UNIVERSAL REPRESENTATIONS FOR ROBUST AI-GENERATED VIDEO DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Methods for detecting AI-generated video fall into three long-standing categories: low-level analysis of generation artifacts, high-level tests of physical plausibility, and learned classifiers over deep embeddings. Detectors in each category generalize poorly beyond their training distribution, and, we argue, published comparisons between them are confounded by backbone size discrepancies. We present, to our knowledge, the first architecture that explicitly unifies all three categories of cues—density-based anomaly scores over frozen encoder features, fused with spectral and motion-coherence streams—and name it SEAM. Trained on a single generator from AIGVDBench, SEAM reaches 0.92 AUC on 19 unseen open-source generators, beating the state-of-the-art (SOTA) performance, and matching SOTA performance (0.87 AUC) on 11 unseen commercial generators, including Sora, Kling, and Pika. We further introduce SLOP, a dataset of 6,000 real and AI-generated videos from Instagram, TikTok, and X, and a taxonomy of such content. SEAM achieves the best in-domain performance on SLOP (0.96 AUC), and a SEAM trained on SLOP transfers to unseen commercial generators better than any baseline trained on the same data (0.89 AUC).

1 INTRODUCTION

Methods for detecting AI-generated or AI-edited images and videos—which we refer to broadly as deepfakes—generally fall into three categories (Lyu, 2020): (1) low-level approaches that exploit generation artifacts, often manifesting as spectral signatures in the frequency domain (Zhang et al., 2019; Durall et al., 2020; Frank et al., 2020; Qian et al., 2020; Li & Lyu, 2018; Agarwal et al., 2019); (2) high-level approaches that assess the semantic, physical, or physiological plausibility of the content (Li et al., 2018; Yang et al., 2019; Ciftci et al., 2020; Hu et al., 2021; Agarwal et al., 2020; Bohacek & Farid, 2022a); and (3) learned, data-driven approaches that train classifiers or leverage embeddings from large-scale models (Afchar et al., 2018; Rossler et al., 2019; Wang et al., 2020; Cozzolino et al., 2021; Ojha et al., 2023).

While these methods perform well on individual academic benchmarks (Li et al., 2020c; Dolhansky et al., 2020; Yan et al., 2023; Barrington et al., 2026b), they generalize poorly to new domains (e.g., deepfakes circulating on social media) (Chandra et al., 2026; Zi et al., 2020; Dell'Anna et al., 2025), unseen generators (Ojha et al., 2023; Corvi et al., 2023b), and content subjected to common pre-processing operations such as re-compression and resizing (Marra et al., 2018; Xuan et al., 2019).

In practice, establishing the presence of AI-generated content therefore combines passive detection with active approaches, including watermarking (Fernandez et al., 2023; Wen et al., 2023; Dathathri et al., 2024; Gowal et al., 2025) and provenance metadata (Coalition for Content Provenance and Authenticity, 2026; Content Authenticity Initiative, 2019; Rosenthol et al., 2020). The efficacy of these active approaches, however, remains limited: watermarks are vulnerable to evasion and deployed by only a subset of proprietary generators (Zhao et al., 2024; Saberi et al., 2024; An et al., 2024), while provenance metadata suffers from slow adoption (Das, 2025) and is routinely stripped by platforms upon upload (Lyu, 2020).

At a time when an ever-growing share of video consumption occurs on social media (Pew Research Center, 2025b;a) and the techniques and formats of deepfakes evolve rapidly (Cardenuto et al., 2023), the need for robust deepfake detection has never been greater.

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.

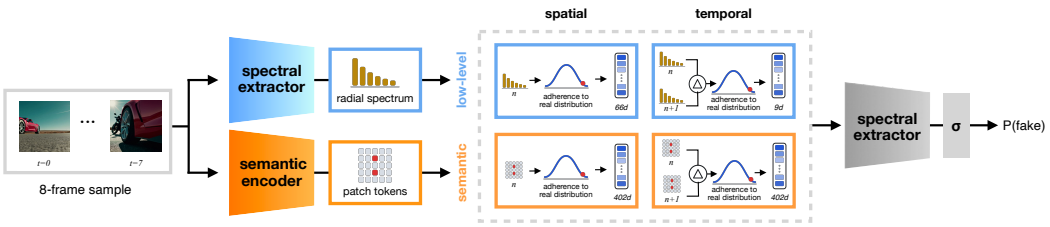


Figure 1: **Overview of SEAM.** Frozen spectral and semantic extractors feed four anomaly streams scored against real-only densities; a lightweight fusion head outputs the probability of the clip being AI-generated.

To address this, we propose, to our knowledge, the first detection architecture that explicitly unifies all three categories of cues and an in-the-wild dataset emphasizing the real-world attributes of AI-generated content, including the conditions under which they are shared (e.g., platform re-encoding, cropping, and re-upload). Our contributions can be summarized as follows:

1. **SEAM, a unifying architecture for AI-generated video detection.** For a given video, it extracts a spectral artifact profile, motion-coherence trajectories, and frozen-encoder embeddings, and combines these signals in a lightweight, per-stream anomaly ensemble. We demonstrate that SEAM leads on unseen open-source generators (0.92 vs. 0.89 AUC baseline), commercial generators including Sora, Kling, and Pika (0.87 AUC), and recent in-the-wild generations (0.80 AUC). Additionally, it adapts well to unseen domains without encoder adaptation.
2. **SLOP, a dataset of real and AI-generated video from Instagram, TikTok, and X.** With 6,000 videos in a 50:50 AI:real split, distributed uniformly across the source platforms, SLOP captures AI-generated content as it actually circulates on social media—preserving platform re-encoding and capturing a variety of AI involvement from edits to whole-cloth AI-persona accounts. Detectors that excel on academic benchmarks struggle with SLOP, while a calibrated SEAM model reaches 0.96 AUC in-domain and transfers back to unseen commercial generators at 0.89 AUC.
3. **An accompanying taxonomy of social-media video.** We annotate the dataset along style (e.g., hyperrealistic, animated, CCTV), format (e.g., talking head, montage, side-by-side), and content (e.g., vlog, advertisement, news), allowing for a category-specific analysis of where detectors fail.

2 Related Work

2.1 Low-Level Artifact Methods

These methods are based on the insight that generative AI models often leave artifacts in their outputs (Zhang et al., 2019; Corvi et al., 2023a), both visible (in the RGB space) (Li & Lyu, 2018; Li et al., 2020a;b) and imperceptible (in high frequencies of the signal) (Durall et al., 2020; Frank et al., 2020; Qian et al., 2020; Corvi et al., 2023b). Moreover, some models have been found to leave unique signatures, which can be used for attribution (Marra et al., 2019; Yu et al., 2019; 2021).

Within this category, several directions have emerged. Frequency-domain methods analyze the spectral statistics of an image, exploiting the observation that upsampling and deconvolution operations in generative pipelines produce characteristic high-frequency peaks that natural images lack; detectors operate directly on Fourier or DCT representations (Durall et al., 2020; Frank et al., 2020) or fuse frequency-aware branches into a learned pipeline (Qian et al., 2020). Blending-based methods instead target the compositing step of face-swap deepfakes: Face X-ray (Li et al., 2020b) localizes the boundary where a synthesized face is blended into a target frame, and self-blended images (Shiohara & Yamasaki, 2022) generate pseudo-fakes from real footage alone to learn generic blending artifacts. A third direction treats generator-specific noise residuals as fingerprints, enabling not only detection but attribution of an image to its source model (Girish et al., 2021; Asnani et al.,

Table 1: **Dataset Comparison.** In-the-wild: collected from circulating posts; social: platform-native processing; AI-gen: full-frame generation (vs. face manipulation); personas: dedicated AI-persona accounts. ∼: partially. Sizes are video counts (WildDeepfake: 707 videos of 7,314 sequences; DF40 adds 1M+ images).

Dataset	Year	Videos	In-the-wild	Social	AI-gen	Personas	Balanced	Splits
WildDeepfake (Zi et al., 2020)	2020	707	✓	✓				✓
DFDC (Dolhansky et al., 2020)	2020	128,154						✓
DF40 (Yan et al., 2024b)	2024	0.1M+			∼		∼	✓
GenVideo (Chen et al., 2026)	2024	2.3M			✓		∼	✓
Deepfake-Eval-2024 (Chandra et al., 2026)	2024	2,036	✓	✓	∼			
DeepSpeak (Barrington et al., 2026b)	2026	13,025			∼		∼	✓
DeepSpeak-Agentic (Barrington et al., 2026a)	2026	200			∼		✓	
MNW (Roca et al., 2026)	2026	320	✓	∼	✓			
Ours	2026	6,000	✓	✓	✓	✓	✓	✓

2023). While these methods can be highly accurate on content matching their assumed generation pipeline, their reliance on fragile low-level statistics makes them sensitive to compression, resizing, and other laundering operations (Marra et al., 2018), and diffusion models have been shown to exhibit substantially different artifact profiles than the GANs on which many of these detectors were developed (Ricker et al., 2022; Wang et al., 2023).

2.2 Semantic Plausibility Methods

These methods exploit the fact that generative models continue to generate outputs that, despite seeming photorealistic, violate certain properties of the physical world (Popov & Tarnopolsky, 2023; Farid, 2022b). Earlier, this manifested in body and structure deformations (e.g., five fingers) (Popov & Tarnopolsky, 2023; Yang et al., 2019), implausible movements (Bansal et al., 2025; Kang et al., 2024) (e.g., gymnast), and others.

As generative models improved, the efficacy of these methods has generally decreased (Nightingale & Farid, 2022), though some plausibility signals persist (e.g., light perspective (Farid, 2022b;a)). While these approaches may not always have enough signal to pick up on, the presence of such physical implausibilities can be a strong signal of AI presence.

Specific directions within this category include physiological methods, which detect the absence or distortion of biological signals such as eye blinking (Li et al., 2018) or the subtle skin-color variations induced by blood flow (Ciftci et al., 2020); geometric methods, which expose inconsistencies in head pose (Yang et al., 2019), facial landmark configurations (Yang et al., 2019), or corneal specular highlights (Hu et al., 2021); and audio-visual methods, which detect mismatches between speech and mouth dynamics (Agarwal et al., 2020; Haliassos et al., 2021). A related line of identity-based work models the facial, gestural, and vocal mannerisms of a specific individual and flags content that deviates from this biometric-behavioral signature (Agarwal et al., 2019; Bohacek & Farid, 2022a;b). Because these methods reason about content rather than signal statistics, they tend to be more robust to compression and laundering than their low-level counterparts.

2.3 Learned Representation Methods

This third category of methods started out by training end-to-end models to predict whether the input is AI-generated or AI-edited. Early examples include MesoNet (Afchar et al., 2018), a shallow CNN targeting mesoscopic image properties, and the XceptionNet baseline popularized by FaceForensics++ (Rossler et al., 2019), which remained the standard supervised recipe for several years and underpinned the ensemble winners of the Deepfake Detection Challenge (Dolhansky et al., 2020). Notably, a classifier trained on a single generator with careful augmentation was shown

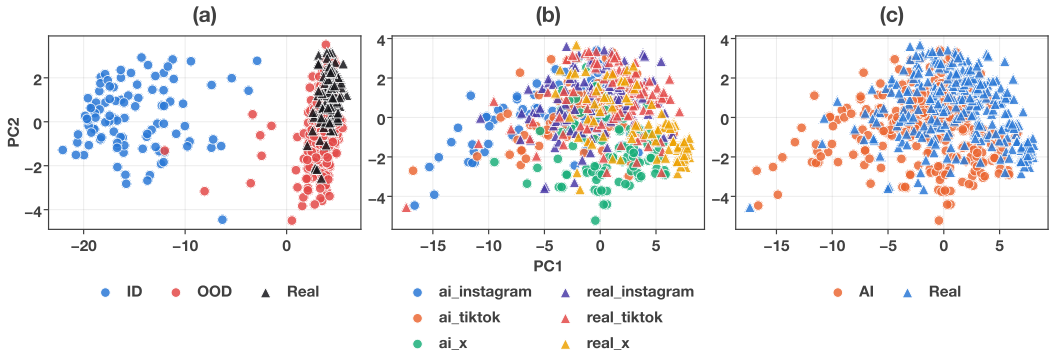


Figure 2: **Data PCA projections.** Leading two PCs of semantic embeddings: (a) AIGVDBench by ID/OOD/real grouping; (b) SLOP real vs. AI per platform; (c) SLOP pooled.

to generalize surprisingly well across contemporaneous GANs (Wang et al., 2020), though this generalization later proved brittle under distribution shift to new generator families (Corvi et al., 2023b).

Later, it has been discovered that training a simple linear probe on top of pre-trained deep learning embeddings such as CLIP or DINO often achieves similar performance at much smaller training cost. Ojha et al. (2023) showed that nearest-neighbor and linear classification in a frozen CLIP feature space outperforms dedicated deepfake classifiers on unseen generative models, and subsequent work demonstrated that such CLIP-based detectors require only a handful of example images and degrade gracefully under compression and resizing (Cozzolino et al., 2024). In the video domain, multi-modal semantic embeddings paired with lightweight classifiers distinguish real from AI-generated human motion while remaining robust to the resolution and compression attacks that confound low- and mid-level methods (Bohacek & Farid, 2024). Related work learns identity-specific embeddings from reference footage of a person and detects deepfakes as anomalies in this learned space (Cozzolino et al., 2021).

2.4 Datasets

Early deepfake detection benchmarks centered on face-swap and reenactment manipulations of curated footage, including FaceForensics++ (Rossler et al., 2019), Celeb-DF (Li et al., 2020c), DeeperForensics-1.0 (Jiang et al., 2020), and the large-scale DFDC dataset (Dolhansky et al., 2020). Subsequent efforts broadened coverage along several axes: WildDeepfake (Zi et al., 2020) collected deepfakes circulating on the internet rather than lab-generated ones; ForgeryNet (He et al., 2021) scaled to tens of manipulation types; and FakeAVCeleb (Khalid et al., 2021) introduced paired audio-visual manipulations. More recent benchmarks target modern synthesis pipelines, including GenImage (Zhu et al., 2023) for text-to-image models, DF40 (Yan et al., 2024b) spanning forty generation methods, DeepAction (Bohacek & Farid, 2024) for text-to-video human motion, DeepSpeak (Barrington et al., 2026b) and DeepSpeak-Agentic (Barrington et al., 2026b) for lip-sync, face-swap, and conversational avatar manipulations, and Deepfake-Eval-2024 (Chandra et al., 2026), an in-the-wild benchmark of deepfakes collected from social media on which state-of-the-art detectors degrade dramatically.

3 Data

3.1 Existing Datasets

Table 1 compares SLOP against prominent deepfake video datasets. The early in-the-wild benchmarks (WildDeepfake, Deepfake-Eval-2024) target face manipulation and predate the current generation of video models, while the large corpora (DF40, GenVideo) are laboratory-generated and never pass through a platform. The most recent efforts either report no real counterparts (MNW) or are recorded under controlled conditions (DeepSpeak, DeepSpeak-Agentic). No existing dataset combines in-the-

Table 2: **Benchmark Comparison.** AUC across AIGVDBench (ID; OOD-C: 19 open-source generators; OOD-F: 11 commercial generators), MNW (200 in-the-wild 2025 commercial clips), and MintVid (talking-head generation). All methods retrained by us with size-matched (∼22M) encoders, each shown at its best encoder.

	AIGVDBench			MNW	MintVid
Method	**ID**	**OOD-C**	**OOD-F**		
Linear probe (UFD) (Ojha et al., 2023)	1.000	0.858	0.819	0.728	0.753
Full fine-tune (Wang et al., 2020)	1.000	0.732	0.821	0.660	0.624
RINE (Koutlis & Papadopoulos, 2024)	1.000	0.816	0.764	0.645	0.679
FatFormer (Liu et al., 2024)	1.000	0.820	0.850	0.496	0.555
C2P-CLIP (Tan et al., 2025)	1.000	0.639	0.622	0.570	0.532
DeMamba (Chen et al., 2026)	1.000	0.858	0.802	0.728	0.727
Human Action CLIPs (Bohacek & Farid, 2024)	1.000	0.885	0.844	0.642	0.769
Effort (Yan et al., 2024a)	1.000	0.824	0.868	0.786	0.715
ForgeLens (Chen et al., 2025)	1.000	0.826	0.870	0.755	0.640
Ours	**1.000**	**0.923**	**0.871**	**0.804**	**0.776**

wild collection, platform-native processing, full-frame AI generation, and AI-persona accounts with a balanced, split-ready release: this is the gap SLOP fills.

We evaluated on AIGVDBench (Ma et al., 2026), a benchmark that spans 31 open- and closed-source video generators, using its cross-generator generalization method: train on a single generator (Open-Sora), test on 19 open-source and 11 unseen closed-source generators.

3.2 Our Dataset: Social-platform Longitudinal Observed Posts (SLOP)

We collect a custom in-the-wild dataset of image and video social media posts from Instagram, TikTok, and X, with annotations indicating whether they were AI-generated. The dataset is predominantly video and has a total of 6,000 datapoints, with a 50:50 real:AI-generated ratio—maintained both globally as well as within the individual social media subsets—and three pre-defined splits: train (0.7), val (0.1), and test (0.2). Splits are account-disjoint: no account contributes to more than one split, so detectors cannot exploit creator-specific cues across the boundary.

Unlike previous datasets, ours targets AI-*generated* content as it actually circulates on social media (Table 1). The genuinely in-the-wild benchmarks are either dated—Deepfake-Eval-2024 predates the current generation of video models—or small and one-sided (MNW ships 320 clips and no real counterparts), while the large corpora are laboratory-generated and never pass through a platform. No existing dataset focuses specifically on social media and the processing that media undergoes there (re-encoding, cropping, captioning, re-upload), nor on fully AI personas and influencers, whose accounts produce a stream of generated content embedded in ordinary creator behavior.

The dataset is constructed through a combination of methods to capture the diversity of AI-generated content on social media: platform-representative random panels (37,100 TikTok and 55,046 X posts over 53 months) mined with detector-assisted candidate ranking and human verification; dedicated AI-persona and AI-influencer accounts; and matched, verifiably human creator accounts for the real class, screened per-video. Notably, the dataset is not intended to be representative of the social media distribution, but rather to capture a breadth of AI-generated content and similar, real counterparts. Figure 2 shows PCA projections of semantic embeddings for both corpora: AIGVDBench separates cleanly along its in-distribution/OOD/real grouping (a), while on SLOP the real and AI classes overlap heavily both per platform (b) and overall (c). This shows that the social-media distribution is semantically far less separable than the benchmark, which motivates the calibration experiments of Section 5.

Table 3: **Stream Ablation.** ✓ marks included streams; silenced streams are zeroed before fusion and the head is retrained (frozen OpenVision configuration, seed 0).

Semantic	Low-level	Physical	OOD-C	OOD-F
✓			0.855	0.793
	✓		0.616	0.578
		✓	0.817	0.648
✓	✓		0.851	0.797
✓		✓	0.925	0.812
	✓	✓	0.827	0.662
✓	✓	✓	0.923	0.813

Table 4: **Encoder Comparison.** SEAM (frozen configuration) across matched-size (∼22M, ViT-S scale) encoders.

	AIGVDBench			MNW	MintVid
Encoder	**ID**	**OOD-C**	**OOD-F**		
CLIP-style (OpenVision ViT-S/16) (Li et al., 2025)	1.000	0.923	0.813	0.721	0.776
DINOv2 (ViT-S/14) (Oquab et al., 2023)	0.999	0.829	0.781	0.717	0.671
DINOv3 (ViT-S/16) (Siméoni et al., 2025)	1.000	0.866	0.785	0.700	0.708

4 METHODS

Our method uses two frozen feature extractors, one for semantic embeddings, and another for low-level embeddings; Figure 1 depicts the full pipeline, from the two extractors through the four anomaly streams to the fusion head. A video is represented by $T = 8$ frames $x_1, \ldots, x_T$ sampled uniformly across the clip. After testing between three different semantic encoders: DINOv2 (Oquab et al., 2023), DINOv3 (Siméoni et al., 2025), and OpenVision (Li et al., 2025), our best performing architecture uses the OpenVision semantic embeddings.

The spectral extractor converts each frame to grayscale, computes a 2-dimensional Fast Fourier Transform of the frame's high-pass residual, and finally radially bins the resulting power spectrum into n_{bins} log-magnitude values, yielding $s_t = \text{Spec}(x_t) \in \mathbb{R}^{d_{\text{spec}}}$ with $n_{\text{bins}} = 64$ and frequencies above $0.6\times$ Nyquist discarded. The semantic extractor passes each frame through the frozen encoder and keeps its N patch-token embeddings $z_{t,1}, \ldots, z_{t,N} \in \mathbb{R}^D$ together with their mean-pooled summary $\bar{z}_t = \frac{1}{N}\sum_i z_{t,i}$ (discarding the CLS token). Temporal cues are carried by the consecutive differences $v_t = s_{t+1} - s_t$ and $\delta_{t,i} = z_{t+1,i} - z_{t,i}$, together with motion-coherence vectors m_t computed from the token field.

Next, inspired by STALL's real-only scoring method (Ben Hayun et al., 2026), our initial phase of training fits a Gaussian reference density to real training features in each cue family: for family $c \in \{s_t,\ z_{t,i},\ v_t,\ \delta_{t,i},\ m_t,\ \Delta x_t\}$ we estimate a mean μ_c and a Ledoit-Wolf shrinkage covariance $\hat{\Sigma}_c$ (Ledoit & Wolf, 2004) from real videos only—no generated content informs any density—and score a feature u by its negative log-likelihood, $a_c(u) = -\log \mathcal{N}(u;\, \mu_c, \hat{\Sigma}_c)$. We uniquely use temporal streams to score raw deltas between one frame and the next, and semantic streams to score individual patch tokens for each frame rather than one total embedding.

Each stream summarizes its scores into a fixed-length descriptor. Writing $\text{top}k$ for the k largest values in descending order, $\text{st}(\cdot)$ for (mean, max), and $\text{st}_3(\cdot)$ for (mean, max, 90th percentile), the

Table 5: **Compression Resilience.** Change in OOD-F AUC after H.264 re-encoding at increasing CRF; closer to zero is better. Two strongest far-OOD baselines shown.

	Clean	Δ 23	Δ 28	Δ 35	Δ 42	Δ 47
Effort	0.868	-0.175	-0.214	-0.215	-0.230	-0.250
ForgeLens	0.869	-0.072	-0.058	-0.054	-0.068	-0.079
Ours	0.871	-0.120	-0.200	-0.244	-0.264	-0.271

Table 6: **Resolution Resilience.** Change in OOD-F AUC after bicubic downscaling of the shorter side; closer to zero is better. Two strongest far-OOD baselines shown.

	Native	Δ 480px	Δ 360px	Δ 240px	Δ 144px
Effort	0.868	-0.117	-0.110	-0.123	-0.142
ForgeLens	0.869	-0.044	-0.051	-0.069	-0.090
Ours	0.871	-0.010	-0.005	-0.005	-0.057

four descriptors (with stream names abbreviated) are

$$d_{\text{ll_sp}} = \left[\tfrac{1}{T}\sum_t s_t;\ \text{st}\{a(s_t)\}\right] \tag{1}$$

$$d_{\text{sem_sp}} = \left[\text{top}k\{a(z_{t,i})\};\ \text{st}\{a(z_{t,i})\};\ \bar{z}\right] \tag{2}$$

$$d_{\text{ll_tmp}} = \left[\text{st}_3\{a(v_t)\};\ \text{st}_3\{a(m_t)\};\ \text{st}_3\{a(\Delta x_t)\}\right] \tag{3}$$

$$d_{\text{sem_tmp}} = \left[\text{top}k\{a(\delta_{t,i})\};\ \text{st}\{a(\delta_{t,i})\};\ \bar{\delta}\right] \tag{4}$$

where $\bar{z}$ is the clip-averaged pooled embedding, and $\bar{\delta}$ the mean token delta. The resulting dimensions are $|d_{\text{ll_spatial}}| = d_{\text{spec}} + 2$, $|d_{\text{sem_spatial}}| = |d_{\text{sem_temporal}}| = k + 2 + D$, and $|d_{\text{ll_temporal}}| = 9$, with $k = 16$ and $D = 384$ for the ViT-S models used throughout.

Each descriptor is z-normalized with statistics recorded from the training reals and passed through its own small MLP, $h_k = \text{MLP}_k(\tilde{d}_k)$, which also emits an auxiliary per-stream logit $\hat{y}_k = w_k^\top h_k$. A linear fusion layer maps the concatenated hidden representations to the final logit, $\hat{y} = w^\top[h_1; h_2; h_3; h_4] + b$, and the head is trained with $\mathcal{L} = \text{BCE}(\sigma(\hat{y}), y) + \lambda \sum_k \text{BCE}(\sigma(\hat{y}_k), y)$, where the auxiliary terms force every stream to remain independently predictive. Our final result is the probability of the clip being fake, combining in the process low-level artifacts, high-level semantic embeddings, and physical plausibility. The inference is forward pass only, and both the densities and the head are frozen from the saved checkpoint; nothing is retrained at the detection time.

A fine-tuned-backbone variant, used where Table 2 reports our best far-OOD configuration, adapts a copy of the encoder with orthogonal-residual fine-tuning in the style of Effort (Yan et al., 2024a) and feeds it to the semantic-spatial stream only—the temporal streams keep the original frozen encoder, whose natural motion statistics discriminative adaptation destroys. The adaptation's co-trained per-frame classification logit is appended to $d_{\text{sem_spatial}}$, and the fusion layer is anchored to the semantic-spatial logit with a zero-initialized correction, $\hat{y} = \hat{y}_{\text{sem_spatial}} + \Delta(h_1, \ldots, h_4)$ with Δ initialized to zero, so the model starts as its best single stream and must earn any departure from it. Because the base encoder is never modified, adapting either variant to a new domain reduces to re-fitting the densities and the small head, which is the mechanism behind the calibrated results of Section 5.

5 Results

We first examine the headline comparison. Table 2 reports AUC for all ten methods across the three AIGVDBench conditions and the two independently-collected sets. SEAM attains the best or tied-best value in every column: 1.000 in-distribution, 0.923 on OOD-C against 0.885 for the next-best method (Human Action CLIPs), 0.871 on OOD-F against 0.870 for ForgeLens and 0.868 for Effort, 0.804 on MNW against 0.786 for Effort, and 0.776 on MintVid against 0.769 for Human Action CLIPs. This shows that the margin is largest on open-source generators, while on commercial generators the leading methods are close.

We next turn to the contribution of the individual streams. Table 3 reports the ablation over the three cue families. The semantic stream is the strongest individual stream (0.855 OOD-C / 0.793 OOD-F), and the low-level spatial stream alone is the weakest (0.616 / 0.578). The semantic+physical pair (0.925 / 0.812) matches the full three-stream model (0.923 / 0.813) within run-to-run noise. This means that the semantic and physical cues carry most of the generalization, while the low-level stream contributes little on clean benchmark data.

Table 7: **Cross-Training Generalization.** Each method is trained on one dataset (rows) and evaluated on five test sets (MNW and VRT are evaluation-only). Two strongest far-OOD baselines shown. Grid means: ours 0.769, Effort 0.742, ForgeLens 0.717.

Method	Trained on	Evaluated on				
		AIGVD	SM	MintVid	MNW	VRT
Ours	AIGVD	0.813	0.569	0.779	0.721	0.574
	SM	0.894	**0.957**	0.714	**0.878**	0.568
	MintVid	**0.904**	0.807	**0.952**	0.818	0.590
Effort	AIGVD	0.868	0.609	0.702	0.786	0.546
	SM	0.853	0.954	0.642	0.743	**0.628**
	MintVid	0.809	0.781	0.948	0.740	0.517
ForgeLens	AIGVD	0.869	0.851	0.626	0.755	0.550
	SM	0.711	0.929	0.624	0.649	0.492
	MintVid	0.869	0.817	0.768	0.703	0.543

Table 8: **In-Domain Performance on SLOP.** Trained on the train split, evaluated on the creator-disjoint test split. AP: average precision. Two strongest far-OOD baselines shown.

Method	AP	AUROC
Ours	**0.961**	**0.957**
Effort	0.956	0.954
ForgeLens	0.929	0.929

Table 4 reports SEAM in its frozen configuration across three pretrained encoders of matched size. The contrastive OpenVision encoder leads in every column (0.923 OOD-C, 0.813 OOD-F, 0.721 MNW, 0.776 MintVid), ahead of DINOv3 (0.866 / 0.785 / 0.700 / 0.708) and DINOv2 (0.829 / 0.781 / 0.717 / 0.671). This shows that language-supervised features are the better substrate for density-based scoring at this scale, with the architecture held fixed.

We next examine resilience to common post-processing. Table 5 reports the change in OOD-F AUC after H.264 re-encoding. SEAM degrades the most of the three methods, from -0.120 at CRF 23 to -0.271 at CRF 47, compared to -0.072 to -0.079 for ForgeLens and -0.175 to -0.250 for Effort. Table 6 reports the complementary experiment under bicubic downscaling, where SEAM is the most resilient method by a wide margin, losing -0.005 to -0.010 down to 240px and -0.057 at 144px, against -0.117 to -0.142 for Effort and -0.044 to -0.090 for ForgeLens. This means that SEAM's weakness is specific to compression artifacts, which corrupt the frame-to-frame deltas its temporal streams read, while genuine resolution loss has little effect.

We next evaluate on SLOP. Table 8 reports methods trained on the SLOP train split and evaluated on its creator-disjoint test split. SEAM reaches 0.961 AP and 0.957 AUROC, ahead of Effort (0.956 / 0.954) and ForgeLens (0.929 / 0.929). Table 7 reports the cross-training matrix over five evaluation sets. SEAM attains the best grid mean (0.769, against 0.742 for Effort and 0.717 for ForgeLens), the best in-domain diagonals on SLOP (0.957) and MintVid (0.952), and the best single transfer cell (0.904, MintVid-trained evaluated on AIGVD). Its weakest cell is AIGVD-trained evaluated on SLOP (0.569); re-fitting the densities and head on SLOP's train split recovers this to 0.957 without touching the encoder. All three methods score low on VRT (0.49–0.63), whose real class is ASMR-only.

Finally, we examine the errors. On SLOP, 11.6% of test clips are misclassified, skewed toward missed fakes rather than false alarms (72% of the errors are AI clips predicted real) and roughly even across platforms (39% on X, 30% on Instagram, 31% on TikTok). On AIGVDBench, errors concentrate in the OOD conditions (872 in other-open-source and 640 in cross-generator, against a single in-distribution error), with LTX (135) and EasyAnimate (85) the largest per-generator sources of false negatives.

6 DISCUSSION

Across the evaluations in this work, the unified representation proves both robust and flexible. Trained on a single generator, it leads or matches the strongest baselines across open-source, commercial, and independently collected test sets, holds up under resolution loss where discriminative baselines degrade, and, because only the real-video reference densities and a small fusion head are ever trained, adapts to a new domain such as social media by re-fitting those components alone: the same architecture moves from 0.60 to 0.96 AUC on SLOP without touching the encoder. We find this combination of generality and adaptability exciting.

At the same time, much work remains. Our method is the most sensitive of the compared detectors to heavy re-compression, and no method we evaluated, ours included, performs well on adversarially curated or content-matched benchmarks. The three cue families we unify are also not the only ways this problem has been represented: watermarking, provenance, biometric, and audio-visual approaches all carry complementary signal. One could imagine unifying all of them under the same anomaly-scoring framing, or, with more nuance, learning which cues to activate for each individual input: a talking-head clip and a screen-recorded montage arguably deserve different evidence. We leave this input-adaptive composition of detection cues as future work, and release SLOP and our code to support it.

REPRODUCIBILITY STATEMENT

The code for the project is fully open-sourced at `https://anonymous.4open.science/r/iclr-dfd-754D`; the dataset and model weights are available through a gated request form at the same location (see the Ethics Statement).

ETHICS STATEMENT

The authors have carefully considered the ethical impacts of this work. We believe the net impact of deepfake detection is positive for society: it upholds media integrity and the ability of individuals to trust the content they see online. That said, there are risks that come with releasing deepfake detectors and code, as they can be used in an adversarial fashion to aid attackers in building models that not only produce realistic content but also evade detectors. For that reason, we decided not to release the model checkpoints and the full dataset snapshot openly, but rather under a request form through which we can review whether requesters have academic or otherwise legitimate evaluation purposes, mitigating negative uses.

AI USE STATEMENT

For some experiments in this work, we used generative AI tools to aid with code implementation, data cleaning, and feedback on experimental design. We have not used generative AI tools for the following tasks, which are not applicable to this work: data generation, development of theoretical models or conceptual frameworks, formulation of mathematical claims, provision of critical ingredients for proving mathematical claims, assistance in the writing of proofs, refinement of hypotheses, translation assistance, and support with qualitative or thematic data analysis. Additionally, we used generative AI tools for help with formatting of LaTeX tables and figures and revisions/suggestions for paper writing. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. MesoNet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pp. 1–7. IEEE, 2018.

Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *CVPR workshops*, volume 1, 2019.

Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 660–661, 2020.

Bang An, Mucong Ding, Tahseen Rabbani, Aakriti Agrawal, Yuancheng Xu, Chenghao Deng, Sicheng Zhu, Abdirisak Mohamed, Yuxin Wen, Tom Goldstein, et al. WAVES: Benchmarking the robustness of image watermarks. *arXiv preprint arXiv:2401.08573*, 2024.

Vishal Asnani, Xi Yin, Tal Hassner, and Xiaoming Liu. Reverse engineering of generative models: Inferring model hyperparameters from generated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15477–15493, 2023.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating physical commonsense for video generation. In *International Conference on Learning Representations*, volume 2025, pp. 102075–102121, 2025.

Sarah Barrington, Maty Bohacek, and Hany Farid. The DeepSpeak-Agentic dataset. *arXiv preprint arXiv:2606.03686*, 2026a.

Sarah Barrington, Maty Bohacek, and Hany Farid. The DeepSpeak dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1893–1902, 2026b.

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*, pp. 16299–16310, 2026.

Matyas Bohacek and Hany Farid. Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms. *Proceedings of the National Academy of Sciences*, 119(48):e2216035119, 2022a.

Matyáš Bohacek and Hany Farid. Protecting president Zelenskyy against deep fakes. *arXiv preprint arXiv:2206.12043*, 2022b.

Matyas Bohacek and Hany Farid. Human action CLIPs: Detecting ai-generated human motion. *arXiv preprint arXiv:2412.00526*, 2024.

João Phillipe Cardenuto, Jing Yang, Rafael Padilha, Renjie Wan, Daniel Moreira, Haoliang Li, Shiqi Wang, Fernanda Andaló, Sébastien Marcel, and Anderson Rocha. The age of synthetic realities: Challenges and opportunities. *arXiv preprint arXiv:2306.11503*, 2023.

Nuria Alina Chandra, Hannah Lee, Ryan Murtfeldt, Lin Qiu, Arnab Karmakar, Emmanuel Tanumihardja, Kevin Farhat, Ben Caffee, Changyeon Lee, Jongwook Choi, et al. Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10668–10678, 2026.

Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, et al. Demamba: Ai-generated video detection on million-scale genvideo benchmark. *Science China Information Sciences*, 69(6):162103, 2026.

Yingjian Chen, Lei Zhang, and Yakun Niu. Forgelens: Data-efficient forgery focus for generalizable forgery image detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16270–16280. IEEE, 2025.

Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. FakeCatcher: Detection of synthetic portrait videos using biological signals. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

Coalition for Content Provenance and Authenticity. Content credentials: C2PA technical specification, version 2.3. `https://c2pa.org/`, 2026.

Content Authenticity Initiative. Content Authenticity Initiative. `https://contentauthenticity.org/`, 2019.

Juliet M Corbin and Anselm Strauss. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, 13(1):3–21, 1990.

Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 973–982, 2023a.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023b.

Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. ID-Reveal: Identity-aware deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15108–15117, 2021.

Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of AI-generated image detection with CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4356–4366, 2024.

Shruti Das. The promise and risk of digital content provenance. Center for Democracy & Technology, December 2025. URL `https://cdt.org/insights/the-promise-and-risk-of-digital-content-provenance/`.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, et al. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823, 2024.

Stefano Dell'Anna, Andrea Montibeller, and Giulia Boato. TrueFake: A real world case dataset of last generation fake images also shared on social networks. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2025.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (DFDC) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7890–7899, 2020.

Hany Farid. Lighting (in) consistency of paint by text. *arXiv preprint arXiv:2207.13744*, 2022a.

Hany Farid. Perspective (in) consistency of paint by text. *arXiv preprint arXiv:2206.14617*, 2022b.

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22466–22477, 2023.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pp. 3247–3258. PMLR, 2020.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021.

Sharath Girish, Saksham Suri, Sai Saketh Rambhatla, and Abhinav Shrivastava. Towards discovery and attribution of open-world GAN generated images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 14094–14103, 2021.

Sven Gowal, Rudy Bunel, Florian Stimberg, David Stutz, Guillermo Ortiz-Jimenez, Christina Kouridi, Mel Vecerik, Jamie Hayes, Sylvestre-Alvise Rebuffi, Paul Bernard, et al. SynthID-Image: Image watermarking at internet scale. *arXiv preprint arXiv:2510.09263*, 2025.

Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5039–5049, 2021.

Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. ForgeryNet: A versatile benchmark for comprehensive forgery analysis. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4358–4367, 2021.

Shu Hu, Yuezun Li, and Siwei Lyu. Exposing GAN-generated faces using inconsistent corneal specular highlights. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2500–2504. IEEE, 2021.

Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2889–2898, 2020.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024.

Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*, 2021.

Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on computer vision*, pp. 394–411. Springer, 2024.

Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, 4 edition, 2018.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

Haodong Li, Bin Li, Shunquan Tan, and Jiwu Huang. Identification of deep network generated images using disparities in color components. *Signal Processing*, 174:107616, 2020a.

Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face X-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5001–5010, 2020b.

Xianhang Li, Yanqing Liu, Haoqin Tu, and Cihang Xie. Openvision: A fully-open, cost-effective family of advanced vision encoders for multimodal learning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3977–3987. IEEE, 2025.

Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018.

Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing AI created fake videos by detecting eye blinking. In *2018 IEEE International workshop on information forensics and security (WIFS)*, pp. 1–7. Ieee, 2018.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3207–3216, 2020c.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10770–10780. IEEE, 2024.

Siwei Lyu. Deepfake detection: Current challenges and next steps. *arXiv preprint arXiv:2003.09234*, 2020.

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. *arXiv preprint arXiv:2601.11035v1*, 2026.

Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva. Detection of GAN-generated fake images over social networks. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, pp. 384–389. IEEE, 2018.

Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do GANs leave artificial fingerprints? In *2019 IEEE conference on multimedia information processing and retrieval (MIPR)*, pp. 506–511. IEEE, 2019.

Sophie J Nightingale and Hany Farid. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8):e2120481119, 2022.

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*, pp. 24480–24489, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Lawrence A Palinkas, Sarah M Horwitz, Carla A Green, Jennifer P Wisdom, Naihua Duan, and Kimberly Hoagwood. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and Policy in Mental Health and Mental Health Services Research*, 42(5):533–544, 2015.

Pew Research Center. Social media fact sheet. `https://www.pewresearch.org/internet/fact-sheet/social-media/`, November 2025a.

Pew Research Center. Social media and news fact sheet. `https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/`, September 2025b.

Fedor K Popov and Grigory Tarnopolsky. Magic angles in equal-twist trilayer graphene. *Physical Review B*, 108(8):L081124, 2023.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pp. 86–103. Springer, 2020.

Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. Towards the detection of diffusion model deepfakes. *arXiv preprint arXiv:2210.14571*, 2022.

Thomas Roca, Marco Postiglione, Chongyang Gao, Isabel Gortner, Zuzanna Wojciak, Pengce Wang, Mahsa Alimardani, Shirin Anlen, Kevin White, Juan Lavista Ferres, et al. The microsoft-northwestern-witness benchmark for deepfake detection. *IEEE Intelligent Systems*, 41(2):15–23, 2026.

Leonard Rosenthol, Andy Parsons, Eric Scouten, Jatin Aythora, Bruce MacCormack, Paul England, Marc Lavallée, Jonathan Dotan, Sherif Hanna, Hany Farid, and Sam Gregory. The Content Authenticity Initiative: Setting the standard for digital content attribution. White paper, Adobe, 2020. URL `https://contentauthenticity.org/`.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1–11, 2019.

Mehrdad Saberi, Vinu Sankar Sadasivan, Keivan Rezaei, Aounon Kumar, Atoosa Chegini, Wenxiao Wang, and Soheil Feizi. Robustness of AI-image detectors: Fundamental limits and practical attacks. In *International Conference on Learning Representations*, volume 2024, pp. 27500–27526, 2024.

Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18720–18729, 2022.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7184–7192, 2025.

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. 8695–8704, 2020.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22445–22455, 2023.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv preprint arXiv:2305.20030*, 2023.

Xinsheng Xuan, Bo Peng, Wei Wang, and Jing Dong. On the generalization of GAN image forensics. In *Chinese conference on biometric recognition*, pp. 134–141. Springer, 2019.

Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. DeepfakeBench: A comprehensive benchmark of deepfake detection. *arXiv preprint arXiv:2307.01426*, 2023.

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*, 2024a.

Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Li Yuan, Chengjie Wang, Shouhong Ding, and Yunsheng Wu. DF40: Toward next-generation deepfake detection. *arXiv*, abs/2406.13495, 2024b.

Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 8261–8265. IEEE, 2019.

Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to GANs: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7556–7566, 2019.

Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *Proceedings of the IEEE/CVF International conference on computer vision*, pp. 14448–14457, 2021.

Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in GAN fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pp. 1–6. IEEE, 2019.

Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang, and Lei Li. Invisible image watermarks are provably removable using generative AI. *Advances in neural information processing systems*, 37:8643–8672, 2024.

Mingjian Zhu, Hanting Chen, Qiang Yan, Xu Huang, Guanyu Lin, Wei Li, Zhaopeng Tu, Hailin Hu, Jie Hu, and Yunhe Wang. GenImage: A million-scale benchmark for detecting ai-generated image. *arXiv*, abs/2306.08571, 2023.

Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. WildDeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 2382–2390, 2020.

A Our Dataset: Social-platform Longitudinal Observed Posts (SLOP)

A.1 Data Collection

SLOP was collected with a two-track protocol designed to make label provenance auditable without per-video annotation, following the documentation practice of datasheets for datasets (Gebru et al., 2021). The first track targets *designated accounts*: for the AI class, accounts that post exclusively AI-generated content (dedicated AI personas, AI-slop channels, and generator showcase pages), and for the real class, accounts of extensively documented humans with posting histories predating modern video generation. Candidate accounts were gathered by the research team under a criterion-based purposive sampling scheme (Palinkas et al., 2015), screened by inspecting the account biography and recent posting grid, and rejected wholesale on any evidence of mixed posting; accounts failing the screen after collection were quarantined rather than relabeled. The second track constructs *keyword queries from popular generator names* and harvests the authors of matching video posts: an account enters the AI class only if it has at least two keyword-matching native-video posts, its profile self-identifies AI content, and it passes a negative filter excluding news, corporate, and commentary pages.

All content was retrieved through platform interfaces: the X API v2 full-archive search, the TikTok Research API, and API-based scraping for Instagram. Videos are stored under their platform-native identifiers, and per-post metadata (author, timestamps, engagement counters where provided) is retained alongside each file, so every clip in the dataset is traceable to a public post and its originating account.

Generator keywords. The keyword track queried the following terms, used both as hashtags and as plain-text search stems: `veo3`, `veo3video`, `veo`, `sora`, `sora2`, `kling`, `klingai`, `runwayml`, `pikalabs`, `hailuo`, `higgsfield`, `midjourney`, `minimax`, `lumalabs`, `cogvideox`, `wan2`, `aivideo`, `aifilm`, `aianimation`, `aiart`, `madewithai`, `aigenerated`.

Accounts. The real class draws on the following account seeds: Instagram: `emrata`, `sarasampaio`, `taylor_hill`, `winnieharlow`, `gigihadid`, `bellahadid`, `karliekloss`, `ashleygraham`, `romeestrijd`, `jasminetookes`, `adrianalima`, `candiceswanepoel`, `irinashayk`, `joansmalls`, `kendalljenner`, `gordongram`, `zachking`, `kingbach`, `lelepons`, `hannahstocking`, `juanpazurita`, `thebodycoach`, `noahbeck`, `kevinhart`, `therock`, `vindiesel`, `willsmith`, `jasonderulo`, `davidbeckham`, `pokimanelol`, `valkyrae`, `ninja`, `loserfruit`, `shroud`, `timthetatman`, `jiffpom`, `itsdougthepug`, `nala_cat`, `tuckerbudzyn`, `mayapolarbear`, `thiswildidea`, `chefjoseandres`, `kayla_itsines`, `massy_arias`, `muradosmann`; TikTok: `drewafualo`, `tinx`, `kallmekris`, `gordonramsayofficial`, `khaby.lame`, `charlidamelio`, `dixiedamelio`, `addisonre`, `zachking`, `mrbeast`, `therock`, `jasonderulo`, `itsjojosiwa`, `lorengray`, `babyariel`, `niasioux`; X: `mkbhd`, `MrBeast`, `GordonRamsay`, `DudePerfect`, `ZachKing`, `Cristiano`, `TheRock`, `SnoopDogg`, `garyvee`, `stephenasmith`, `rickygervais`, `jimmyfallon`, `SerenaWilliams`, `usainbolt`, `kevinhart4real`, `DanielRicciardo`. The AI class draws on curated accounts: TikTok: `grannyspills`, `angelica.ai`, `sebinsol`, `aicenturyclips`, `aisha.robot`, `bella.hueman`, `lilmiquela`; X: `teamzatsu`, `VideosAIpro`, `NeoSynthVisions`, `EBONYMANLEY_22`, `kellyeld`, `Megalania34`, `GarotasAI`, `Danielvisualsai`, `Christian_Luoma`, `TonyAngelAI`, `jojojojo822777`, `333mic999`, `XCallofShadows`, `PrometheanAIX`, `prompt_soru`, `picatrix_picori`, `pixelbouncer`, `zpipechan`, `AIPointView`, `RockGrokAI`, `RUN_Creators`, `YuuAki101`: plus 427 additional TikTok accounts surfaced by the keyword track; the complete per-clip account roster is included in the dataset release manifest.

A.2 Statistics

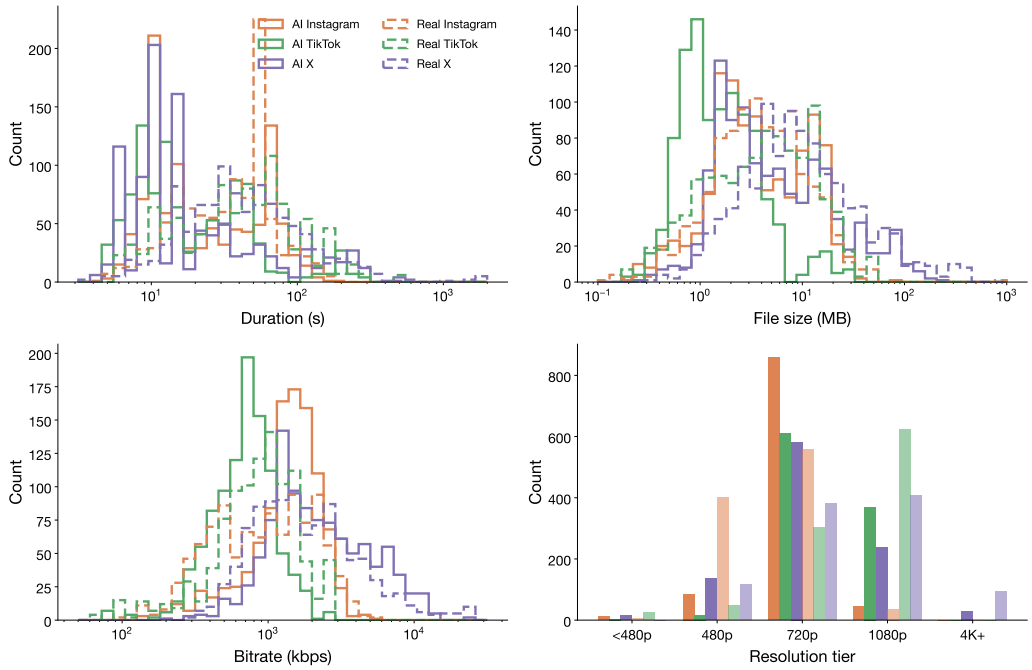


Figure 3: **Per-cell distributions.** Duration, file size, bitrate, and resolution per platform-class cell (solid: AI; dashed: real).

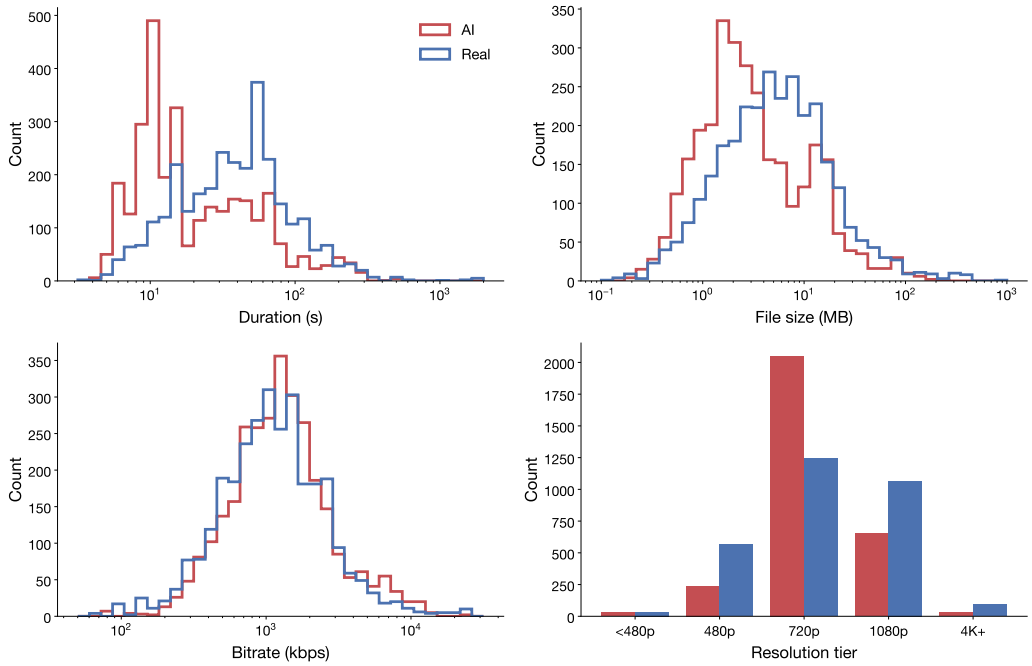


Figure 4: **AI vs. real, pooled across platforms.**

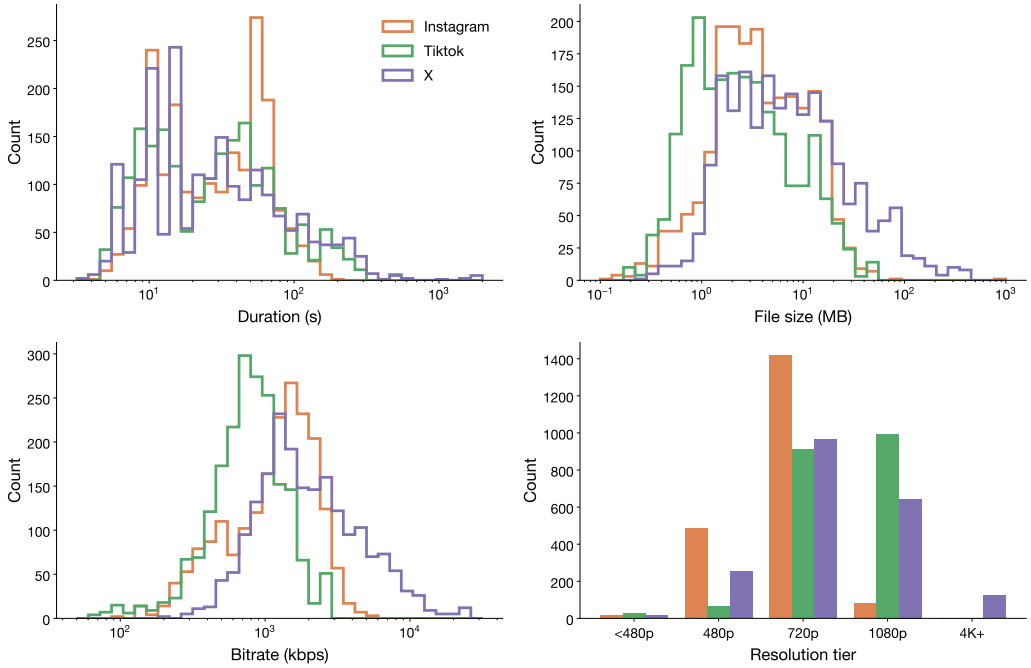


Figure 5: **By platform, pooled across classes.**

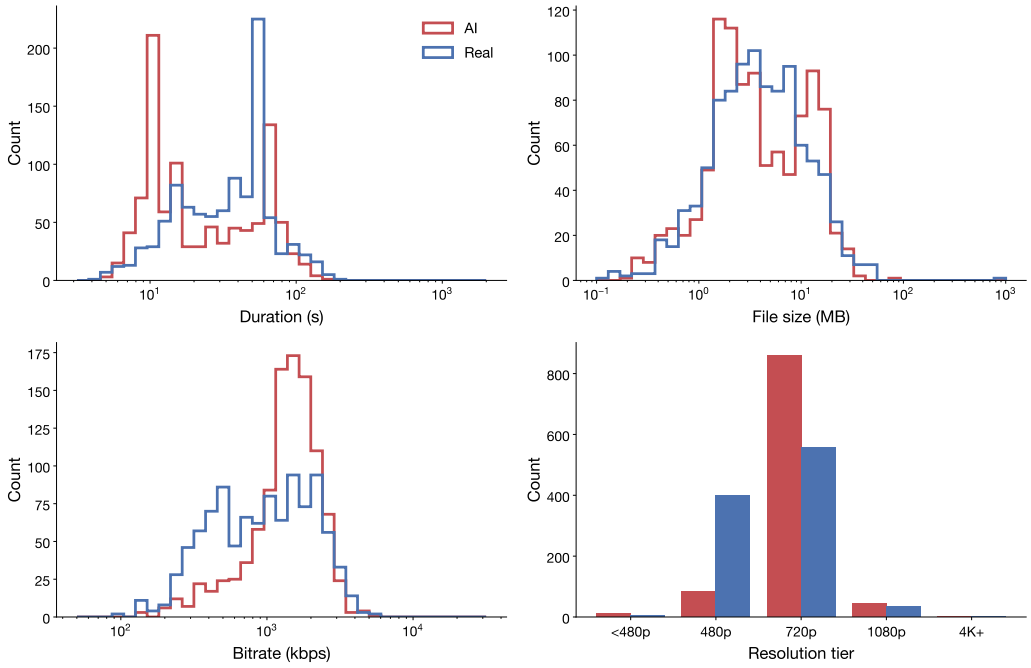


Figure 6: **Instagram: AI vs. real.**

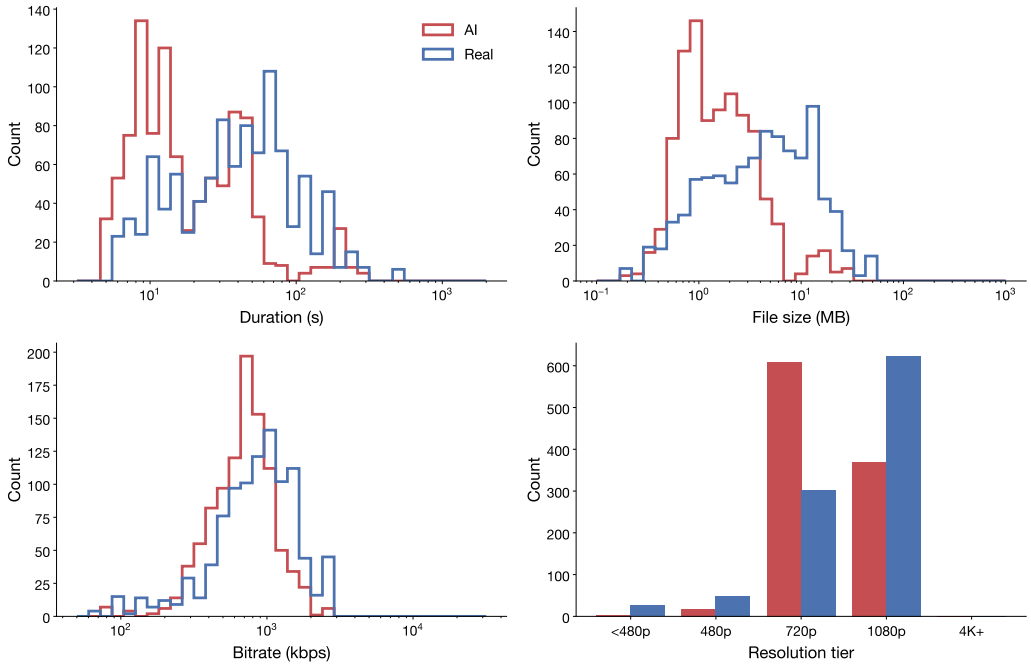


Figure 7: **TikTok: AI vs. real.**

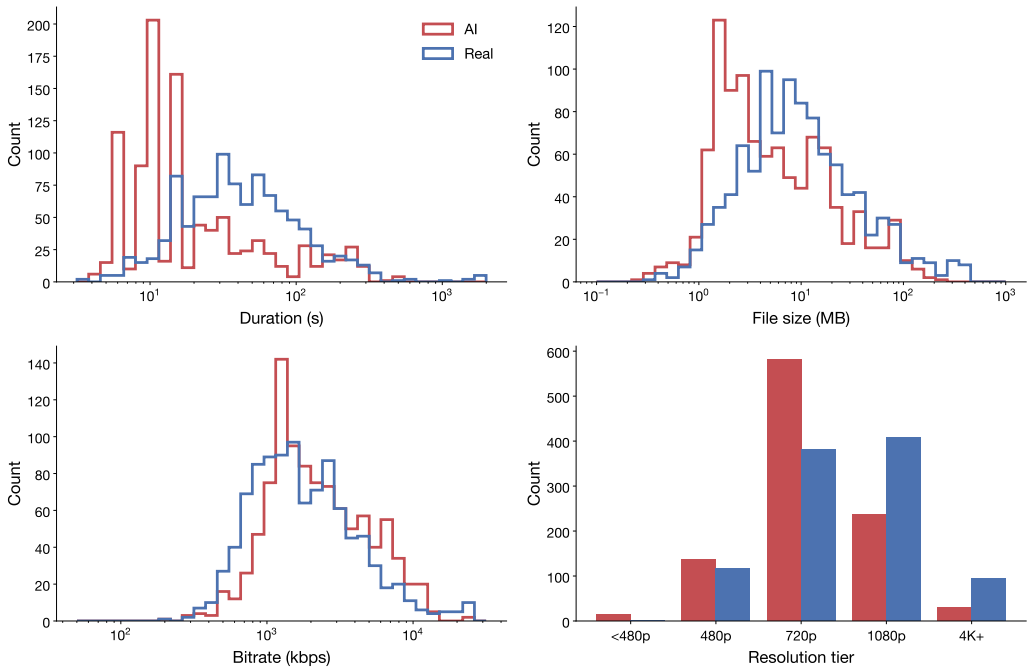


Figure 8: **X: AI vs. real.**

A.3 TAXONOMY

We can characterize the content diversity found on social media, via the sample of our dataset, along three orthogonal axes: *style*, *format*, and *content*. The taxonomy was derived by open coding (Corbin & Strauss, 1990) a random sample of videos drawn from every platform–class cell of SLOP: free-form descriptive tags were assigned per video, affinity-grouped into axes chosen so that every video receives exactly one category per axis, and the category set was expanded until newly sampled videos introduced no categories not already covered, following standard codebook construction practice in content analysis (Krippendorff, 2018).

Style describes the way in which a video is rendered or captured, independent of what it depicts (examples in Appendix A.4.7):

1. **Animated/cartoon**: non-photorealistic rendering with stylized characters or drawn textures.
2. **Realistic/hyperrealistic**: photorealistic rendering presented as camera capture.
3. **CCTV/security camera**: a fixed, wide-angle surveillance viewpoint, typically with timestamp or infrared artifacts.
4. **Found-footage/handheld**: unstabilized first-person camera work such as dashcams or phone-in-hand recording.
5. **Screen-recording**: capture of another device's screen, including its interface elements.

Format describes the overall spatial structure of the frame, independent of rendering and subject (examples in Appendix A.4.8):

1. **Talking head**: a single framed speaker addressing the camera.
2. **Full-form scene**: a full-body subject embedded in a complete scene.
3. **Side-by-side horizontal / vertical**: two simultaneous panels split across or stacked along the frame.
4. **Macro**: an extreme close-up dominating the frame.
5. **Single object**: one centered object of focus in a static composition.
6. **Multi-cut montage**: a rapid sequence of distinct shots.

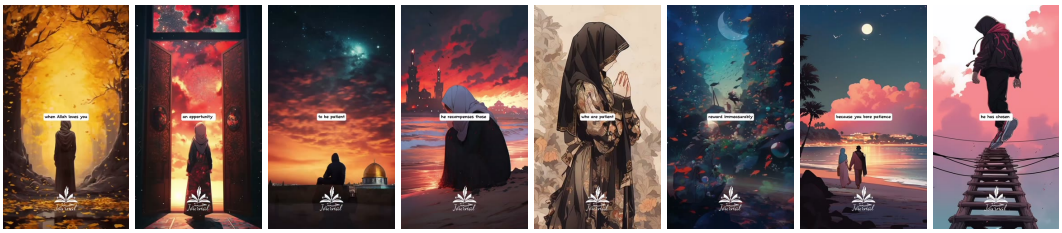
Content describes the subject matter (examples in Appendix A.4.9):

1. **Storytelling**: a narrated fictional or dramatized arc.
2. **Vlog**: a creator documenting their own activity.
3. **Advertisement**: promotion of a product or service.
4. **Interview**: an on-camera exchange between an interviewer and a subject.
5. **Pet/animal**: animals as the primary subject.
6. **Nature**: landscapes or wildlife without a human protagonist.
7. **Inspirational**: motivational messaging addressed to the viewer.
8. **Satisfying**: repetitive, sensory-pleasing processes such as ASMR.
9. **News**: reporting or presenting of current events.

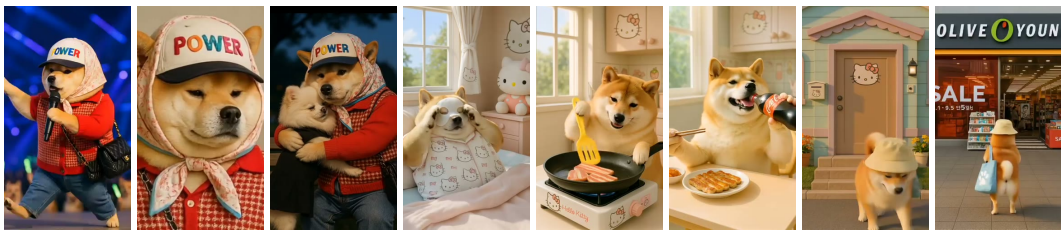
Unlike a flat categorical taxonomy, these dimensions combine freely: the same format (side-by-side) co-occurs with unrelated content (gaming commentary, podcast conversation), and the same content (storytelling) appears in multiple styles (animated, hyperrealistic). This categorization also sharpens our explanation of VRT's difficulty (Table 7): its ASMR-only real class combines a macro/single-object format with a satisfying-content profile that barely exists in the talking-head-heavy distributions detectors are typically trained on.

A.4 EXAMPLES

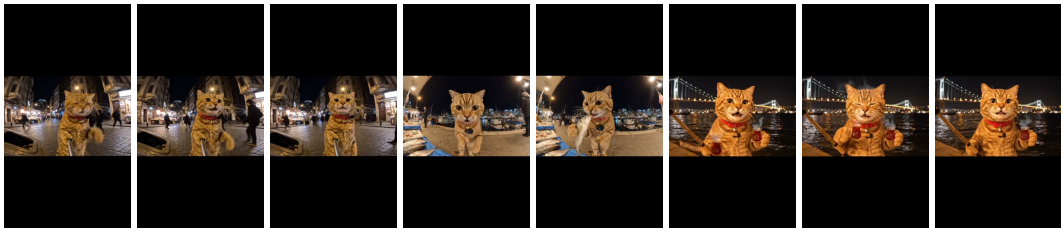
A.4.1 INSTAGRAM - AI

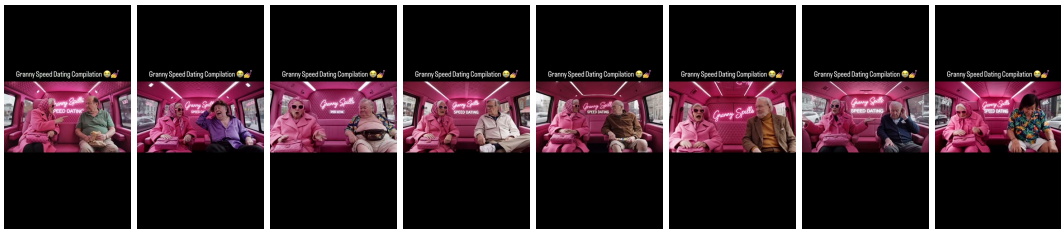


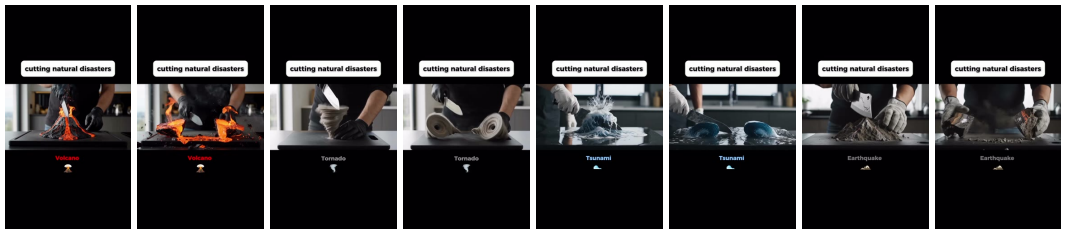






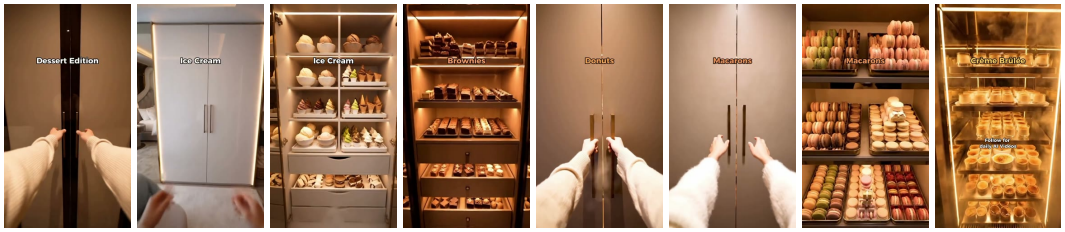








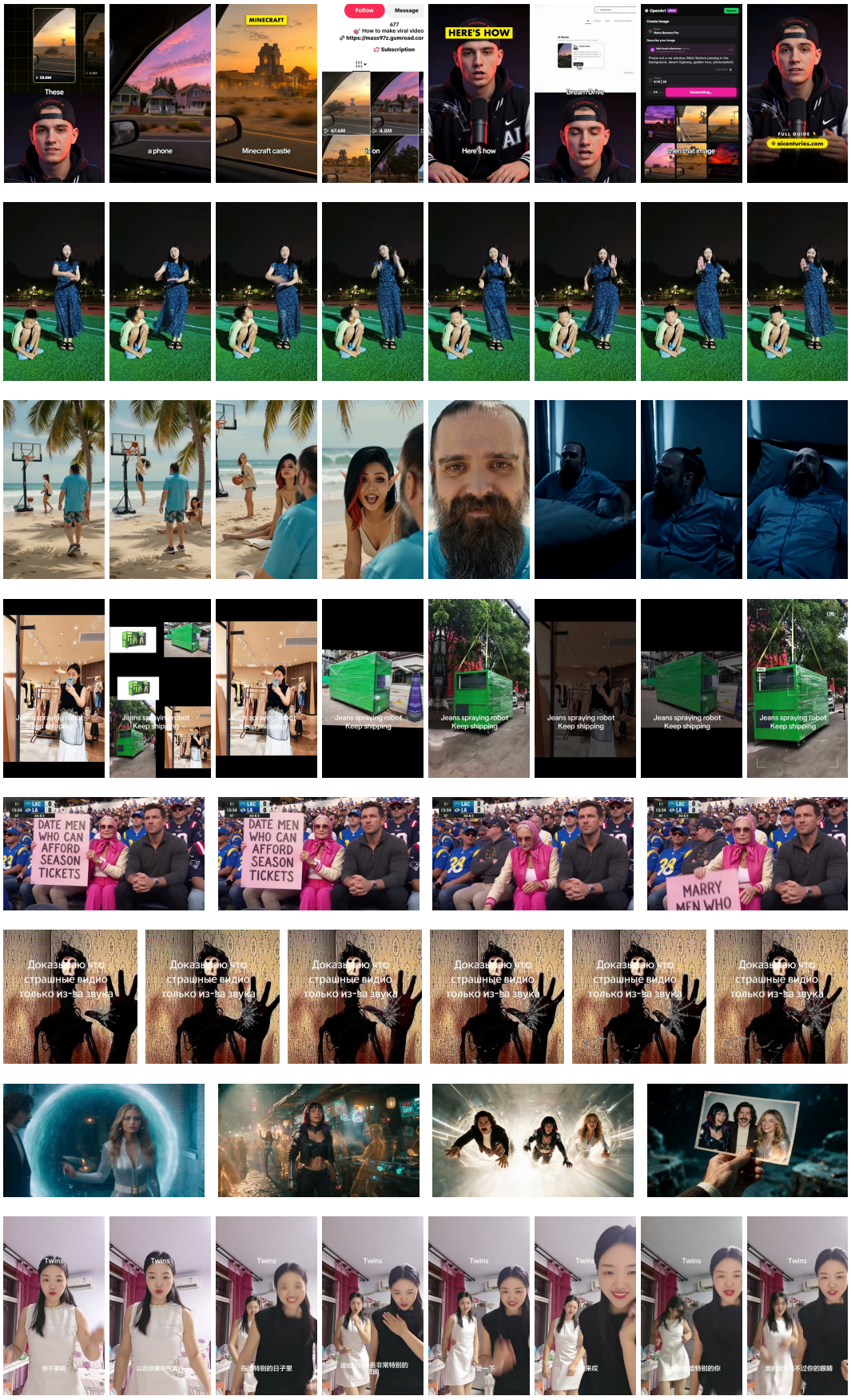






A.4.2 TIKTOK - AI




These
a phone
MINECRAFT
Minecraft castle
Follow
Message
How to make viral video
Subscription
HERE'S HOW
Here's how
Dream Drive
then that image
FULL GUIDE
aicenturies.com
Jeans spraying robot
Keep shipping
DATE MEN WHO CAN AFFORD SEASON TICKETS
MARRY MEN WHO
Доказываю что страшные видио только из-ва звука
Twins







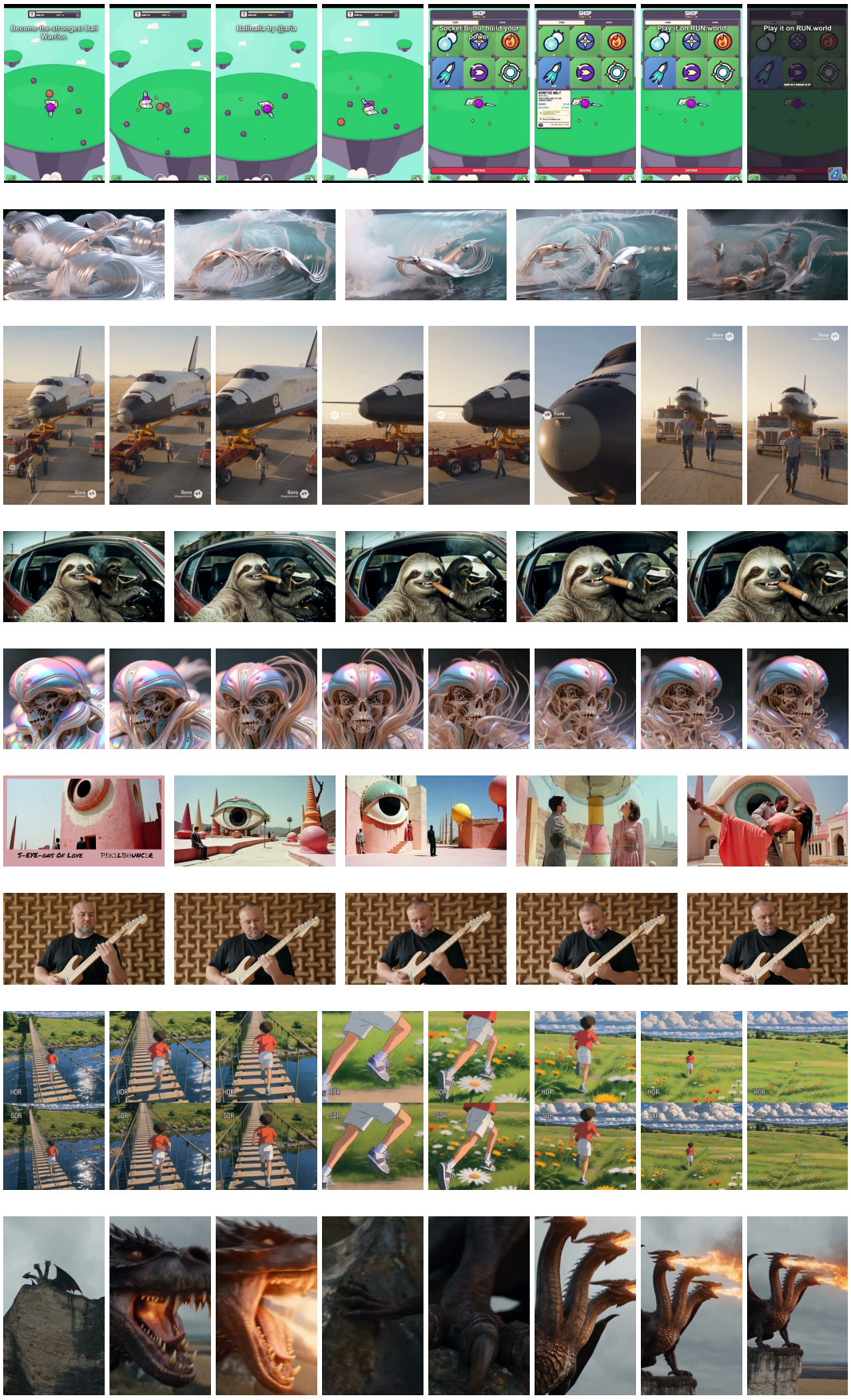
A.4.3 X - AI


IN CONVERSATION WITH
IN CONVERSATION WITH
KELLY BOESCH
21 APRIL
RSVP NOW
LINK IN BIO



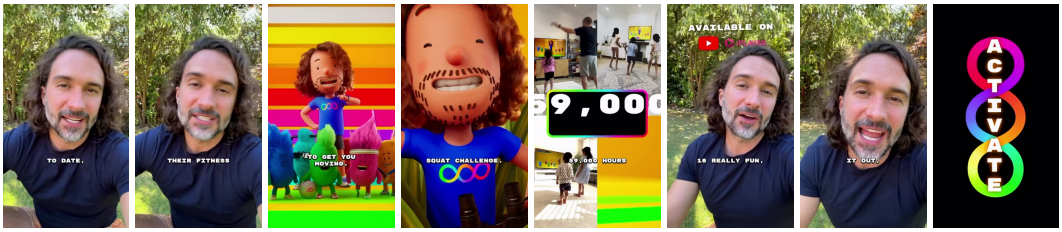


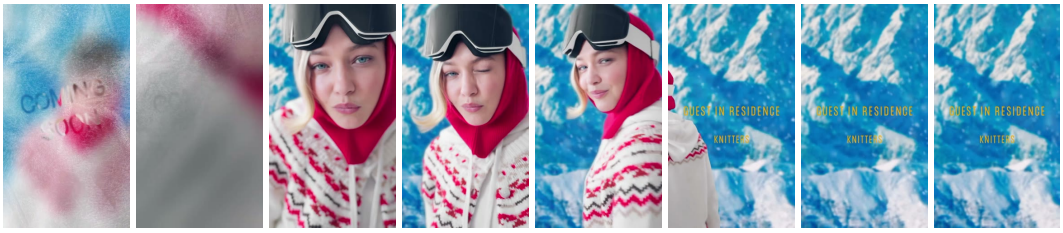



Become the strongest Ball Warrior
Ballhalla by @aria
SHOP
Socket Bijou, build your power
SHOP
Play it on RUN.world
SHOP
Play it on RUN.world
Sora
5-EYE-GUS OF LOVE
PANZLBOUNCER
HDR
SDR



A.4.4 INSTAGRAM - REAL



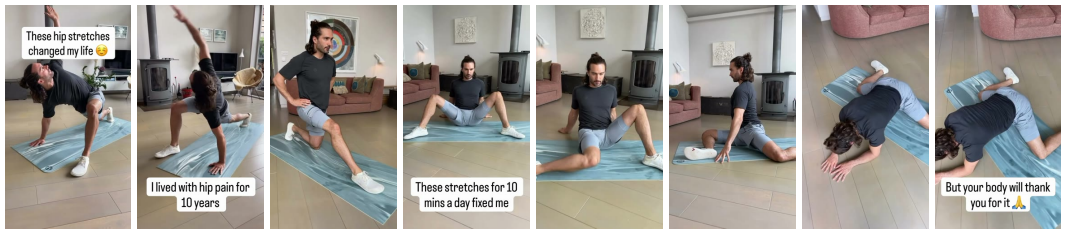














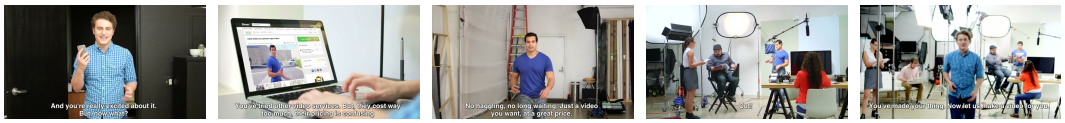


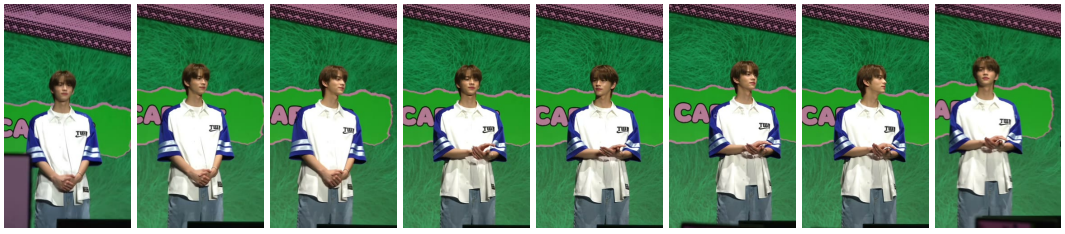


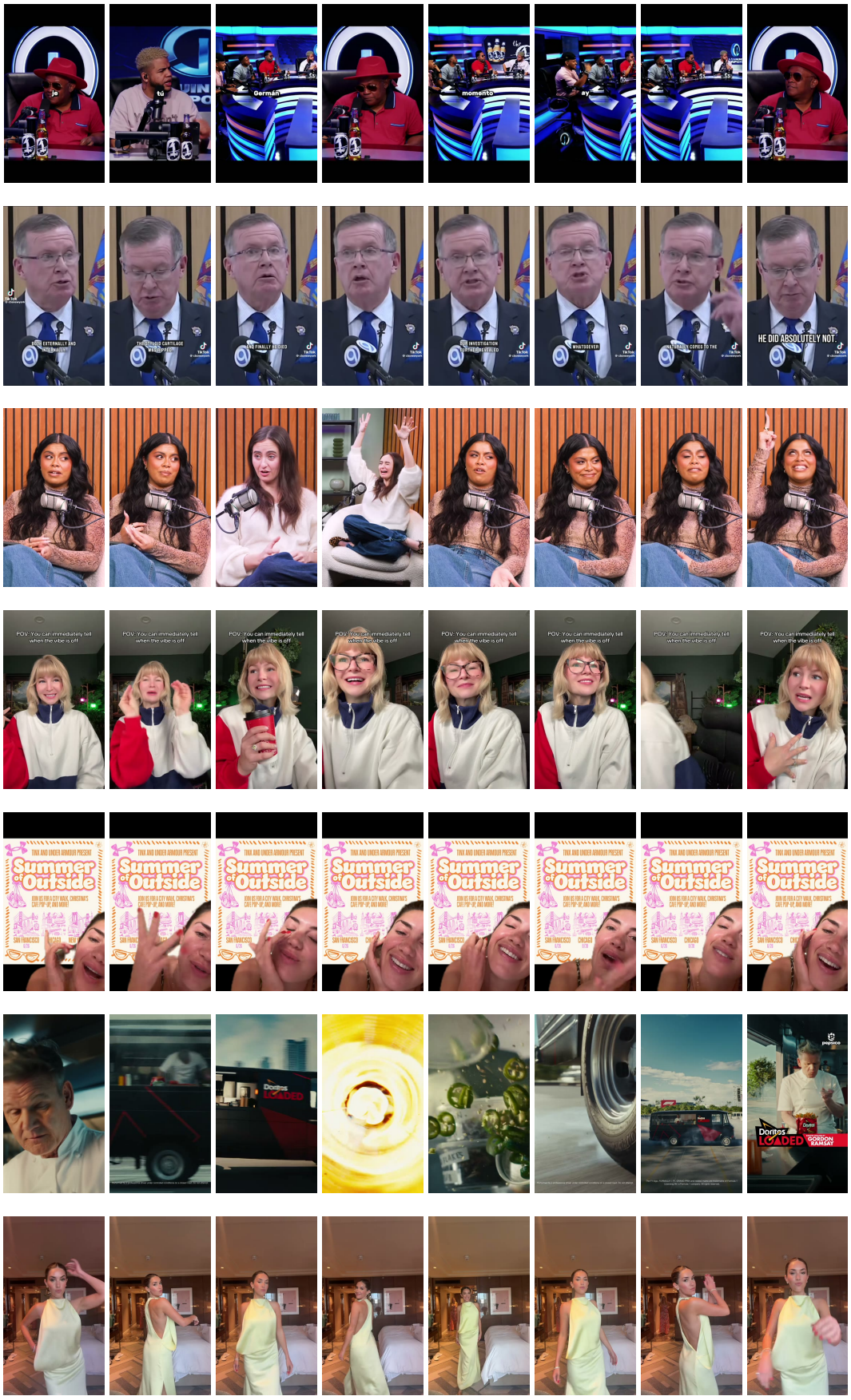


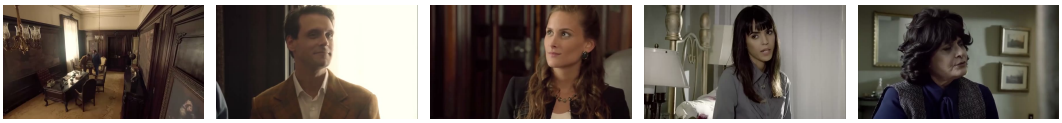
A.4.5 TIKTOK - REAL








HE DID ABSOLUTELY NOT.
POV: You can immediately tell when the vibe is off.
Summer Outside
Doritos LOADED
GORDON RAMSAY



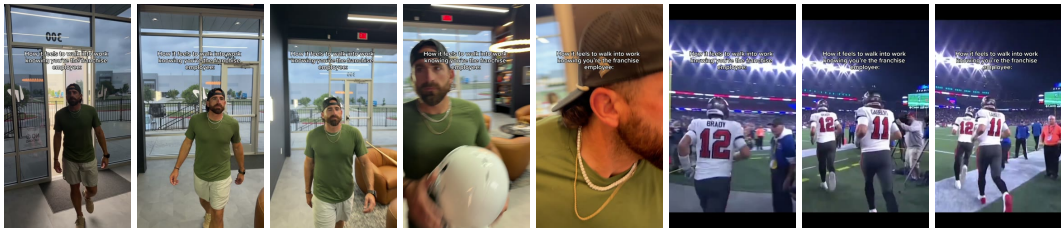


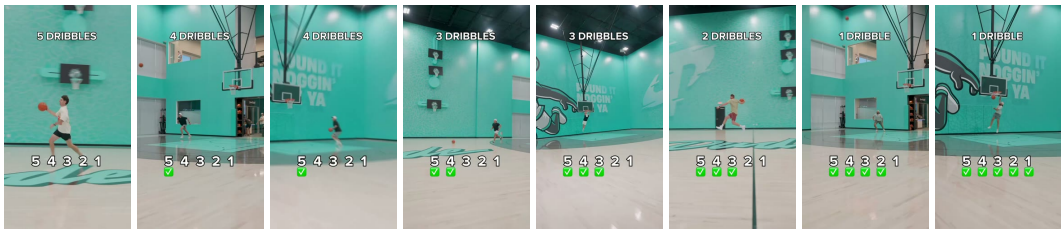
A.4.6 X - REAL

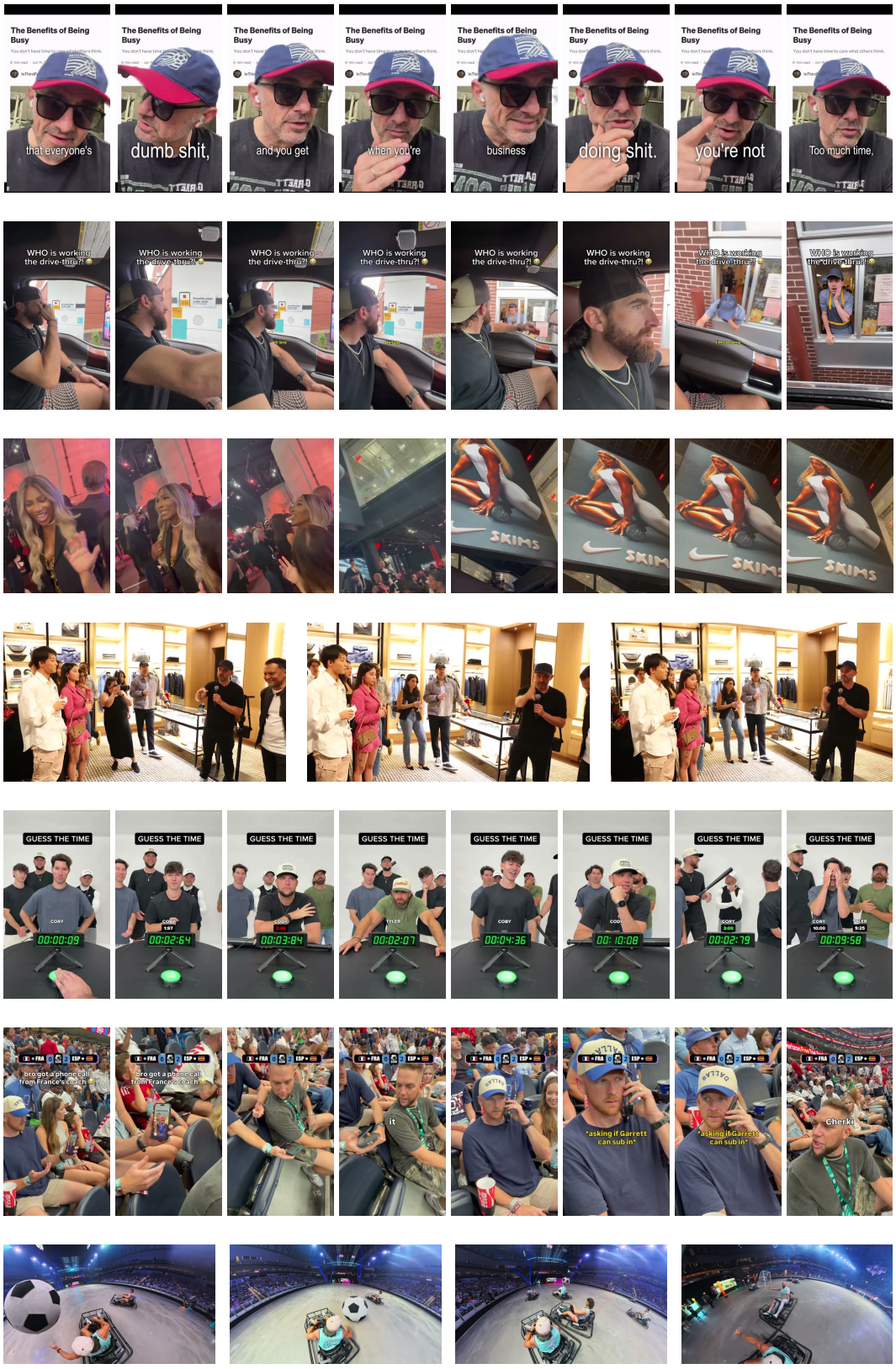









5 DRIBBLES
4 DRIBBLES
4 DRIBBLES
3 DRIBBLES
3 DRIBBLES
2 DRIBBLES
1 DRIBBLE
1 DRIBBLE



A.4.7 STYLE

Animated


Platform: Instagram, Label: AI

Hyperrealistic


Platform: Instagram, Label: AI


Platform: X, Label: AI

Security Camera


Platform: Instagram, Label: AI

Found-footage / Handheld


Platform: Instagram, Label: Real


Platform: Instagram, Label: Real

A.4.8 FORMAT

Full-form Scene

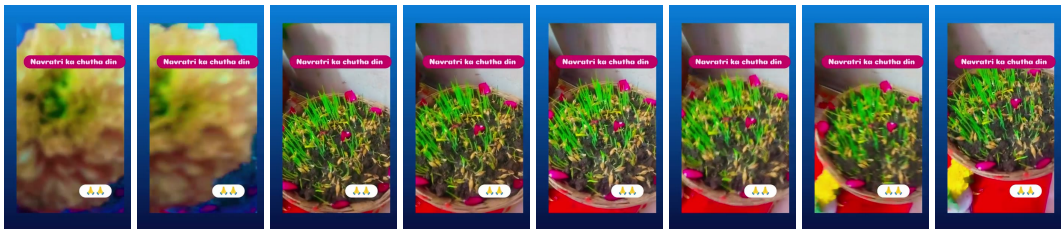


Platform: TikTok, Label: Real



Platform: TikTok, Label: Real

Single Object



Platform: Instagram, Label: Real



Platform: Instagram, Label: Real

A.4.9 CONTENT

Vlog



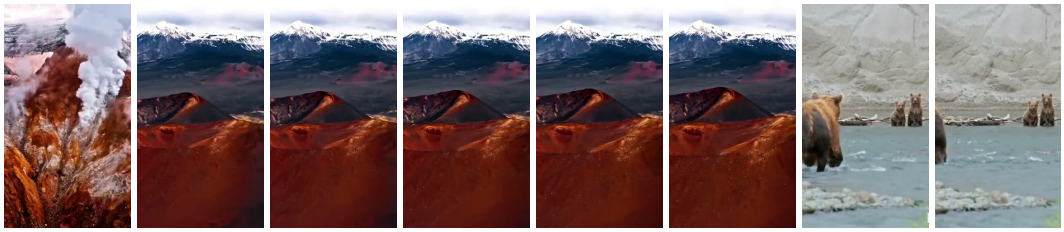
Platform: Instagram, Label: Real


Platform: Instagram, Label: Real

Nature


Platform: TikTok, Label: Real


Platform: TikTok, Label: Real


Platform: TikTok, Label: Real

: content / interview :

Interview



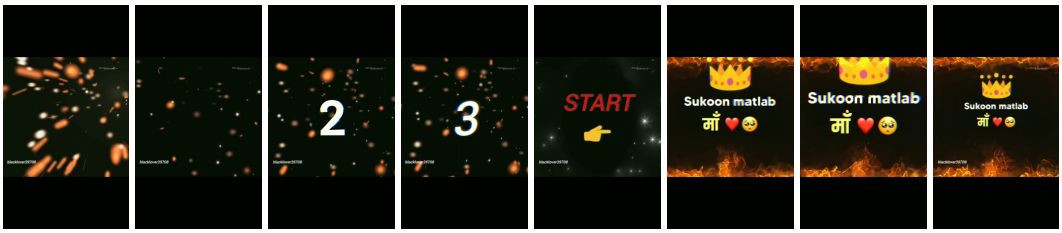
Platform: TikTok, Label: Real

News


Platform: TikTok, Label: Real

B FAILURE CASES

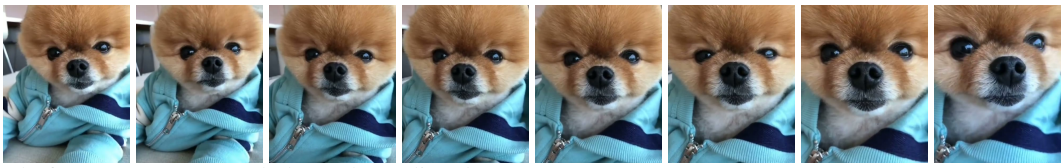
B.1 FALSE POSITIVES



Platform: Instagram, Label: Real, Predicted: AI



Platform: Instagram, Label: Real, Predicted: AI



Platform: Instagram, Label: Real, Predicted: AI



Platform: Instagram, Label: Real, Predicted: AI

B.2 FALSE NEGATIVES



Platform: Instagram, Label: AI, Predicted: Real



Platform: Instagram, Label: AI, Predicted: Real



Platform: Instagram, Label: AI, Predicted: Real



Platform: TikTok, Label: AI, Predicted: Real