As image-to-video and text-to-video generators matured, recent work broadened scope beyond face-centric manipulations to general (non-face) synthetic video. The GenVideo benchmark reports large-scale real and generated videos spanning diverse categories and generation techniques to study cross-source generalization in AI-generated video detection [6]. DF40 aggregates content from 40 distinct deepfake techniques to facilitate next-generation cross-technique evaluation at scale [7]. Complementary evaluations such as VANE-Bench probe large multimodal models on anomaly and inconsistency detection (including synthetic videos), illuminating sensitivity to subtle generative artifacts even when the task is framed as question answering rather than binary forensics [8].

Recent datasets have further emphasized open-set, multilingual, and multimodal evaluation. Deepfake-Eval-2024 curates deepfakes from 88 websites across 52 languages, demonstrating substantial performance drops for existing detectors when tested on contemporary content compared to prior academic benchmarks [9]. SocialDF focuses on social-media contexts, compiling 2k real and synthetic videos and proposing multimodal detection approaches aligned with user-generated media verification [10]. MAVOS-DD extends this trend to multilingual, audio–visual deepfake detection, covering 8 languages and over 250 hours of content; results show persistent challenges in generalizing to unseen languages and generator architectures under open-world conditions [11].

Another important evaluation dimension is robustness to post-processing. Videos disseminated online are often compressed, resized, re-encoded, or filtered, masking the forensic traces that detectors rely on. The FaceForensics++ authors specifically highlighted degradation in performance under compression and lower-quality settings [1].

Community analyses around the DFDC effort (e.g., the Partnership on AI report) highlighted deployment-relevant considerations that differ from purely academic settings, motivating evaluations that better reflect operational constraints and real content flows [12]. Collectively, these works motivate the design decisions in SAFE: a blind, model-submission evaluation on hidden data with standardized compute, and task definitions that explicitly measure generalization to unseen generators (Task 1) and robustness to common post-processing operations (Task 2).

Taken together, these findings suggest 3 major challenges: (i) *generalization*: detectors must work across unknown generators and domains, (ii) *robustness*: detectors must tolerate common post-processing and real-world distribution shifts, and (iii) *evaluation realism*: existing benchmarks often release fixed training and test sets that may be overfitted by the research community, reducing their real-world relevance. These challenges motivated the design of our competition.

3. Competition Overview

The SAFE: Synthetic Video Detection Challenge was hosted on the Hugging Face Hub. Running for four

months, the competition ended at a media forensic conference at a major computer vision conference. Its overarching goal was to benchmark and advance methods for detecting AI-generated and manipulated video content under realistic, blind evaluation conditions. The challenge focused on two central aspects of video forensic performance: the ability to generalize across unseen generation models and domains, and the robustness of detection methods to real-world post-processing operations such as compression and resizing. All evaluations were conducted through containerized submissions executed in the same, offline compute environments to ensure fairness and reproducibility.

3.1. Tasks

Our challenge consisted of two synthetic video detection tasks with increasing difficulty.

Task 1 – Detection of Synthetic Video Content focused on identifying video clips generated from various high performing video generation models. No post processing operations were performed on the synthetic video clips. As for the real video clips, audio was stripped and each clip was split into 5-10 second chunks to mirror the synthetically generated clips. A comprehensive list of the 21 video generation sources and 13 video generation models is provided in Tables 1 and 2.

Task 2 – Detection of Synthetic Video Content after Post-Processing assessed the ability to detect synthetic content from post-processed video clips. Fourteen commonly used post processing techniques include upscaling, downscaling, motion blur, etc. A detailed list of the post processing techniques is provided in Table 3.

3.2. Evaluation Criteria

The main metric for the competition was area under curve (AUC), balanced accuracy (BAC), defined as an average of the true positive (TPR) and true negative rates (TNR). The competition maintained a public and private leaderboard for each task where participants had access only to the public leaderboard for the duration of the competition. The public leaderboard provided a breakdown of the evaluation metrics across the entire public dataset as well as across each real and generated source. The names of each source were anonymized on the public leaderboard. This limited feedback and visibility allowed us to maintain a black-box testing scenario. The final evaluation for each task was based on BAC over the private dataset.

The dataset for each task consisted of a public and private set where the public set was a subset of the private dataset. The public dataset consisted of 10 out of 21 real sources and 6 out of 13 generators. The private set contained the entire dataset, including sources which were withheld from the public set. This design choice was made to incentivize participants to develop detectors that did not overfit on the public dataset.

3.3. Setup and Rules

The SAFE: Video Detection Challenge was a script-based, fully-blind competition format where no data was released for the duration of the competition. The competition was hosted through a custom version of the the Hugging Face competition framework [13]. Participants submitted their model by creating a private repository on huggingface.co with no restrictions on what their repo may contain. Participants could have included any code, packages, model weights, or environment within their repository. The only requirement was that participants had to submit a script file that read in our competition dataset and outputted a submission file in a specific format. Participants submitted their model by logging into our competition and authorized our framework to access their model. Our testing framework then evaluated participants' models with the following procedure as shown in Figure A.7: (i) environment is installed, (ii) network access is disabled, (iii) submission model and dataset are pulled and (iv) the evaluation script is executed and the results submission file is uploaded and evaluated.

During the evaluation step, we sequestered the participants' models from the Internet to prevent ex-filtration of the competition data or their potential use of externally hosted services. To level the field, all submissions were executed on the same compute resources of an NVIDIA L40S GPU with 48GB VRAM, 62GB RAM and 8 CPUs. All submissions were limited to 20,000 seconds and participating teams were limited to three submissions per day. Submission logs were automatically analyzed with the Qwen3 480 billion parameter coder model [14] to provide teams with feedback related to errors during their submission process. Since we did not release any data during the competition, we provided several pointers to publicly available open source synthetic datasets to assist participants.

To further simulate real world settings, we required models to make a binary decision of either "real" or "generated" in their submission file. Therefore, the evaluation was performed at a confidence threshold chosen by each participant in each test sample. However, to facilitate further analysis, the submission file also had to contain a decision score and an average inference time for each test data point input file to allow us to compute AUCs.

4. Dataset Description

For the challenge, we constructed a comprehensive dataset consisting of 13 state of the art Text+Image-To-Video (TI2V) models and other generator models and wide range of real videos from 21 different sources, totaling 6,000 samples and 20 hours of video. We designed the synthetic data to be of high quality and representative of what people would likely encounter in the wild, while being semantically aligned with the real videos (i.e. have the same content). For Task 2, we reused a subset of Task 1 data to apply a common set of post-processing operations.

Source	Dur	Res	FPS	Description	Enc	Split
vsp-trail-cam	7.27	0.23	30.00	Wildlife trail cam	H.264	public
vsp-drone	7.37	0.23	29.89	Aerial drone footage	H.264	public
vlog	9.78	6.95	26.52	Vlog content	H.264	public
time-lapse	9.12	3.99	27.05	Time-lapse footage	H.264	public
social-media-text-overlays	7.47	0.90	28.96	Shortform videos with subtitles	H.264	public
social-media-dance	7.51	2.07	30.00	Short form dance videos	H.264	public
social-media-composite-video	7.07	0.89	30.00	Social media content	H.264	public
lecture	7.54	0.23	30.00	Academic lecture recordings	H.264	public
drone	7.11	2.02	30.00	Aerial drone footage	H.264	public
documentary-2	8.66	4.54	30.29	Documentary clips	H.264	public
documentary-1	6.68	2.07	29.97	Documentary clips	H.264	public
vsp-tv-talk-show	6.64	0.23	25.00	TV talk show clips	H.264	private
vsp-travel	6.96	1.76	25.00	Travel and tourism	H.264	private
vsp-selfie	6.96	0.21	29.87	Self-recorded short form content	H.264	private
vsp-production	7.60	0.23	25.00	Broadcast footage	H.264	private
sports	7.07	4.00	30.94	Sports highlights	H.264, MJPEG	private
social-media-phone-camera	7.17	0.88	29.34	Short form content	H.264	private
internet-archive-sports	7.22	0.11	26.92	Sports highlights	H.264	private
internet-archive-pre-digital	7.32	1.28	25.00	Pre-digital age footage	H.264	private
dashcam	7.60	2.16	10.00	Car dashcam footage	MPEG-4	private
academic-selfie	7.66	2.07	50.00	Talking head footage	H.264	private

Table 1: Real Video Sources in the challenge split by public and private sets. Dur(ation) is average duration of video in seconds. Res(olution) is average resolution in megapixels.

Source	Dur	Res	FPS	Description	Enc	OS	Split
wan-2_1-i2v-480p [15]	5.06	0.43	16	Open-source image-to-video model	H.264	Yes	public
veo-2 [16]	5.00	0.92	24	High-quality cinematic generation	H.264	Yes	public
pixverse-v4_5 [17]	5.37	0.61	30	General-purpose video generator	H.264	No	public
kling-v2_0 [18]	5.04	0.92	24	Commercial text-to-video model	H.264	No	public
hunyuan [19]	5.37	0.86	30	Open-source high-resolution generator	H.264	Yes	public
framepack [20]	5.01	0.92	30	Lightweight long-video generation framework	H.264	Yes	public
vidu-2.0 [21]	4.03	0.92	32	Stylized and animated video synthesis	H.264	No	private
video-01 [22]	5.64	0.90	25	Photorealistic cinematic animation	H.264	No	private
seedance-1-pro [23]	5.04	2.09	24	High-resolution multi-shot generation	H.264	No	private
runway_gen4_turbo [24]	10.08	0.92	24	Professional cinematic video generation	H.264	No	private
physicalai [25] [26] [27]	10.00	2.07	30	Multi-view synthetic human activity dataset	H.264	Yes	private
hunyuan-avatar [28]	5.15	0.62	25	Audio-driven human animation	H.264	Yes	private
cosmos-drive-dreams [29]	5.04	0.90	24	Synthetic driving scene generation	H.264	Yes	private

Table 2: Synthetic Video Sources in the challenge split by public and private sets. Dur(ation) is average duration of video in seconds. Res(olution) is average resoltion in megapixels.

4.1. Real Video Sourcing

The real videos in Task 1 of the challenge were curated from 21 diverse sources, encompassing a broad spectrum of content including recent videos such as those found on social media, drone footage, sports, time lapse videos, smartphone footage, and shows of various production quality available on video sharing platforms (VSPs). We designed the dataset to be as representative of the real distribution videos found on the Internet, so we sourced from a wide variety of platforms such as VSP, social media, the Internet archive, etc. To make the data more challenging, we included more obscure content such as pre-digital age videos, trail cam footage, dashcam videos, and academic selfie data. We also included altered content such as video with text overlays and videos that were a composite of multiple shots (all very common on the Internet). The

Technique	Description	Parameters
border	Add padded black border	width = min video dim*0.1
color correction	Adjusts colors using LUTs	—
compr. libaom-av1	AV1 Compression	crf=10-50
compr. libx264	H.264 Compression	crf=10-40
compr. libx265	H.265 Compression	crf=10-40
upscale	increase res by 75%, center-crop to o.g. res	—
downscale	decrease resolution by 75%	—
frame interpolation	increased fps using minterpolate algorithm	fps=[16,24,30,60,2*current fps]
gaussian blur	gaussian blur with sigma $\propto$ video res	ref sigma=3 ref res=1080p
motion blur	Temporally averaging consecutive frames	—
noise	Adds gaussian noise	target PSNR=36dB
real camera	Recording of video using webcam	—
speed adjustment	Alters playback speed	scale factor $\in$ [0.25, 2.0]
text overlay	Superimpose text captions onto video	font size=min video dimension * 0.05

Table 3: Post-processing and augmentation techniques used in Task 2, descriptions and parameters (if any)

.

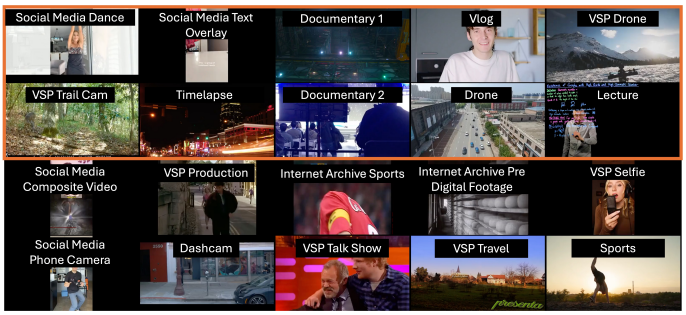

Figure 1: Real Video Sources: a sample frame from each video source is shown. The top examples within the orange outline were in the public split and the remaining examples were in the private.

dataset comprised 1,064 real videos with approximately 50 videos per source. We partitioned the data such that 10 sources were allocated to the public data split, with the remaining 11 sources reserved for the private evaluation split. The video sources, description, and partition are listed in Table 1. To ensure consistency with the synthetically generated videos, all real videos were processed by removing audio content and truncating to segments of 5-10 seconds in duration. The spatial resolutions of these videos varied considerably, ranging from 240p to 4K. Figure 1 shows a sample of frames per each source with the orange outlined frames indicating the public set.

4.2. Synthetic Video Generation

In constructing the synthetic video data, we focused on (i) replicating commonly used video generation tools used in the wild by including high quality, readily available closed and open source models, and (ii) matching the content of the real videos by seeding the generation with a starting frame and descriptions derived from real content.

Generators: In Task 1, we utilized 13 state-of-the-art models, including 10 diffusion based TI2V, one Audio-Driven model (Hunyuan-Avatar), and two physical environment simulation models (PhysicalAI-Autonomous-Vehicle-Cosmos-Drive-Dreams, PhysicalAI-SmartSpaces). 100 image-text pairs served as inputs to each of the 10 generative models to create 1000 synthetically generated videos with a distribution of content similar to the real videos (see details on our video generation pipeline below). Out of the 10 model sources, 6 were used for the public dataset, and 4 were used in the private split. The private dataset was supplemented with 100 videos generated using audio driven Hunyuan-Avatar, 100 synthetic dash cam videos sampled from PhysicalAI-Autonomous-Vehicle-Cosmos-Drive-Dreams, and 100 synthetic indoor workspace videos sampled from PhysicalAI-SmartSpaces. Each video source used in the synthetically generated portion of the dataset is listed in Table 2. After generation, the videos were left as is with no modifications or post processing. The generated videos varied in duration from 5 to 10 seconds and in resolution from 360p to 4K. Figure 2 shows a sample of frames per each source with the orange outlined box indicating the public set.

Video Generation Pipeline: To align the content of synthetic and real videos, we developed an automated generation pipeline shown Figure 3. TI2V models require a starting image and a text description of the desired content for video generation. To ensure that the distribution of content in generated videos matched that of real videos, we designed the following process. We uniformly sampled 100 videos across all categories from the real video dataset. For each video, we extracted a clear and semantically meaningful frame by instructing Qwen2.5-VL [30] to identify event boundaries within the video, where an event is defined as a segment containing a unique action. For each identified event, Qwen2.5-VL selected one representative frame, then performed a quality assessment based on clarity and ab-

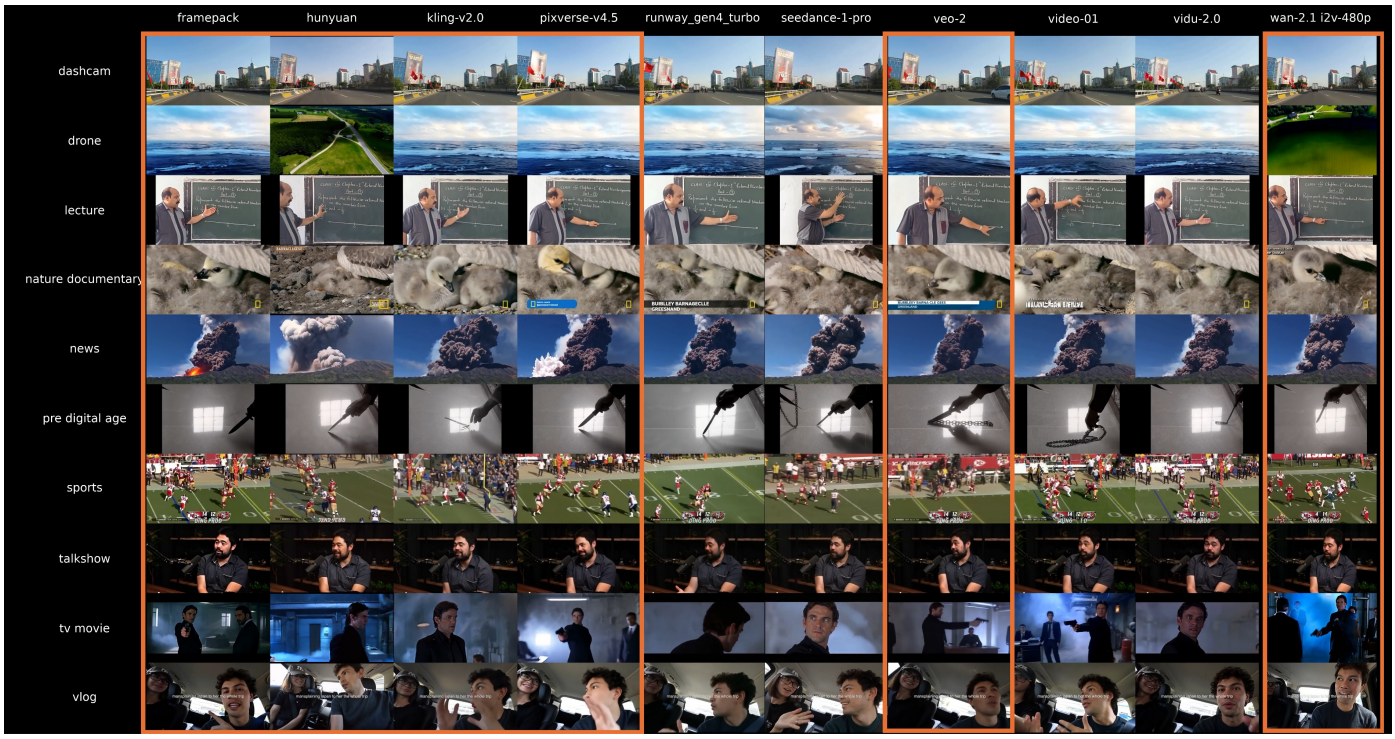

Figure 2: Synthetic Video Sources: 10 different frame samples matching the content of the real videos for each TI2V generator (columns). The orange outlined sources are in the public split and the rest are in private.

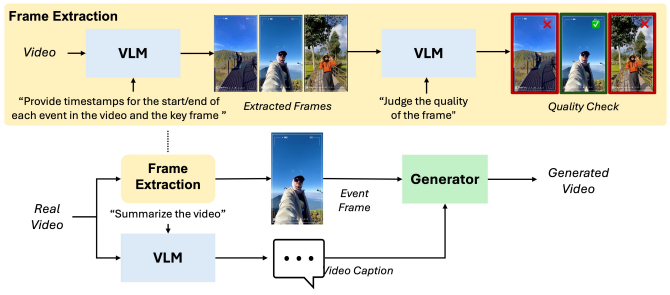

Figure 3: Generation pipeline used for creating the dataset that matches real video content distribution.

Figure 4: A sample of frames showing the visual quality per augmentation applied in Task 2

.

Team Name	Organization
GRIP-UNINA	University of Naples Federico II
ISPL-Realynx	ISPL, Politecnico di Milano
DASH	DASH Lab / Sungkyunkwan University
LEmma1727	DASH Lab / Sungkyunkwan University
TRustbees	Truebees
SHahidmuneer	DASH Lab / Sungkyunkwan University
DX	DX Lab / Sungkyunkwan University
BAseline	MISL / Drexel University

Table 4: Team names and institutional affiliations for the top participating teams in the SAFE Video Detection Challenge.

sence of artifacts, ultimately selecting the frame it deemed best matched the criteria as input for video generation. For the text prompt, we instructed Qwen2.5-VL to generate a detailed description of each video's content. This method produced 100 text-image pairs matching the category distribution of real videos, which served as inputs for the text/image-to-video models. By using identical text-image input pairs across all generators, we isolated generator performance from content variability, enabling more accurate comparative evaluation.

Post-Processing: To test robustness of detectors to post-processing, in Task 2 we used 14 post-processing techniques that are commonly performed on videos found on the Internet. These techniques included resolution manipulation with upscaling and downscaling (aspect ratio fixed; upscale = height x 1.75; downscale = height x .75), and

Metric	Task	GR	IS	LE	DA	TR	*BA*
BAC	1	**0.86**	0.85	0.82	0.84	0.77	0.68
	2	**0.74**	0.70	0.70	0.73	0.66	0.59
AUC	1	**0.93**	0.92	0.88	0.88	0.85	0.74
	2	**0.82**	0.77	0.77	0.79	0.72	0.64
TPR	1	0.84	**0.86**	0.81	0.77	0.66	0.68
	2	0.66	**0.82**	0.74	0.71	0.59	0.44
TNR	1	0.88	0.84	0.83	**0.90**	0.88	0.68
	2	**0.82**	0.59	0.66	0.75	0.73	0.75

Table 5: Detection performance (on the full private set) with a comparison of the two tasks, the top teams and the baseline across four metrics: balanced accuracy (BAC), area under the curve (AUC), true positive rate (TPR), and true negative rate (TNR)

.

common compression algorithms: H.264, H.265, and AV1. We also included other techniques such as adding a border, color correction, frame interpolation, gaussian blur, motion blur, noise, speed adjustment, and text overlaying. Finally, we included one laundering technique ("real camera") where we recorded the video playing on a screen with another camera. Each technique along with the description is listed in Table 3. Videos from Task 1 were uniformly sampled from each source, resulting in a total of 78 synthetic videos and 84 real videos. Each post-processing technique was applied to every video, for a total of 1092 synthetic videos and 1176 real videos. The partitioning of Task 2 videos to either public or private datasets corresponded to the partitioning of their original videos as used in Task 1. Note that the augmentation did not result in noticeable quality degradation or the presence of artifacts (see Figure 4) thus qualifying as reasonable processing that one could do to videos before sharing.

5. Results

Here we present an overview of the results from the SAFE Video Detection Challenge[1]. The challenge ran for 90 days and attracted participation from over 12 teams worldwide, resulting in more than 600 model submissions across both tasks. To contextualize leaderboard results, Table 4 maps the Hugging Face submission handles used during evaluation to the corresponding institutions for top performing teams that presented results at the workshop. Due to space limitations, result tables show only the top 5 teams (abbreviated by the first two letters). The charts show additional teams. Also note that we have incomplete details of participants' approaches since they were not required to share any specific details (see Table 9). For a baseline (BA), we used a synthetic video detector from [31] that has been trained on several older synthetic video models such Stable Video Diffusion and COG.

[1]Competition details: `https://safe-video-2025.dsri.org`

Source	μ	GR	IS	DA	LE	TR
`vsp-trail-cam`	0.97	0.98	**1.00**	0.94	0.93	0.99
`vsp-drone`	0.91	**0.99**	**0.99**	0.84	0.84	0.92
`social-media-dance`	0.91	**0.98**	0.92	0.90	0.87	0.90
`drone`	0.91	0.95	**0.97**	0.83	0.83	**0.97**
`social-media-text-overlays`	0.91	**0.97**	0.90	0.92	0.86	0.88
`lecture`	0.90	0.96	**0.98**	0.94	0.88	0.80
`vlog`	0.90	0.95	**0.99**	0.92	0.86	0.90
`social-media-composite-video`	0.90	**0.97**	**0.97**	0.89	0.88	0.89
`documentary`	0.90	**0.97**	0.95	0.88	0.86	0.89
`time-lapse`	0.90	0.95	**0.96**	0.91	0.88	0.84
`Public Avg`	0.89	0.92	**0.94**	0.87	0.87	0.86
`academic-selfie`	0.90	**1.00**	0.97	0.85	0.87	0.91
`vsp-selfie`	0.90	**1.00**	0.96	0.85	0.84	0.88
`vsp-travel`	0.90	**0.99**	0.98	0.83	0.88	0.88
`vsp-production`	0.90	**0.98**	**0.98**	0.83	0.88	0.87
`social-media-phone-camera`	0.89	**0.98**	0.96	0.88	0.86	0.91
`internet-archive-sports`	0.88	0.87	**0.98**	0.94	0.88	0.83
`internet-archive-pre-digital-footage`	0.85	0.85	**0.99**	0.92	0.87	0.61
`sports`	0.85	0.87	**0.91**	0.81	0.86	0.84
`vsp-tv-talk-show`	0.80	0.85	**0.92**	0.82	0.84	0.90
`dashcam`	0.78	0.88	0.49	**0.95**	0.90	0.68
`Private Avg`	0.89	**0.95**	0.90	0.88	0.87	0.85

Table 6: AUC conditioned on the real source. Only the top five teams and their means are shown for brevity.

From the competition results, we observe several important takeaways. Table 5 shows performance across the top teams, as well as several metrics across the two tasks.

Strong Performance of Unmodified Synthetic Videos: Performance on detection of unmodified synthetic video (Task 1) was notably strong. The top teams, **GR**IP-UNINA, **IS**PL-Realynx, **DA**SH, and **LE**mma1727 achieved an AUC of 0.93, 0.92, 0.88, and 0.88, respectively, substantially outperforming the baseline model (0.74). Generalization of submissions on both private and public splits was surprising given that the participants had no

Source	μ	GR	IS	DA	LE	TR
veo-2	0.97	**1.00**	0.95	0.99	0.99	0.94
wan-2_1-i2v-480p	0.97	**1.00**	0.99	0.98	0.97	0.93
pixverse-v4_5	0.96	**0.99**	0.95	0.97	0.98	0.89
kling-v2_0	0.95	**0.99**	0.95	0.91	0.95	0.94
hunyuan	0.94	**0.99**	0.92	0.95	0.95	0.90
framepack	0.85	**0.93**	0.81	0.92	0.92	0.77
Public Avg	0.94	**0.98**	0.93	0.95	0.96	0.89
runway-gen4-turbo	0.98	**0.99**	0.98	0.99	0.99	0.95
cosmos-drive-dreams	0.92	**0.94**	0.92	0.88	0.86	0.88
vidu-2.0	0.88	0.84	**0.90**	0.81	0.77	0.81
video-01	0.81	0.87	**0.96**	0.62	0.62	0.70
hunyuan-avatar	0.79	0.92	**0.95**	0.62	0.69	0.61
seedance-1-pro	0.77	0.83	**0.87**	0.78	0.69	0.90
physicalai	0.75	0.87	0.77	**0.98**	0.72	0.82
Private Avg	0.84	0.89	**0.91**	0.81	0.80	0.81

Table 7: AUC conditioned on the generator. Only the top five teams and their means are shown for brevity.

knowledge of synthetic and real sources in either split.

Detectors are Not Robust to Post-processed Content: Across all teams, performance was consistently lower on Task 2 (detection after post-processing). This suggests that even the best techniques are still not robust to common operations that videos undergo when shared across social media and the Internet.

Pre-trained Image Backbones and Autoencoding for Training Good Detectors: Top performers shared several important aspects in their model training and architecture. First, all top performers utilized a strong backbone (DINOv2) specifically designed for general purpose computer vision tasks (object detection, segmentation, etc.) and pre-trained on large amount of real images. Second, most performers incorporated autoencoding augmentation during their training. A real image or video is reconstructed using the VAE creating a synthetic example perfectly matched to a real one w.r.t. content. Table 9 shows different aspects of participants' approaches [2]

Surprising Generalizability to Novel Generators: Table 9 shows that the top performing teams had little or no overlap between the challenge and training data while still achieving very high performance. This suggests a significant degree of transferability to unknown detectors. For example, **DA**SH only used a single model (WAN2.2) for training while achieving good performance across the 12 other models in our test set. This suggests that the architecture and the training setup is just as important as having a wide coverage of a diverse array of generators.

[2]Not all teams shared approach details with the organizers.

Augmentation	μ	GR	DA	IS	LE	TR
none	0.88	**0.93**	0.87	0.90	0.84	0.84
speed-adjustment	0.83	0.87	0.84	**0.89**	0.82	0.75
upscale	0.83	**0.89**	0.82	0.83	0.80	0.81
text-overlay	0.83	0.86	0.83	**0.89**	0.82	0.75
border	0.82	**0.88**	0.80	0.85	0.79	0.76
frame-interpolation	0.82	0.85	0.83	**0.86**	0.81	0.73
motion-blur	0.81	0.84	0.82	**0.86**	0.80	0.72
compression-libx265	0.79	**0.87**	0.79	0.83	0.74	0.71
downscale	0.79	0.81	**0.82**	0.81	0.76	0.72
color-correction	0.78	**0.85**	0.79	0.77	0.78	0.69
compression-libaom-av1	0.77	**0.85**	0.76	0.74	0.75	0.75
compression-libx264	0.77	0.75	**0.81**	0.75	**0.81**	0.70
gaussian-blur	0.74	0.73	0.75	**0.79**	0.76	0.67
noise	0.72	**0.83**	0.75	0.63	0.68	0.71
camera	0.62	0.66	**0.67**	0.51	0.63	0.63

Table 8: Task 2 Results: AUC conditioned on augmentation (for the top team and their means) are shown. The rows are sorted by decreasing AUC. The "none" row corresponds to unmodified videos.

5.1. Task 1 Discussion

Tables 6 and 7 show AUC conditioned on real sources and video generators, respectively. During the competition, participants had access to these fine-grained performance metrics but only on the public split of the data (the top half of the tables). However, the names of the sources remained anonymous. Performance on the private set remained hidden until the end of the competition. Even under these restrictions, the results did not show any large performance gaps between private and public sources.

In Table 7 showing performance conditioned on the video generator, the mean AUCs (μ) of top teams was lower on average between public (.94) and private (.84). The generators (hunyan-avatar, seedance-1-pro, physicalai) had the lowest performance in the mid to upper 70's compared to scores in the 90's for most generators in the public split. This is likely due to hunyan-avatar being an audio+text+image to video model and seedance-1-pro being a newer generator. The physicalai data is a CGI-like physical motion generator model that is somewhat different from standard I2V models. For details on different synthetic video sources see Table 2.

In Table 6, the mean AUCs conditioned on real sources have the same averages for private and public splits among

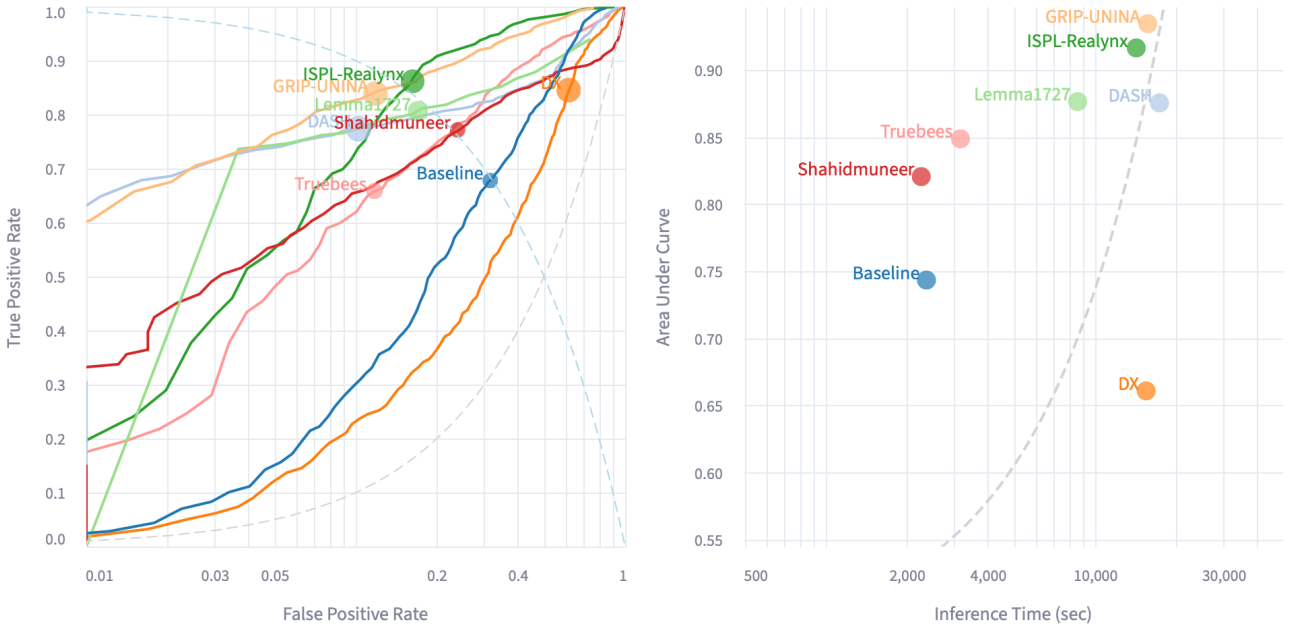

Figure 5: Task 1 Results: [Left] ROC Curve for all teams (with FPR plotted on a log scale. [Right] AUC vs Inference time (on a log scale)

Team	Synthetic	Real	Overlap	Augmentation	Architecture
BA [31]	Video Crafter, Luma, SVD, CogVideo	VideoACID, Moments in Time	None	..	MISLNet
GR [32]	Pyramid Flow, COG Video X, Open Sora, Allegro	..	No Synthetic	autoencoding, wavelet replacement	Ensemble w. DINOv2/3 backbone, Frame Average
IS	45+ different generators	variety of video frames, images	Limited	Autoencoding, Other	EfficientNet, Frame Average
DA/LE	Wan2.2	2k HQ Youtube Vids	Only WAN2.2	Autoencoding, Real Frame to Video	DINOv2 Backbone + Transformer over Frames

Table 9: Overview of training strategies, data augmentation, and architectures by team. Note that this table is incomplete since teams were not required to share the details of their approach. (The DASH and LEmma teams had very similar approaches).

the top teams. The only outliers appear to be the "dash-cam" and the "tv-talk-show" source (with .78 and .80 AUCs). The former is both semantically and forensically different than other the sources which could explain performance drops.

Figure 5 shows the ROC curves for the best submission. Note the x-axis (False Alarm Rate) is on the log scale. A large fraction of submissions fell on the equal error rate curve suggesting balanced true and false positive rates. However, none of the detectors can operate in very low false alarm regimes (<.01). Therefore, while the overall performance is promising, we are still far from being able to handle large (i.e. Internet) scale detection of synthetic videos without being flooded with false alarms. The right panel shows AUC vs inference time (on a log scale). Since all submissions used the same underlying hardware, this figure reflects realistic inference times. The main trend is that a longer inference time (likely implying a model with more parameters) improves detection. However, we see, for example, that the Lemma and DX team achieved the same AUC with the Lemma being significantly faster.

5.2. Task 2 Discussion

Recall that in Task 2, we applied a set of augmentation and post-processing operations to the subset of video data from Task 1 to simulate common operations that occur when sharing videos on the Internet. Table 8 shows the AUC conditioned on the augmentation type for the top submissions (sorted by the mean μ performance). The main observation is that all augmentations reduced the detection accuracy by 5 to 25 AUC points when compared

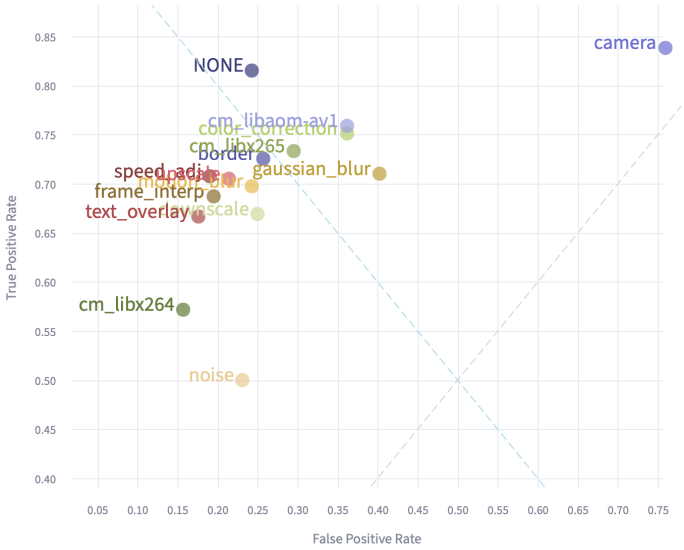

Figure 6: True vs False Positive Rates for each augmentation type in Task 2 for the best GR are shown. Details are in Table 3.

to unaltered video (the "none" category in the table [3]).

In Figure 4 we also see that visually most of the augmentations applied are very mild in terms of reducing visual quality or adding semantic artifacts. Of note is the significant drop in performance due to video re-compression with "libx264 and libx265, and av1" reducing performance from .88 to .77, for libx264 and av1 and to .79 for libx265. This is likely due to the additional corruption of forensic traces present in the video. Gaussian blurring and white noise augmentations resulted in average AUC scores of .74 and .72, respectively, likely for the same reason. The last and most difficult augmentation was the laundering "camera" process where we recorded a video playing on a screen with another camera as a digital-to-analog-to-digital operation thus the drop to .62 from .88 is not surprising.

Figure 6 shows a scatter plot of the average true and false positive rates color coded by the augmentation type. We observed that different augmentations resulted in different error types. For example, some operations such as "noise" only degraded true positive rates (which makes sense since real videos have naturally occurring noise due to the physical process of imaging the real world while synthetic videos do not). Some degraded both error types such as gaussian blur. Surprisingly, the laundering "camera" augmentation increased false alarm rates, indicating that at the detectors' operating points, laundered real videos shift closer in score to synthetic videos, contrary to our expectation that the opposite effect would occur.

5.3. Discussion of Participants' Approaches

Table 9 contains partial details on submission approaches including training data used, augmentation techniques applied, and architecture of detection models. Participants were not required to provide all details of their approach and, by design, the competition did not have access to the submission code or models; therefore, the table covers only a subset of teams. As mentioned earlier, we observed interesting patterns across the top submissions. A few submissions such as **GR**IP-UNINA and **DA**SH had almost no overlap with competition in their training while achieving very good detection accuracy suggesting strong transferability between different synthetic video generators. The use of autoencoding on real video images to generate synthetic training examples emerged as an effective augmentation strategy for improving model performance. Lastly, starting from a large scale general purpose computer vision backbone such DinoV2 that had been trained on large amounts of real images was considered an important component for training synthetic image or video detectors.

[3]For real video, we do not know the complete history of processing.

6. Acknowledgments

This work and the competition are sponsored by ULRI Digital Safety Research Institute.

References

[1] A. Rössler, et al., Faceforensics++: Learning to detect manipulated facial images, in: Proc. ICCV, 2019, pp. 1–11.

[2] B. Dolhansky, et al., The deepfake detection challenge (dfdc) dataset, arXiv (2020). arXiv:2006.07397.

[3] L. Jiang, W. Wu, C. C. Li, C. Qian, C. C. Loy, Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection, in: Proc. CVPR, 2020, pp. 2889–2898.

[4] P. Kwon, et al., Kodf: A large-scale korean deepfake detection dataset, in: Proc. ICCV, 2021, pp. 10744–10753.

[5] B. Zi, Z. Song, K. Zhang, W. Luo, Wilddeepfake: A challenging real-world dataset for deepfake detection, arXiv (2021). arXiv:2101.01456.

[6] H. Chen, X. Hong, et al., Demamba: Ai-generated video detection on million-scale genvideo benchmark, arXiv (2024). arXiv:2405.19707.

[7] Z. Yan, et al., Df40: Toward next-generation deepfake detection, in: NeurIPS Datasets and Benchmarks, 2024.

[8] R. Bharadwaj, et al., Vane-bench: Video anomaly evaluation benchmark for conversational lmms, arXiv (2024). arXiv:2406.10326.

[9] N. A. Chandra, et al., Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024, arXiv (2025). arXiv:2503.02857.

[10] A. Batra, S. Kumar, et al., Socialdf: Benchmark dataset and detection model for mitigating deepfakes on social media, arXiv (2025). arXiv:2506.05538.

[11] F.-A. Croitoru, et al., Mavos-dd: Multilingual audio-video open-set deepfake detection benchmark, arXiv (2025). arXiv:2505.11109.

[12] Partnership on AI, The deepfake detection challenge: Insights and recommendations for ai and media integrity, Tech. rep., Partnership on AI, accessed 2025 (2020).

[13] Hugging face competitions (2025).
URL https://huggingface.co/docs/competitions/en/index

[14] A. Yang, et al., Qwen3 technical report, arXiv (2025). arXiv:2505.09388.

[15] Team Wan, et al., Wan: Open and advanced large-scale video generative models, arXiv (2025). arXiv:2503.20314.

[16] Veo 2 text+image-to-video generator (2024).
URL https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/veo/2-0-generate

[17] Pixverse 4.5 text+image-to-video generator (2025).
URL https://platform.pixverse.ai/onboard

[18] Kling 2.0 text+image-to-video generator (2025).
URL https://app.klingai.com/

[19] W. Kong, et al., Hunyuanvideo: A systematic framework for large video generative models, arXiv (2024). arXiv:2412.03603.

[20] L. Zhang, et al., Frame context packing and drift prevention in next-frame-prediction video diffusion models, in: NeurIPS, 2025.

[21] F. Bao, et al., Vidu: A highly consistent, dynamic and skilled text-to-video generator with diffusion models, arXiv (2024). arXiv:2405.04233.

[22] Minimax video-01 text+image-to-video generator (2024).
URL https://platform.minimax.io

[23] Y. Gao, et al., Seedance 1.0: Exploring the boundaries of video generation models, arXiv (2025). arXiv:2506.09113.

[24] Runway gen-4.0 turbo text+image-to-video generator (2025).
URL https://docs.dev.runwayml.com/api/

[25] Z. Tang, et al., The 9th ai city challenge, in: Proc. ICCV Workshops, 2025, pp. 5467–5476.

[26] S. Wang, et al., The 8th ai city challenge, in: Proc. CVPR Workshops, 2024, pp. 7261–7272.

[27] Y. Wang, et al., Mcblt: Multi-camera multi-object 3d tracking in long videos, arXiv (2024). arXiv:2412.00692.

[28] Y. Chen, et al., Hunyuanvideo-avatar: High-fidelity audio-driven human animation for multiple characters, arXiv (2025). arXiv:2505.20156.

[29] X. Ren, et al., Cosmos-drive-dreams: Scalable synthetic driving data generation with world foundation models, arXiv (2025). arXiv:2506.09042.

[30] J. Bai, et al., Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, arXiv (2023). arXiv:2308.12966.

[31] D. S. Vahdati, et al., Beyond deepfake images: Detecting ai-generated videos, in: CVPRW, 2024, pp. 4397–4408.

[32] R. Corvi, et al., Seeing what matters: Generalizable ai-generated video detection with forensic-oriented augmentation, in: NeurIPS, 2025.

Appendix A. Additional Competition Details

Additional details about the competition are provided below.

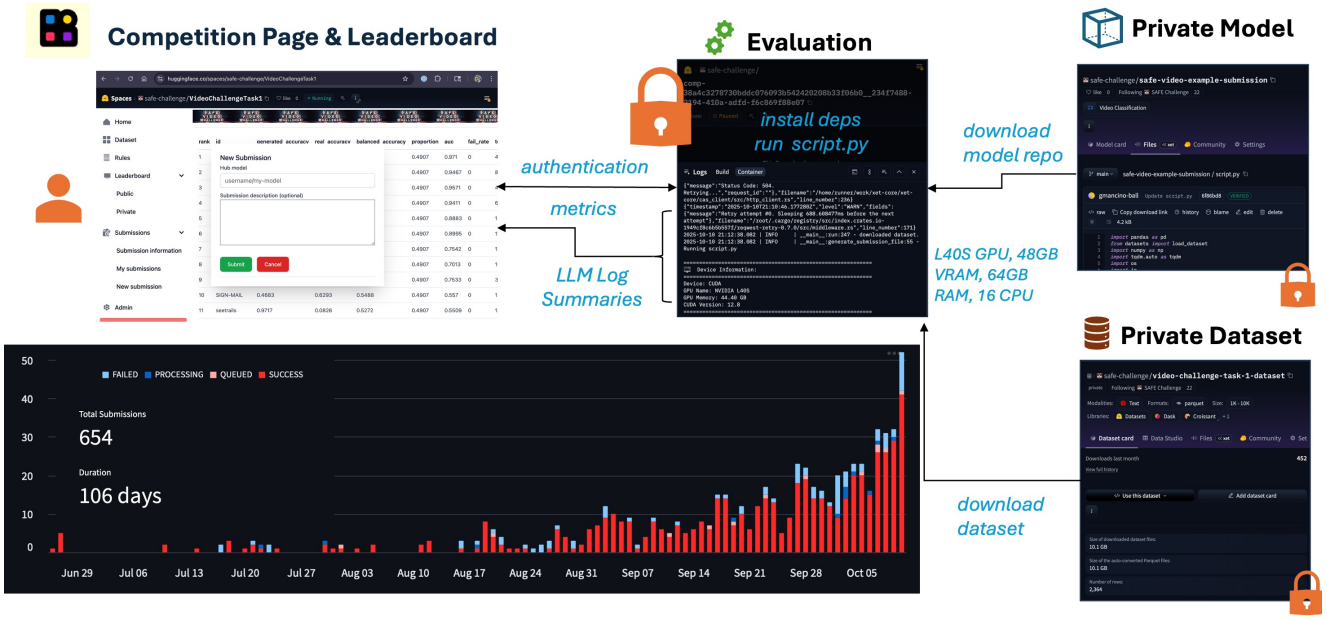

Figure A.7: Competition Setup: participants log in with their Hugging Face credentials and provide the submission repo. In the backed, our platform pulls the model, installs the requirements, pulls the private dataset and runs the evaluation without any network access. The bottom left shows the volume of our competition submissions, which spiked a month before the end of the challenge.