# VIDTIDE: A LIVING IN-THE-WILD BENCHMARK FOR AI-GENERATED VIDEO DETECTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Video detectors that succeed on laboratory benchmarks can fail on videos shared online, where generators, editing, and platform processing change over time. We introduce VIDTIDE, a benchmark that tracks this changing population through successive, fixed releases. Its April 2026 slice contains 21,504 clips with provenance-based labels; a blind audit finds 96.4% agreement and 0.6% clear label errors. Five detector families score 52–71% AUROC on a balanced evaluation subset, compared with their published 93–96% on GenVideo. Standard backbones trained on VIDTIDE reach 90.5% mean AUROC with linear probes and 91.7% when selecting the better of probing and fine-tuning. Controlled evaluations examine compression, metadata shortcuts, and newer multimodal models; temporal and cross-dataset tests assess transfer. We release versioned manifests and download scripts, and measure URL availability to document reconstruction limits. The results support updating training-data coverage while preserving fixed test sets. Code and data: `https://anonymous.4open.science/r/vidtide-anon-C0B8/`.

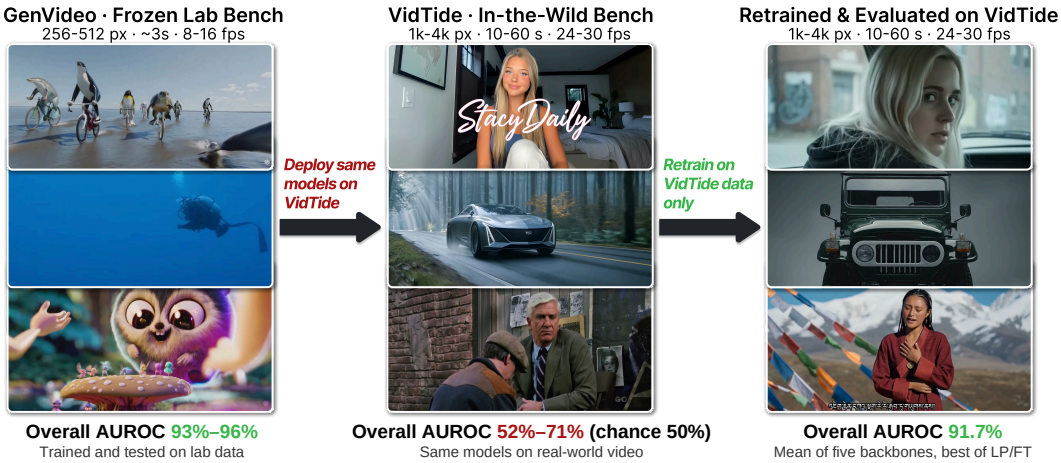



Figure 1: **The static benchmark illusion.** Performance falls from laboratory benchmarks to platform video, while training on VIDTIDE improves detection. The right panel reports the mean of the five backbones' best LP/FT results (Table 6). Adaptation uses a separate model group and evaluation split.

## 1 INTRODUCTION

Can a detector trained on yesterday's generated videos recognize what people share today? Creators edit videos, platforms compress them, and users repost them. New generators such as Sora 2 (OpenAI, 2025), Veo 3 (Google DeepMind, 2025), and Runway Gen-4 (Runway Research, 2025) also change the content itself. Figure 1 illustrates why evaluation on directly generated laboratory clips may not predict performance on current platform video.

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.

A fixed collection captures this mismatch once. VIDTIDE is designed to measure it repeatedly. We collect full-frame text-to-video and image-to-video content from public platforms and official galleries, then freeze each release as a *slice*. Methods can be compared on identical clips within a slice and tested again on later slices. Published slices remain fixed; new collection adds another slice rather than changing an earlier test set.

The first slice, M0, contains 21,504 clips from the April 2026 release. Labels retain their source evidence and are checked through human and automated audits. We publish a *manifest* of URLs, labels, hashes, and metadata with retrieval scripts. Since source videos can disappear, reproducibility includes measuring availability and identifying the clips actually tested.

Existing detectors score 52–71% AUROC on our evaluation pool, versus 93–96% in their GenVideo publications: a mean difference of 31.0 percentage points. AUROC measures how well generated videos are ranked above real ones (50% is chance; 100% is perfect). The comparison documents poor transfer, while controlled tests examine possible explanations. Training standard backbones on VIDTIDE improves detection, but transfer back to laboratory data remains incomplete.

Our contributions are (1) **fixed releases** from recurring, auditable collection; (2) **audited evaluation** of existing detectors with label, compression, and metadata controls; and (3) **adaptation and transfer** experiments that measure the value and limits of the collected training data.

## 2 RELATED WORK

**Benchmarks and the deployment gap.** FaceForensics++ (Rössler et al., 2019) and DeepfakeBench (Yan et al., 2023) focus on face manipulation. GenVideo (Chen et al., 2024) and GenVidBench (Ni et al., 2026) extend evaluation to full-frame generation. AEGIS (Li et al., 2025) studies realistic generated content, while RobustSora (Wang et al., 2025) examines watermark shortcuts. Deepfake-Eval-2024 (Chandra et al., 2025), our closest comparator, already demonstrates large performance drops on web media. VIDTIDE adds recurring releases, recorded label evidence, and generator-resolved analysis within the narrower scope of full-frame video generation (Appendix Table 11).

**Detectors and continuing evaluation.** Our experiments cover Mamba models (Chen et al., 2024), spatial and temporal inconsistency methods (Gu et al., 2021), thumbnail layouts (Xu et al., 2023), frequency cues (Tan et al., 2024), diffusion priors (Zhang et al., 2025), and prompted vision-language models (Qwen Team, 2025a;b; 2026). LiveBench (White et al., 2025) and LiveCodeBench (Jain et al., 2024) illustrate continuing evaluation in other domains. Video forensics adds challenges in source-label reliability, platform processing, and reconstruction of URL-based releases; these are part of our evaluation protocol.

Table 1: **Benchmark landscape.** "Gen." = named generators with per-clip provenance; "GT" = ground-truth source (*lab* = generated in-house, *hand* = hand-curated, *platform* = platform-mandated AI disclosure, $\times 3$ = three-tier hierarchy); "Wild" = in-the-wild sourcing; "Upd." = ongoing updates. Only VIDTIDE combines wild sourcing, platform-anchored multi-tier ground truth, and monthly frozen slices. Deepfake-Eval-2024's video count is back-computed from 45 h at $\sim$30 s mean duration; its image and audio modalities are not comparable to video benchmarks.

| Benchmark | Year | Total | Gen. | GT | Res. | Dur. | Wild | Upd. |
|---|---|---|---|---|---|---|---|---|
| GenVideo (Chen et al., 2024) | 2024 | 18.6K | 10 | lab | 256–512 | $\sim$3 s | ✗ | ✗ |
| GenVidBench (Ni et al., 2026) | 2026 | 6.78M | 11 | lab | 256–720 | $\sim$5 s | ✗ | ✗ |
| AEGIS (Li et al., 2025) | 2025 | 10K+ | 8 | lab+hand | 512–1080 | $\sim$8 s | part. | ✗ |
| RobustSora (Wang et al., 2025) | 2025 | 6.5K | 4 | lab | 480–1080 | $\sim$8 s | ✗ | ✗ |
| Deepfake-Eval (Chandra et al., 2025) | 2025 | $\sim$5K (45 h) | — | hand | mixed | $\sim$30 s | ✓ | ✗ |
| **VIDTIDE (Ours)** | 2026 | 21.5K / 36.5K[†] | 24 | **platform $\times 3$** | 720–1080 | $\sim$15 s | ✓ | **Mo.** |

[†]Frozen M0 slice (21.5K) / live crawl at submission (36.5K).

# 3 THE VIDTIDE BENCHMARK

## 3.1 COLLECTION AND PUBLICATION

VIDTIDE separates a changing collection process from fixed evaluation releases. Figure 2 follows a clip from its public source to a published manifest. The pipeline retains both the video identifier and the evidence supporting its label, so later audits can distinguish known provenance from weaker metadata-based judgments.

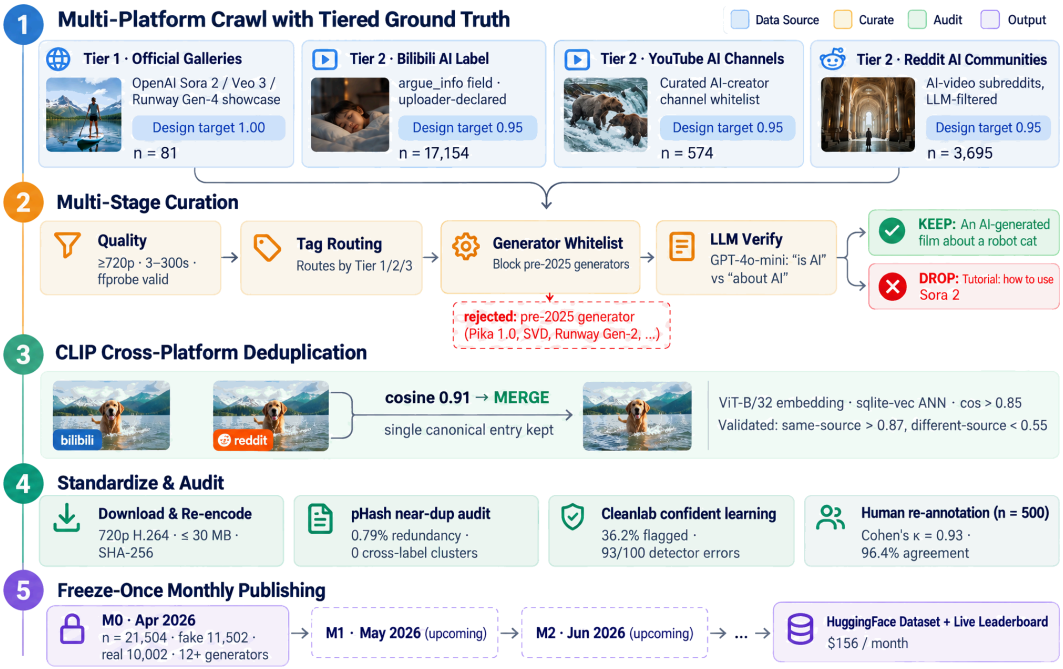



Figure 2: **The VIDTIDE pipeline.** Collection from four platform categories, quality and metadata filtering, CLIP deduplication, standardization, human review, and frozen monthly manifests. HuggingFace hosts metadata, URLs, hashes, and retrieval scripts; videos are fetched from their original platforms. Tier 3 (64 clips) consists of YouTube and Bilibili posts whose titles and descriptions are checked by GPT-4o-mini.

**Sources and evidence.** Generated-video candidates come from official galleries, Bilibili AI disclosures, curated AI channels on YouTube, and AI-video communities on Reddit. Real sources include Kinetics-400 (Kay et al., 2017) and platform content without an AI tag. Missing disclosure is weaker evidence than a known source: it requires auditing and is not treated as proof of authenticity (§4). Generator identity is recorded when the source provides it, rather than inferred for every clip.

**Curation and duplicate removal.** Candidates are checked for video validity, quality, and source evidence. A generator whitelist targets recent systems. For community posts, GPT-4o-mini reviews titles and descriptions to distinguish generated videos from discussions *about* generation tools. This step uses metadata, not independent visual ground truth. Reposts are matched using CLIP ViT-B/32 thumbnail embeddings at cosine similarity 0.85, complementing perceptual-hash checks that can miss re-encoded copies.

**Publication.** Retained clips are encoded to 720p H.264 for review, hashed with SHA-256, and linked to logged collection runs. Publication fixes the clip list and recorded labels. Future collection produces another slice; availability changes are recorded separately so that a retired URL does not erase the record of an earlier evaluation. Filtering counts and crawler settings are in Appendix Tables 13 and 15.

### 3.2 The First Slice and Its Evaluation Subsets

The April 2026 slice, M0, contains 11,502 generated and 10,002 real clips (Table 2). Bilibili supplies 79.8% of the release. Most generated clips (74.1%) carry no generator tag. These properties limit the release's coverage and motivate the platform and generator analyses in §5.5.

Table 2: **Composition of the frozen release.** All counts refer to the same 21,504 clips in the released manifest. Physical video properties from the larger crawl pool are documented separately in Appendix Table 14.

| Source | Clips | Share | Labels present |
|---|---|---|---|
| Bilibili | 17,154 | 79.8% | Generated and real |
| Reddit | 3,695 | 17.2% | Generated only |
| YouTube | 574 | 2.7% | Generated only |
| Official galleries | 81 | 0.4% | Generated only |

Experiments use a nearly class-balanced evaluation pool and a stratified train/test split:

$$\underbrace{\mathcal{D}_{\text{eval}}}_{9{,}956} = \underbrace{\mathcal{D}_{\text{train}}}_{7{,}957} \dot{\cup} \underbrace{\mathcal{D}_{\text{test}}}_{1{,}999}. \tag{1}$$

Here $\dot{\cup}$ denotes disjoint union. Existing detectors use M0-Eval without adaptation; adapted models train on M0-Train and are evaluated on M0-Test. Smaller controlled tests state their own sample sizes. The split files are released with the manifest (Appendix A.5).

### 3.3 Reconstruction and Availability

A fixed manifest does not guarantee a fixed set of retrievable videos. Three months after the freeze, 1,060/1,100 crawled URLs remained retrievable (96.36%), compared with 88/100 Kinetics-400 reference URLs (88.00%; Figure 3). The 40 crawl failures comprised 13 recoverable and 27 permanent failures. Access depends on network location as well as source deletion (Appendix C.5).

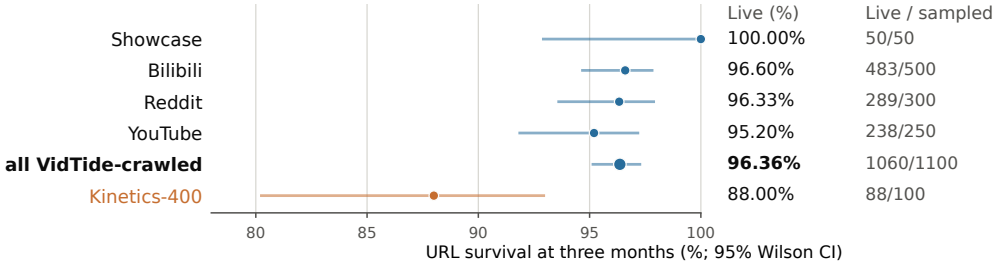



Figure 3: **Measured URL availability after three months.** Points show successful retrieval fractions with Wilson 95% intervals. Kinetics-400 is a separately sampled reference. These are audit outcomes, not guaranteed future survival rates.

Each evaluation records the manifest version and retrieved clip list. Comparisons across retrieval snapshots use their shared clips and report this subset. Hashes check byte-identical files; frame fingerprints check content after re-encoding. Platform-specific AUROC and their macro-average expose source imbalance, separately from pooled AUROC. A source with only one class requires an explicit real reference to define an AUROC. Leaderboard comparisons use a common retrieval snapshot and include a private subset.

## 4 Label Evidence and Quality

Public-video labels have different evidence sources. Tier 1 covers official galleries and imported dataset labels (81 clips). Tier 2 uses platform evidence (21,359 clips): AI disclosures (7,134 generated), curated AI channels and subreddits (4,223 generated), and the weaker absence-of-tag evidence for real clips (10,002). Tier 3 uses titles and descriptions reviewed by GPT-4o-mini (64). Tiers record provenance; they are not measured probabilities that a label is correct.

Table 3: **Ground-truth evidence tiers.** Counts refer to the frozen M0 release; independent checks follow in Table 4.

| Tier | Evidence source | Clips | Examples and limitations |
|---|---|---|---|
| 1 | Known provenance | 81 | Official galleries; imported real datasets |
| 2 | Platform evidence | 21,359 | AI disclosures; curated AI channels and subreddits; missing tags are weaker evidence for real clips |
| 3 | Title/description review | 64 | GPT-4o-mini checks titles and descriptions |

**Independent human review.** Three trained annotators reviewed 500 stratified clips, equally divided between benchmark labels, without seeing titles, platforms, or tiers. Table 4 separates agreement from clear errors: ambiguous cases and shared human mistakes can remain even when annotators agree (Appendix B.2).

Table 4: **Label checks answer different questions.** Human agreement measures consistency with independent judgments; clear errors are a narrower category. The absence-of-tag audit targets undisclosed generation in the real class.

| Check | Sample | Result |
|---|---|---|
| Majority agreement with pipeline | 500 clips | 482/500 (96.4%) |
| Clear label errors | Same 500 | 3/500 (0.6%) |
| Mean pairwise Cohen's $\kappa$ | Three annotators | 0.92 |
| Undisclosed generation estimate | 200 untagged Bilibili clips | 1.5–2.5% |

The separate untagged-video audit has a reported 95% interval of [0.8%, 5.7%]. It supports retaining these real labels at Tier 2 and repeating audits as collection changes. Cohen's $\kappa$ is a chance-adjusted agreement measure, not an accuracy estimate.

**Automated checks.** A perceptual-hash scan of 2,281 clips found 0.79% within-label redundancy and no cross-label duplicate clusters at the tested threshold. A separate *cleanlab* (Northcutt et al., 2021) pass flagged 36.2% of 13,835 scored clips; 93 of its 100 most suspicious examples were real-labelled clips (85 Bilibili, 8 Kinetics-400) scored as generated. Such disagreement can arise from either detector or label error, so the flag rate is not a label-error estimate (Appendix Table 18).

**Sensitivity and scope of the bound.** Suppose a fraction $\varepsilon \leq \frac{1}{2}$ of each observed class comes from the opposite true class, independently of detector score within each true class. Then

$$\mathrm{AUROC}_{\mathrm{observed}} \leq 1 - \varepsilon, \qquad 0.948 \leq 1 - \varepsilon \implies \varepsilon \leq 0.052. \tag{2}$$

The 94.8% result is from the later temporal slice, so the bound applies to that slice under this model, not to every possible error in M0. The human audit does not establish the model's independence assumption. Derivations and the corruption test are in Appendix C.1. Re-encoding removes container metadata, but does not prove the absence of pixel-level watermarks; source controls are discussed in Appendix B.3.

## 5 Experiments

We evaluate GenVideo-trained DeMamba, NSG-VD, STIL, NPR, and TALL checkpoints, plus zero-shot CLIP. Separately, five Kinetics-400-pretrained backbones are adapted on M0-Train: DeMamba, VideoSwin-Tiny (Liu et al., 2022), TSM (Lin et al., 2019), I3D, and SlowFast. AUROC is in percent; pp denotes percentage points. Intervals use 5,000 clip-level bootstrap resamples, paired for identical test clips (Appendix Table 24).

### 5.1 Existing Detectors Transfer Poorly

Table 5 compares published laboratory results with measurements on M0-Eval. Pika and Crafter denote checkpoint training variants, not test subsets. The five Pika variants average 63.5% on

VIDTIDE, a 31.0 pp gap from their published GenVideo results. This comparison includes differences in data and source protocols.

Table 5: **Published laboratory performance and measured platform-video performance.** A/b/F = AUROC/balanced accuracy/F1 (%); bACC and F1 use threshold 0.5. Both checkpoint variants use the same 9,956-clip M0-EVAL. GenVideo reports published AUROC; $\Delta$ is measured minus published AUROC (pp). Pika/Crafter denote training variants, not test subsets.

| Model | GenVideo | Pika checkpoint (A/b/F) | $\Delta^{\text{AUROC}}_{\text{Pika}}$ | Crafter checkpoint (A/b/F) | $\Delta^{\text{AUROC}}_{\text{Crafter}}$ |
|---|---|---|---|---|---|
| DeMamba | 93.9 | **71.1** / 55.8 / 23.4 | −22.8 | **67.4** / 54.1 / 20.3 | −26.5 |
| NSG-VD | 96.1 | 52.5 / 52.3 / 39.9 | −43.6 | 48.8 / 49.3 / 41.8‡ | −47.3 |
| STIL | 93.5 | 61.2 / 50.4 / 2.4 | −32.3 | 63.5 / 51.6 / 10.3 | −30.0 |
| NPR | 93.0 | 69.7 / 65.0 / 59.5 | −23.3 | 49.0 / 48.4 / 47.1 | −44.0 |
| TALL | 95.7 | 62.8 / 52.4 / 12.3 | −32.9 | 53.3 / 51.4 / 12.2 | −42.4 |
| *Generator-agnostic baseline (no Pika/Crafter distinction):* | | | | | |
| CLIP ViT-L/14 (zero-shot) | — | VIDTIDE 60.5 / 53.8 / 65.0 | | (A / b / F, same evaluation pool) | |
| **Avg (5 supervised)** | **94.4** | **63.5** / 55.2 / 27.5 | −31.0 | **56.4** / 51.0 / 26.3 | −38.0 |

‡NSG-VD Crafter checkpoint is not publicly released; we trained it ourselves on the official GenVideo Crafter split (details in Appendix D.2) following the authors' protocol unchanged. The 31.0 pp headline number is the Pika-checkpoint average gap. Per-checkpoint bACC and F1 are also isolated in Appendix Table 25.

At threshold 0.5, the five Pika checkpoints average 55.2% balanced accuracy and 27.5% F1 (Appendix Table 25). Ranking degradation therefore coincides with poor decisions at this operating point.

### 5.2 LEARNING FROM THE COLLECTED VIDEO

A *linear probe* (LP) trains a classifier on frozen features; full fine-tuning (FT) also updates the backbone. Both use M0-TRAIN and M0-TEST. Table 6 reports the full results, including variability, available historical checkpoints, and gains. Data and training procedure change together.

Table 6: **Changing the training data, not the architecture.** AUROC (%) on M0-TEST (1,999 clips; mean ± std over 3 seeds for LP). "Original" = GenVideo-trained checkpoint; LP and FT use M0-TRAIN. Available original checkpoints are scored on the same test clips. Training data and procedure change together, so the paired differences do not isolate the effect of data alone. The final column pair reports best-of-LP/FT results.

| Model | Original | LP | $\Delta_{\text{LP}}$ | Best LP/FT | $\Delta_{\text{best}}$ |
|---|---|---|---|---|---|
| DeMamba (Chen et al., 2024) | 69.4 | **94.7**±0.3 | **+25.3** | 94.7‡ | +25.3‡ |
| VideoSwin-Tiny | N/A* | **92.8**±0.4 | N/A | 92.8‡ | N/A |
| TSM | N/A* | 89.8±0.5 | N/A | **95.4**±0.6 | N/A |
| I3D | N/A* | 86.4±0.7 | N/A | 86.4‡ | N/A |
| SlowFast | N/A* | 89.0±0.4 | N/A | 89.0‡ | N/A |
| NSG-VD (Zhang et al., 2025)† | 52.5 | 52.3 | −0.2 | N/A | N/A |
| **Average** (excl. NSG-VD) | – | **90.5** | – | **91.7** | – |

*No pre-existing GenVideo-trained checkpoint. ‡FT did not surpass LP under our single-GPU budget (A100-40GB, 20 epochs); we report $\max(\text{LP, FT})$ as a descriptive best-of-two summary, not a preselected method. An lr sweep ($\{5\times10^{-5}, 10^{-4}, 3\times10^{-4}\}$) on TSM confirmed $10^{-4}$; other backbones diverged or plateaued at LP level by epoch 5. †NSG-VD has no fully fine-tunable pathway: retraining its MMD discriminator on VIDTIDE collapsed to AUROC 0.50, consistent with its diffusion-prior design.

### 5.3 TRAINING-DATA COVERAGE AND TRANSFER

Figure 4 varies training-set size on a fixed test split, then evaluates fixed linear probes on different test populations.

**How much data is useful?** With 10% of M0-TRAIN (approximately 800 clips), DeMamba reaches 90.2% AUROC. All three tested backbones improve with more training clips. Sample count, generator, platform, and content diversity change together, so this ablation does not isolate their individual effects (Appendix Table 33).

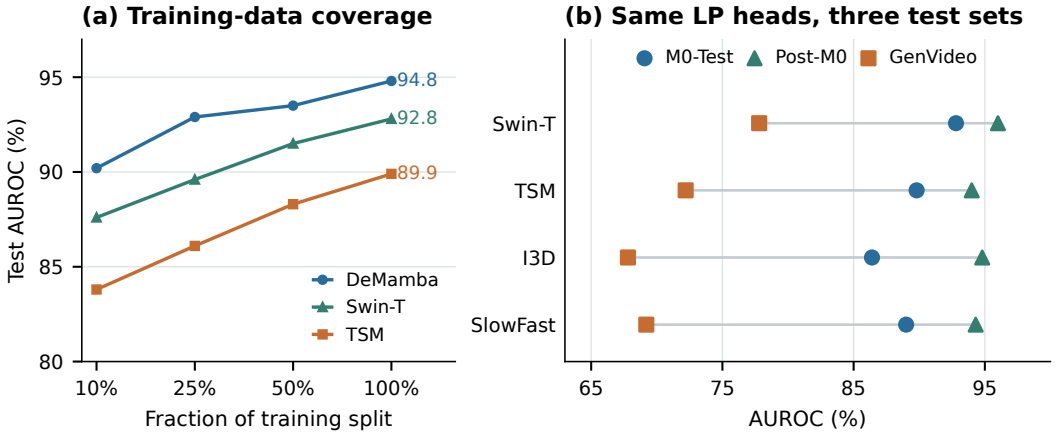



Figure 4: **Training coverage helps, but transfer depends on the target.** (a) Existing data-scaling ablation on M0-TEST (Appendix Table 33). (b) Four fixed linear probes evaluated on M0-TEST, a 99-clip temporal holdout, and GenVideo-Val. Each marker is a separate AUROC; connecting lines group the same backbone and do not imply interpolation between test sets.

**Does the benefit persist over time?** We evaluate four LP heads—Swin, TSM, I3D, and SlowFast—without retraining on a later collection of 50 generated and 49 real clips. These clips were collected after the M0 split closed and deduplicated against it. Mean AUROC is 94.8%, suggesting useful temporal transfer in this collection (Appendix Table 29). The set is small, however, and may share platform and content structure with M0. Its high score should not be read as evidence that all future generators will be equally easy to detect.

**Does the benefit transfer to laboratory video?** The same four LP heads average 71.7% on GenVideo-Val, with individual scores of 67.8–77.8% (Appendix Table 30). Their corresponding mean on M0-TEST is 89.5%. Comparing these fixed models exposes a remaining cross-dataset limitation. The historical GenVideo-to-VIDTIDE experiment uses a different detector group and source protocols, so differences between the two transfer directions do not establish a universal ranking of the training datasets.

**Is platform identity enough?** Restricting M0-TEST to its 1,721 Bilibili clips changes the four probes by at most 0.8 pp (Appendix Table 32). This reduces one source of between-platform confounding. It does not remove differences within Bilibili, such as subject matter or editing style. The next experiments control measured metadata and additional compression more directly.

### 5.4 CONTROLLED EVALUATIONS

The following tests vary one part of the evaluation at a time. They use different clip subsets, so Figure 5 shares an AUROC scale but does not pool their measurements. Each panel identifies the tested factor; the text states what remains outside that control.

**Additional compression.** We take a fixed 16-second window from each of 800 label-balanced test clips and re-encode it at eight bitrate settings, keeping the content and detector heads fixed. The realized range is approximately 228–10,892 kbps. Across this 48-fold range, the mean endpoint drop for four VIDTIDE-trained backbones is 1.73 pp (95% CI [0.84, 2.66]); CLIP varies by 0.55 pp. Most change occurs at the lowest bitrates. These shifts are smaller than the historical 31.0 pp gap, but the model groups differ. Since platform processing predates collection, this experiment cannot isolate damage from earlier compression, editing, or reposting (Appendix C.2).

**Measured metadata shortcuts.** A classifier using only platform, duration, size, resolution, frame rate, and bitrate scores 83.85% AUROC on the broader evaluation pool. For the controlled test we use M0-TEST, where its starting score is 77.50%. We stratify by duration and bitrate, add platform to the strata, then match real and generated clips on all three properties, retaining 486 pairs. Metadata-only

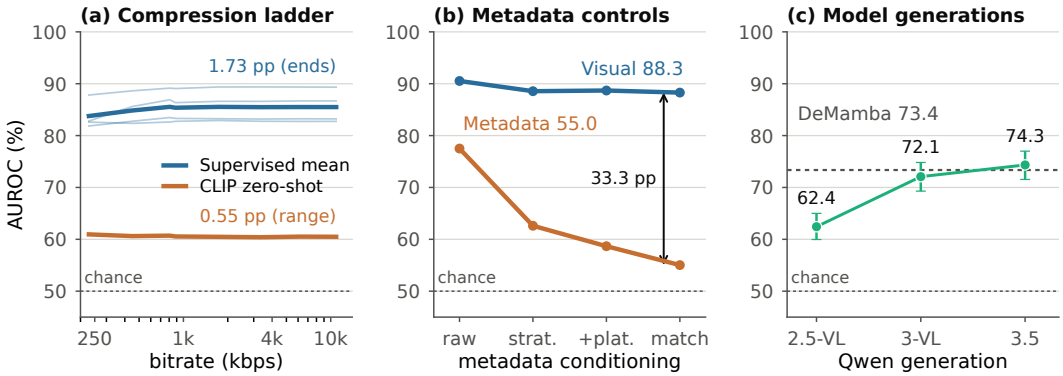



Figure 5: **Compression, metadata, and model-generation controls.** (a) The same 800 clips at eight bitrates. (b) Metadata conditioning on M0-TEST, ending with 486 matched real/generated pairs. (c) Zero-shot Qwen models on the same 1,000 clips; error bars are paired-bootstrap 95% intervals, and the reference is DeMamba-Pika on those clips.

AUROC falls to 55.03% (95% CI [51.47, 58.58]), while the five visual probes fall from 90.53% to 88.29% (95% CI [86.61, 89.89]). Their matched advantage is 33.26 pp (95% CI [29.51, 37.02]). Useful visual signal remains after these controls; unmeasured content and style differences may still matter (Appendix C.3).

**Newer multimodal models.** Qwen2.5-VL-7B (Qwen Team, 2025a), Qwen3-VL-8B (Qwen Team, 2025b), and Qwen3.5-9B (Qwen Team, 2026) receive the same prompt and eight frames per clip on a balanced 1,000-clip subset. With deterministic decoding and no adaptation, AUROC rises from 62.43% to 72.08% to 74.34%. The earliest-to-latest improvement is 11.91 pp (95% CI [8.64, 15.16]). The newest model is close to DeMamba-Pika at 73.36% on identical clips: their difference is 0.98 pp (95% CI [−2.90, 4.71]). Model generation matters, though size and other design changes accompany release date. Adaptation scores use another test set and are not drawn as a same-pool reference (Table 7).

Table 7: **All detectors re-scored on the shared 1,000-clip subset.** AUROC (%). Specialist detectors are GenVideo-trained checkpoints evaluated zero-shot; MLLM and CLIP rows are zero-shot general-purpose models. Adaptation results use a different test split and are reported separately in Table 6.

| Model | AUROC | 95% CI | Model | AUROC |
|---|---|---|---|---|
| **Qwen3.5-9B** (2026-02) | **74.34** | [71.54, 77.00] | STIL-Crafter | 62.65 |
| DeMamba-Pika (2024) | 73.36 | [70.24, 76.47] | Qwen2.5-VL-7B (2025-01) | 62.43 |
| NPR-Pika (2024) | 72.58 | — | STIL-Pika | 60.98 |
| Qwen3-VL-8B (2025-10) | 72.08 | [69.29, 74.82] | NSG-VD-Crafter | 56.15 |
| DeMamba-Crafter | 67.93 | — | TALL-Crafter | 52.32 |
| TALL-Pika | 64.49 | — | NSG-VD-Pika | 51.97 |
| CLIP ViT-L/14 zero-shot | 64.04 | — | NPR-Crafter | 49.36 |

### 5.5 VARIATION ACROSS GENERATORS AND PLATFORMS

An overall score hides changes in both model ranking and difficulty. Table 8 compares detectors across generators and platforms. Each cell is a separate evaluation slice; the same detector can behave very differently across columns.

Averaged over the eleven rows of Table 8, AUROC is 58.0% for Dreamina 3, 58.6% for Sora 2, 65.0% for Kling 2.1, and 72.1% for Veo 3. Most generated clips lack a generator tag, so we also expand the analysis with high-confidence family predictions. Six of eight detectors retain rank correlation of at least 0.9. This is a sensitivity check, not verified generator provenance (Appendix E.9).

The corresponding platform means are 68.5% for Reddit, 58.1% for Bilibili, and 57.0% for YouTube. Content, generators, and processing change together across sources, so this ordering cannot identify a single cause. In a separate ensemble analysis, 60.5% of generated and 55.4% of real clips have mean

Table 8: **Cross-generator and cross-platform detection.** AUROC on a 0–1 scale; color ranges from red (lower) to green (higher), with numeric values in every cell. P/C denote Pika/Crafter training variants. Columns are evaluation slices, not controlled interventions; each column compares that slice's generated clips with the full real pool.

| | Kling 2.1 | Veo 3 | Sora 2 | Dreamina 3 | Runway G4 |
|---|---|---|---|---|---|
| CLIP | 0.63 | 0.74 | 0.53 | 0.58 | 0.75 |
| DeMamba / C | 0.73 | 0.77 | 0.63 | 0.64 | 0.75 |
| DeMamba / P | 0.83 | 0.95 | 0.71 | 0.65 | 0.92 |
| NPR / C | 0.55 | 0.61 | 0.54 | 0.52 | 0.54 |
| NPR / P | 0.76 | 0.75 | 0.68 | 0.71 | 0.75 |
| NSG-VD / C | 0.46 | 0.59 | 0.45 | 0.49 | 0.43 |
| NSG-VD / P | 0.51 | 0.55 | 0.53 | 0.55 | 0.53 |
| STIL / C | 0.69 | 0.69 | 0.61 | 0.65 | 0.59 |
| STIL / P | 0.70 | 0.74 | 0.63 | 0.56 | 0.68 |
| TALL / C | 0.57 | 0.70 | 0.52 | 0.46 | 0.65 |
| TALL / P | 0.72 | 0.84 | 0.62 | 0.57 | 0.83 |

(a) Detector × generator.

| | Bilibili | Reddit | YouTube |
|---|---|---|---|
| CLIP | 0.57 | 0.68 | 0.79 |
| DeMamba / C | 0.66 | 0.77 | 0.50 |
| DeMamba / P | 0.66 | 0.92 | 0.79 |
| NPR / C | 0.48 | 0.57 | 0.42 |
| NPR / P | 0.69 | 0.76 | 0.51 |
| NSG-VD / C | 0.49 | 0.50 | 0.41 |
| NSG-VD / P | 0.52 | 0.54 | 0.49 |
| STIL / C | 0.65 | 0.64 | 0.50 |
| STIL / P | 0.60 | 0.69 | 0.57 |
| TALL / C | 0.50 | 0.64 | 0.60 |
| TALL / P | 0.57 | 0.82 | 0.69 |

(b) Detector × platform.

correct-class confidence below 0.6; these examples are difficult for the ensemble, but confidence also depends on calibration (Table 9; Appendix E.8).

Table 9: **Per-generator hard-case breakdown.** Hard fakes (mean detector confidence <0.6) on the VIDTIDE M0 fake pool, broken down by claimed generator. "Other / unidentified AI" aggregates platform-confirmed-AI clips for which our metadata pipeline could not pin down a specific generator (typically Bilibili `argue_info` positives without a recognizable model name in title/description).

| Generator | # Hard | # Total | Hard % | Dominant Failure Mode |
|---|---|---|---|---|
| Dreamina 3 | 113 | 172 | **65.7%** | Hyper-realistic, professional grading |
| Pika 2 | 39 | 61 | **63.9%** | Short clips, low motion |
| Sora 2 | 197 | 311 | **63.3%** | Hyper-realistic, cinematic camera work |
| Runway Gen-4 | 61 | 133 | 45.9% | Stylized, painterly artifacts |
| Kling 2.1 | 229 | 513 | 44.6% | Stylized, exaggerated motion |
| Veo 3 | 134 | 370 | 36.2% | Identifiable temporal inconsistencies |
| Other / unidentified AI | 3,933 | 6,182 | 63.6% | Mixed: Bilibili re-encoding + unknown model |
| **Total (fakes)** | **4,825** | **7,973** | **60.5%** | — |

## 6 LIMITATIONS AND CONCLUSION

**Limits of the current evidence.** M0 is Bilibili-dominant, most generated clips lack verified generator identity, and disclosures can be incomplete. The temporal test has only 99 clips; reverse transfer remains below in-distribution performance. Controls cover selected metadata and encoding, not every content, editing, or watermark cue. URL availability limits reconstruction.

**Conclusion.** VIDTIDE combines recurring collection with fixed, auditable releases. Existing detectors transfer poorly to the first slice; standard backbones benefit from training on its data. Controlled and transfer evaluations delimit this benefit. Frozen slices support repeated, comparable testing on newly collected video without guaranteeing future generalization.

## ETHICS STATEMENT

VIDTIDE uses publicly posted videos and public platform metadata. The release contains source URLs and labels rather than video files. Creator handles are not separate indexed fields, but source URLs can reveal the account that posted a video; the release therefore does not anonymize the underlying posts. The opt-out policy provides for requests to be honored within 24 hours, with affected rows marked as retired. Source deletion also limits future retrieval, although it cannot remove copies already downloaded by users.

Crawling uses public endpoints and documented APIs with rate limiting. Collection and release policies are documented in Appendix A.10. Three trained graduate-student annotators performed the human audit on public video content. A possible misuse of the benchmark is to optimize generated videos against the evaluated detectors. A private leaderboard subset reduces direct exposure of evaluation examples but does not eliminate this risk.

## REPRODUCIBILITY STATEMENT

The anonymous project repository provides the pipeline, slice manifests, evaluation scripts, and experimental artifacts: `https://anonymous.4open.science/r/vidtide-anon-C0B8/`. Appendix Tables 24 and 26 document models, splits, and training settings. The compression, metadata, multimodal-model, label-noise, and availability analyses are detailed in Appendix C.2, C.3, C.4, C.1, and C.5, respectively. Published GenVideo scores are attributed to their original papers and distinguished from measurements made in this study.

Performance intervals use 5,000 clip-level bootstrap resamples with a fixed seed; paired comparisons resample the same clips for both methods. URL-survival intervals use the Wilson method, and small audit proportions use Clopper–Pearson intervals. Reconstructing an evaluation requires its manifest version, retrieval status, and access to the source platforms. Section 3.3 specifies how to compare partially retrievable slices without silently changing the test population.

## AI USE STATEMENT

We used large language models as writing assistants. They helped edit wording and grammar, tighten paragraphs and captions, format tables and the conference template, and check that numbers and cross-references were consistent across the main text and appendix. The authors reviewed and revised all suggested text.

## REFERENCES

Nuria Alina Chandra, Ryan Murtfeldt, Lin Qiu, Arnab Karmakar, Hannah Lee, Emmanuel Tanumihardja, Kevin Farhat, Ben Caffee, et al. Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. *arXiv preprint arXiv:2503.02857*, 2025.

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

Cyberspace Administration of China and Ministry of Industry and Information Technology and Ministry of Public Security. Provisions on the administration of deep synthesis internet information services. `https://www.cac.gov.cn/2022-12/11/c_1672221949354811.htm`, 2022. Issued 25 November 2022; effective 10 January 2023.

Cyberspace Administration of China and others. Measures for labeling artificial intelligence-generated synthetic content. `https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm`, 2025. Issued 14 March 2025; effective 1 September 2025.

Timnit Gebru et al. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Google DeepMind. Veo 3: High-fidelity video generation. `https://deepmind.google/technologies/veo/`, 2025.

Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Spatiotemporal inconsistency learning for DeepFake video detection. In *Proceedings of the 29th ACM International Conference on Multimedia (ACM MM)*, 2021. arXiv:2109.01860.

Naman Jain et al. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.

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

Jieyu Li, Xin Zhang, and Joey Tianyi Zhou. AEGIS: Authenticity evaluation benchmark for ai-generated video sequences. In *Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM)*, 2025. arXiv:2508.10771.

Ji Lin, Chuang Gan, and Song Han. TSM: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

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

Zhenliang Ni, Qiangyu Yan, Mouxiao Huang, Tianning Yuan, Yehui Tang, Hailin Hu, Xinghao Chen, and Yunhe Wang. GenVidBench: A 6-million benchmark for ai-generated video detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2026. arXiv:2501.11340.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021.

OpenAI. Sora 2. `https://openai.com/index/sora-2/`, 2025.

Qwen Team. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025a.

Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025b.

Qwen Team. Qwen3.5: Towards native multimodal agents. `https://qwen.ai/blog?id=qwen3.5`, February 2026.

Andreas Rössler et al. FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Runway Research. Introducing gen-4. `https://runwayml.com/research/introducing-runway-gen-4`, 2025.

Chuangchuang Tan, Huan Liu, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Wenhao Wang and Yi Yang. VidProM: A million-scale real prompt-gallery dataset for text-to-video diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. arXiv:2403.06098.

Zhuo Wang, Xiliang Liu, and Ligang Sun. RobustSora: De-watermarked benchmark for robust ai-generated video detection. *arXiv preprint arXiv:2512.10248*, 2025.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, et al. LiveBench: A challenging, contamination-limited LLM benchmark. In *International Conference on Learning Representations (ICLR)*, 2025. Spotlight. arXiv:2406.19314.

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

Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. DeepfakeBench: A comprehensive benchmark of deepfake detection. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023. arXiv:2307.01426.

Shuhai Zhang, ZiHao Lian, Jiahao Yang, Daiyuan Li, Guoxuan Pang, Feng Liu, Bo Han, Shutao Li, and Mingkui Tan. Physics-driven spatiotemporal modeling for ai-generated video detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. Spotlight. arXiv:2510.08073.

# A Benchmark Construction and Documentation

## A.1 Operating Cost and Sustainability

A living benchmark is only valuable if it can be maintained. We address sustainability through cost engineering, graceful degradation, and open governance.

**Monthly cost breakdown.** Table 10 itemizes the monthly operating budget.

Table 10: **Monthly operating cost.** VidTide operates at ∼$156/month in its standard configuration, well below the $1,500–2,000/month budget typical of industry-curated video datasets.

| Component | Cost | Notes |
|---|---|---|
| Crawling compute | $0 | Negligible CPU; no API fees |
| LLM review (Tier 3) | $50 | GPT-4o-mini, ∼1,000 candidate screenings; ∼6% pass confidence threshold (M0 retained 64) |
| Automated audits | $0 | pHash + cleanlab, CPU-only |
| GPU (detector evaluation) | $100 | ∼50 GPU-hours × $2 |
| HuggingFace hosting | $0 | Free for public datasets |
| Infrastructure (Fly.io) | $6 | Shared-CPU VM + 5 GB volume |
| **Total** | **$156** | |

**Three budget modes.** These are budget settings, unrelated to the ground-truth tiers of Section 4: VidTide is designed to degrade gracefully:

- **Minimal ($50–100/month)**: Only open-source generators and Tier 1+2 ground truth (Section 4), no LLM review. Sustainable by a single researcher.
- **Standard (∼$156/month)**: Multi-platform crawling, full three-tier GT, automated pHash + cleanlab audits. The target operating mode.
- **Full ($1,500–2,000/month)**: All platforms, watermark removal experiments, cross-encoding normalization analysis, plus optional two-person re-annotation. Industry-level budget.

**Freeze-once guarantee.** Each monthly slice, once published to HuggingFace, is immutable. If monthly updates cease, every previously published slice remains a valid, reproducible, standalone benchmark. This structural guarantee means that *sustainability failures do not damage existing scientific outputs*.

**Open-source governance.** The entire pipeline—crawlers, filters, dedup, evaluation scripts—is open-source and released at the anonymous repository `https://anonymous.4open.science/r/vidtide-anon-C0B8/`, with a live leaderboard at `https://vidtide.fly.dev/`. Any research group can fork the repository and run an independent instance. New generators and platforms are proposed via GitHub issue templates. Slice provenance is verifiable through HuggingFace's git commit history.

## A.2 Benchmark Landscape (Extended)

Table 1 summarises the landscape; it omits face-manipulation benchmarks and detailed column descriptions for space. DeepfakeBench (Yan et al., 2023) is a unified evaluation framework spanning 9 face-manipulation datasets (FF++, Celeb-DF, DFDC, etc.) with 15 detection methods. GenVideo's Sora entry refers to Sora v1 (2024), not Sora 2 (2025). Duration and FPS values are approximate

where not explicitly reported in the original papers. The key takeaway: only Deepfake-Eval (Chandra et al., 2025) and VIDTIDE source content from the wild, but only VIDTIDE provides ongoing monthly updates with verified ground truth.

**Detailed contrast with Deepfake-Eval-2024.** Because Deepfake-Eval-2024 is the closest prior in-the-wild benchmark, we restate the differentiation in tabular form (Table 11). The two benchmarks complement rather than substitute for each other: Deepfake-Eval-2024 establishes *that* an in-the-wild gap exists across modalities and languages, while VIDTIDE provides the *generator-resolved, rolling, platform-anchored* infrastructure that lets the community keep measuring the gap as new generators come online.

Table 11: **Detailed contrast: Deepfake-Eval-2024 vs. VIDTIDE.** Both benchmarks crawl in-the-wild content; the rest of the column structure is where they differ.

| **Axis** | **Deepfake-Eval-2024 (Chandra et al., 2025)** | **VIDTIDE (Ours)** |
|---|---|---|
| Modalities | Video (45 h) + audio (56.5 h) + image (1,975) | Video only (21,504 clips, ∼15 s mean) |
| Sourcing scope | 88 websites, 52 languages, 2024 social-media + detection-service uploads | YouTube, Bilibili, Reddit, official showcase galleries (4 platforms) |
| Deepfake type | "Circulating" deepfakes: face swap, voice clone, lip-sync, full-frame mixed | Full-frame T2V/I2V from named modern generators |
| Generator metadata | Not disaggregated by generator | 24 named generators (Sora 2, Veo 3, Kling 2.1, Dreamina 3, Pika 2, Runway Gen-4, Hailuo, Luma, Wan 2.1, Vidu, Pixverse, Hunyuan, . . . ) |
| Ground-truth source | Hand-curated single-tier labels per upload | Three-tier confidence hierarchy: Showcase > Platform-disclosed (YouTube AI tag, Bilibili `argue_info`, China DSR) > Reverse-evidence (Section 4) |
| Update cadence | Fixed 2024 snapshot | Monthly slices (M0, M1, M2, . . . ) |
| Cross-dataset metric | 45–50 pp AUC drop on prior models | 31 pp AUC drop on GenVideo-trained models (Table 5); recoverable with retraining (Table 6) |

### A.3 DATASET DOCUMENTATION (DATASHEET)

Following Gebru et al. (Gebru et al., 2021), we provide a datasheet for VIDTIDE.

**Motivation.** VIDTIDE was created to address the temporal staleness of existing AI-generated video detection benchmarks. Its purpose is to serve as a continuously updated evaluation resource for the video forensics community.

**Composition.** The M0 slice contains 21,504 videos: 11,502 AI-generated (53.5%) and 10,002 real (46.5%); subsequent monthly slices target a similar real≈fake balance. AI-generated videos are sourced from YouTube, Bilibili, Reddit, and official generator showcase galleries. Real videos are sourced from Kinetics-400, Pexels, and same-platform non-AI content. Each video entry includes: video ID, source platform, source URL, title, label (`real`/`fake`), label source (tier), label confidence, claimed generator, duration, resolution, frame rate, SHA-256 hash, and crawl timestamp. Videos are re-encoded to a standardized 720p H.264 format.

**Collection process.** Videos are collected via automated crawlers that query platform APIs and web endpoints. Labels come from three evidence tiers: official galleries and imported datasets such as Kinetics-400 (Tier 1); platform-mandated AI disclosure, curated AI channels, and the absence of an AI tag for platform reals (Tier 2); and LLM-verified community labels (Tier 3). Label quality is validated by automated audits (pHash near-duplicate scan + cleanlab confident-learning flags; see Section 4).

**Preprocessing.** Videos are re-encoded to 720p maximum, H.264 codec, with a 30 MB size cap. Thumbnails are extracted for CLIP-based deduplication. No content modification (cropping, color correction) is applied beyond re-encoding.

**Uses.** VIDTIDE is intended for: (1) evaluating AI-generated video detectors on current-generation content, (2) fine-tuning existing detectors with in-the-wild data, and (3) studying the temporal evolution of detector performance. It should *not* be used for: training AI video generators, surveillance, or any purpose that could harm individuals depicted in videos.

**Distribution.** VIDTIDE is distributed via HuggingFace Datasets under CC-BY-NC 4.0. We publish only *metadata* (platform, label, generator, source URL, timestamps, hash, audit-tier) and *download scripts*; we do *not* redistribute the video files themselves. End users run the provided crawler against the original platform URLs to obtain the videos at evaluation time. Tier 1 showcase clips are accessed via the same per-URL mechanism as Tier 2/3 (no special HuggingFace-side video hosting). This minimizes copyright exposure (Section A.10), respects creator opt-outs (a removed source URL in turn removes the clip from any reproduced slice), and follows the URL-based distribution of Kinetics (Kay et al., 2017).

**Maintenance.** Monthly slices are published on a rolling basis. Each slice is immutable once published. The pipeline code is open-source; the community can fork and operate independent instances. A GitHub issue tracker handles opt-out requests (24-hour removal guarantee), error reports, and new platform/generator proposals.

### A.4 M0 SLICE COMPOSITION

Table 12: **M0 slice composition.** Distribution of videos by platform, label, generator, and ground truth tier in the first monthly VIDTIDE slice.

| Dimension | Category | Count |
|---|---|---|
| Platform | Bilibili | 17,154 |
| | Reddit | 3,695 |
| | YouTube | 574 |
| | Showcase | 81 |
| Label | AI-generated (fake) | 11,502 |
| | Real | 10,002 |
| Top Generators (fake) | Kling 2.1 | 883 |
| | Veo 3 | 674 |
| | Sora 2 | 488 |
| | Runway Gen-4 | 228 |
| | Dreamina 3 | 189 |
| | Pika 2 | 82 |
| | Other / Unlabeled | 8,958 |
| GT Tier | Tier 1 (gallery, dataset) | 81 |
| | Tier 2 (platform tag, channel, tag absence) | 21,359 |
| | Tier 3 (LLM) | 64 |
| **Total** | | 21,504 |

### A.5 M0 DATA FLOW: FROM RAW CRAWL TO EVAL SPLITS

The headline experimental tables (5, 6) use the pools summarized in Table 13. The evaluation pool and its train/test split are released as self-contained files in which every record carries its label and source URL; 3,179 evaluation clips, including the 198 Kinetics-400 reals, are listed only in these split files.

The detector-scored pool (13,835) is wider than the active eval pool (9,956) because the multi-detector hard-case analysis (Appendix E.8) opportunistically pooled all videos that had passed through any of the seven detectors over the M0 collection window, including videos later excluded from the class-balanced 5K-real / 5K-fake bench slice so that headline AUROC comparisons would not be affected by class imbalance. The 80/20 LP / FT split is computed on the 9,956-video active pool with stratification along label $\times$ generator so that every named generator appears in both train and test.

Table 13: **VIDTIDE M0 data flow.** "Real / Fake" show the per-class counts at that stage; the bench-5K and 80/20-split rows are class-balanced by design.

| Stage | Total | Real | Fake | Used in | Selection rule |
|---|---|---|---|---|---|
| Released M0 manifest | 21,504 | 10,002 | 11,502 | Datasheet (Sec. A.3) | All released records passing pipeline filters (Sec. 3.1) |
| Detector-scored pool | 13,835 | 5,862 | 7,973 | Hard-case (Tab. 9) | Subset with cached scores from $\geq$7 detectors |
| Active eval pool | 9,956 | 4,972 | 4,984 | Bench-5K (Tab. 5) | Class-balanced 5K / 5K target; 44 inference errors removed |
| $\hookrightarrow$ Train (80%) | 7,957 | 3,972 | 3,985 | LP / FT training (Tab. 6) | Stratified by label $\times$ generator |
| $\hookrightarrow$ Test (20%) | 1,999 | 1,000 | 999 | LP / FT evaluation (Tab. 6) | Disjoint from train; same stratification |

Table 14: **Per-platform physical-dimension composition of the M0 crawl database** (26,908 videos with probe metadata; the live VidTide crawl at submission contains 36,519 videos at `https://vidtide.fly.dev/`), before the class-balanced filtering that produces the 21,504-video published M0 slice (Table 12). Counts ($n$), clip duration (mean / min – max, seconds), average resolution (width$\times$height, pixels), and frame rate (fps). "fake" and "real" sub-rows split each platform by label. Lower half: label-source (GT tier) breakdown aggregated across all platforms.

| Platform | Label | $n$ | Dur. mean / range (s) | Avg. res. (W$\times$H) | Avg. fps |
|---|---|---|---|---|---|
| Bilibili | fake | 14,555 | 92.4 / 5–600 | n/a | n/a |
| | real | 9,336 | 197.1 / 6–600 | n/a | n/a |
| Kinetics-400 | real | 500 | 10.0 / 10–10 | n/a | n/a |
| Reddit | fake | 1,968 | 39.3 / 3–120 | 1322 $\times$ 1087 | n/a |
| YouTube | fake | 549 | 125.6 / 3–598 | 1671 $\times$ 960 | 27.4 |
| **Total** (by label) | fake | 17,072 | 86.0 (pooled) | – | – |
| | real | 9,836 | 187.7 (pooled) | – | – |
| **Total** (all) | – | 26,908 | – | – | – |

*Ground-truth label-source breakdown (all platforms)*

| | | | |
|---|---|---|---|
| tier1_dataset | real | 500 | Kinetics-400 import, provenance-by-construction |
| tier1_gallery | fake | 87 | OpenAI Sora / Google Veo / Runway / DeepMind gallery downloads |
| tier1_platform_absence | real | 9,336 | Bilibili uploads without `argue_info` AI-disclosure tag |
| tier2_channel_whitelist | fake | 4,040 | YouTube / Bilibili AI-channel content (curated 10-channel whitelist) |
| tier2_platform_tag | fake | 14,536 | Bilibili `argue_info` AI-generated flag set by uploader |
| tier3_llm | fake | 63 | Reddit / YouTube title-description LLM classification (GPT-4o-mini) |

*Notes.* The tier-source rows are not disjoint with the per-platform rows above: `tier2_channel_whitelist` (4,040) and `tier2_platform_tag` (14,536) overlap heavily on Bilibili AI-channel uploads that also carry the `argue_info` tag, so they sum to more than the per-platform Bilibili-fake count (14,555). Field names keep their collection-time prefixes; under the evidence tiers of Section 4, `tier1_platform_absence` reals belong to Tier 2. The Bilibili-fake snapshot here (14,555) is the M0 freeze; the live VidTide crawl has since grown to 23,021 Bilibili fakes / 36,519 total videos at `https://vidtide.fly.dev/`.

### A.6 M0 PHYSICAL-DIMENSION COMPOSITION

Table 12 reports the headline volume counts for M0. Table 14 supplies the complementary physical-dimension breakdown (clip duration range, resolution, and frame rate) computed from the 26,908 videos in the M0 database for which probe metadata survived the crawl pipeline (some early Bilibili crawls did not emit resolution/fps probes, indicated by "n/a"). The per-label-source row at the bottom decomposes the same pool along the three-tier GT axis (Section 4).

Two axes of asymmetry are visible. (i) *Duration*: Bilibili reals are on average 2.1$\times$ longer than Bilibili fakes (197 s vs. 92 s), consistent with the Bilibili platform mixing documentary-style real content with short-form AI reels; the 10-second Kinetics-400 reals are far shorter than either. (ii) *Resolution*: Reddit uploads peak at roughly 1080p-landscape (1322$\times$1087 avg. widths skewed by vertical short-form content), YouTube skews toward 16:9 1080p (1671$\times$960), and Bilibili's probe coverage is incomplete in our current crawl (scheduled for M1). These distributional differences motivate the frame-rate and resolution normalization that every detector in our zoo applies in preprocessing (Appendix D.3), and they are the axis along which the per-platform differences of Table 8 operate. As the controlled ladder of Appendix C.2 shows, those differences are largely compositional rather than an effect of re-encoding itself.

Table 15: **Crawler configuration.** Key parameters for each platform's crawling module.

| Platform | Parameter | Value |
|---|---|---|
| YouTube | Search queries | 15 generator-specific queries |
| | Channel whitelist | 41 channels (vendor + prosumer AI creators) |
| | Max videos per run | 600 |
| Bilibili | API endpoint | `/x/web-interface/search/type` |
| | Pages per query | $3 \times 50$ results |
| | Sort order | Publication date (descending) |
| Reddit | Subreddits | 62 AI-video communities |
| | API | Public JSON API |
| Showcase | Vendors | OpenAI Sora 2, DeepMind Veo 3, Runway Gen-4 |
| | Method | Static gallery page scraping |

Table 16: **Platform re-encoding characteristics.** Typical encoding parameters observed for videos uploaded to each platform, measured via ffprobe on a sample of 50 videos per platform.

| Platform | Codec | Typical Bitrate | Max Resolution | Frame Rate |
|---|---|---|---|---|
| YouTube | H.264/VP9 | 2–8 Mbps | 1080p | Original |
| Bilibili | H.264 | 1–4 Mbps | 1080p | Original |
| Reddit | H.264 | 0.5–2 Mbps | 720p | Original |

### A.7 Crawling Configuration

The crawling pipeline is configured via a single YAML file. Key parameters for each platform crawler are documented in Table 15.

### A.8 Platform Re-Encoding Specifications

### A.9 Copyright Safeguards and Quality-Control Layers

Table 17: **Copyright safeguards and quality-control layers in VidTide.** Each row is one defensive layer; section / table references point to the underlying mechanism, audit number, or compliance argument.

| Copyright safeguards | Quality-control layers |
|---|---|
| Metadata-only HuggingFace release; videos fetched from origin URL at eval time, never redistributed (§3.1; Appendix A.3). | Pipeline filters: $\geq$720p, 3–300 s, ffprobe structural validation, generator whitelist, LLM verification (§3.1). |
| 24 h opt-out via GitHub issue tracker; manifest deletion immediately revokes the clip from all reproduced slices (Appendix A.3). | pHash near-duplicate audit on a 2,281-video hold-out: 0.79% within-label redundancy, **0** cross-label clusters (§4, Appendix Table 18). |
| Per-jurisdiction compliance argument: EU GDPR Art. 89 (research exemption), China DSL / PIPL (public AI-disclosure metadata only), US fair-use 4-factor (Appendix A.10). | cleanlab confident-learning audit on 13,835 detector-scored videos; top-100 inspection: 93/7 *detector-error*-vs-label-error skew (§4, Appendix Table 18). |
| Platform Terms-of-Service & `robots.txt` compliance for each crawler (Bilibili, YouTube, Reddit, Showcase; Appendix A.10). | Blinded human re-annotation of 500 stratified videos by 3 trained annotators: Cohen's $\kappa$=0.92, 96.4% agreement with pipeline label, 0.6% clear errors (§4; Appendix B.2). |
| VidProM (Wang & Yang, 2024) CC-BY-NC 4.0 licensing precedent followed for monthly slice releases. | Tier-2 false-negative-rate audit on 200 Bilibili no-tag reals: 1.5–2.5% point estimate; 95% Clopper–Pearson CI $[0.8\%, 5.7\%]$ ("likely AI") and $[0.3\%, 4.3\%]$ ("definitely AI") (Appendix B.2). |

Table 18: **Automated label-quality audits on VIDTIDE M0.** Three complementary audits, each at the bottom of an independent pipeline: **(A) Cleanlab** confident-learning flags on ensemble fake-probability vs. crawled label (top-100 "most-suspicious" inspected manually); **(B) pHash near-duplicate** detection (perceptual-hash Hamming-distance$\leq$6 clusters, which bound the redundancy rate of the crawl); **(C) Detector-ensemble hard-case** analysis (mean confidence across 7 detectors $<$0.6). All three audits converge on the same story: Bilibili is the dominant platform in both volume and difficulty, re-encoding is the primary difficulty driver, and cross-label duplicates are effectively zero.

| Audit & metric | Value | Notes |
|---|---|---|
| *(A) Cleanlab confident-learning flags (ensemble of 10 detectors)* | | |
| Videos audited | 13,835 | Detector-scored pool (Tab. 13) |
| Videos flagged as potential label issues | 5,005 (36.2%) | Default cleanlab threshold; conservative, not all true mislabels |
| Top-100 "most-suspicious" breakdown | 93 real vs. 7 fake | Flags concentrate on the real side, matching the Kinetics-400 domain-shift story (Sec. 5.5) |
| by platform (top-100) | Bilibili 92, Kinetics 8 | No YouTube or Reddit appears in top-100 |
| by label source (top-100) | tier1_abs. 85, tier1_ds. 8, tier2_tag 7 | Tag-absence reals (Tier 2 in Sec. 4) dominate |
| *(B) pHash near-duplicate scan (Hamming $\leq$ 6)* | | |
| Total videos in M0 manifest | 21,504 | Full M0 release |
| Holdout sample hashed | 2,281 of 2,286 | 5 extraction misses (corrupted streams) |
| Multi-video clusters | 8 | Out of 2,263 clusters total |
| Videos in multi-clusters | 26 | 18 redundant after dedup; **0.79%** redundancy rate |
| Cross-platform dup clusters | 1 | Bilibili$\leftrightarrow$Reddit (user re-uploads) |
| Cross-*label* dup clusters | **0** | **No label-confusion duplicates detected** |
| Largest cluster | size 10 (showcase, all fake) | Identical upload of a Sora 2 gallery reel |
| *(C) Detector-ensemble hard-case analysis (mean confidence $<$ 0.6)* | | |
| Hard-fake rate overall | 4,825 / 7,973 = **60.5%** | Confirms Sec. 5.5 headline |
| YouTube | 239 / 299 = **79.9%** | Heaviest re-encoding $\rightarrow$ hardest |
| Bilibili | 3,931 / 5,913 = 66.5% | Dominant by volume |
| Reddit | 655 / 1,699 = 38.6% | Lightest re-encoding; easiest |
| Showcase | 0 / 62 = **0%** | Native gallery uploads are never hard |
| Hard-real rate overall | 3,249 / 5,862 = 55.4% | Calibration warning, not detection failure |
| Kinetics-400 | 221 / 234 = **94.4%** | Kinetics $\rightarrow$ platform-native domain shift |
| Bilibili reals | 3,028 / 5,628 = 53.8% | Reference distribution |

### A.10 LEGAL COMPLIANCE ANALYSIS

**EU GDPR (General Data Protection Regulation).** Under Art. 89, processing of publicly available data for scientific research purposes is permitted without explicit consent, provided appropriate safeguards (pseudonymization, data minimization) are in place. VIDTIDE distributes only video metadata (platform, label, generator, URL) without creator handles, usernames, or biometric data.

**China's Data Security Law (2021) and Personal Information Protection Law (PIPL).** We process only publicly disclosed platform-level labels (the `argue_info` field) that platforms are legally required to expose under the Deep Synthesis Provisions (Cyberspace Administration of China and Ministry of Industry and Information Technology and Ministry of Public Security, 2022) and the AI-Generated Synthetic Content Labeling Measures (Cyberspace Administration of China and others, 2025). No restricted or non-public data is accessed. Video content is accessed only through public URLs.

**US Copyright and Fair Use.** Distribution of metadata and download scripts (not video files) constitutes transformative academic use. The four-factor fair use analysis favors our use: (1) purpose is non-commercial research; (2) nature of works is factual metadata; (3) amount used is minimal (metadata only); (4) no market harm to original works (we do not redistribute content).

**Platform Terms of Service.** Bilibili's public API endpoints do not require authentication for search queries. YouTube data is accessed through public video metadata. Reddit's public JSON API is used per their API terms. We comply with all platforms' robots.txt directives.

## B LABEL QUALITY: AUDITS AND CONFOUND ANALYSES

### B.1 DATASET LABEL AUDITS: CLEANLAB, NEAR-DUPLICATE, AND HARD-CASE AGGREGATES

Section 4 argues that VIDTIDE's three-tier ground-truth is sound. Here we supply the raw automated-audit numbers that back up that claim, consolidated in Table 18. All three audits run over the 13,835-video detector-scored pool (Table 13); the near-duplicate audit additionally runs over a 2,286-video random holdout of the M0 crawl for pHash coverage.

Three conclusions triangulate the same underlying picture of VIDTIDE's label quality. *First*, the pHash audit provides a clean upper bound on dataset redundancy: 0.79% within-label and *zero* cross-label duplicates, which directly rules out the worst-case failure of a real-vs-fake benchmark (having the same clip in both classes). *Second*, cleanlab flags 36.2% of the pool as potential label issues, but the top-100 inspection reveals these are almost entirely (93%) *real*-labeled videos misread as "AI" by detectors that over-trigger on platform-native footage; the platform breakdown (Bilibili 92, Kinetics-400 8) further pinpoints Bilibili reals lacking the `argue_info` tag (85 / 100, tier1_platform_absence) as the dominant failure mode, i.e., a *detector* calibration issue surfacing as a cleanlab flag, not a *label* error. *Third*, the hard-case rate tracks the platform axis with near-perfect monotonicity (Showcase 0% $<$ Reddit 39% $<$ Bilibili 67% $<$ YouTube 80%), which is the cleanest experimental confirmation we have that re-encoding is the dominant driver of detector difficulty on the M0 slice. The hard-case analysis thus validates the heatmaps of Section 5.5 with a threshold-free signal derived from an independent ensemble.

### B.2 HUMAN ANNOTATION VALIDATION DETAILS

Three trained annotators (graduate students in computer vision with $>$1 year experience in video forensics) independently viewed 500 stratified-random videos from the M0 active eval pool (250 fake, 250 real; stratified by platform and GT tier). Each annotator received the following instructions: "Watch the full video. Judge whether this video was generated entirely or predominantly by an AI model (label: fake) or is a recording of real events, possibly with minor post-processing (label: real). Assign a confidence: high, medium, or low." No metadata (title, platform, tier) was shown—only the video itself.

Pairwise Cohen's $\kappa$ values: Annotator A–B: 0.93; A–C: 0.91; B–C: 0.92 (mean 0.92). Majority-vote vs. pipeline-assigned labels: agreement 482/500 (96.4%), $\kappa = 0.93$.

For the false-negative rate estimation, annotators additionally reviewed 200 random Bilibili videos that *lacked* the `argue_info` AI tag (i.e., presumptively real). Of these, 5 (2.5%) were judged by majority vote as "likely AI-generated or heavily AI-assisted," and 3 (1.5%) as "definitely AI-generated." This gives an estimated Tier 2 false-negative rate of 1.5–2.5%.

**Confidence intervals.** A Clopper–Pearson 95% binomial CI for 5/200 "likely AI" is $[0.8\%, 5.7\%]$ and for 3/200 "definitely AI" is $[0.3\%, 4.3\%]$. We plan to expand the audit to 500 stratified Bilibili-no-tag reals; at a 2.5% point estimate (12/500), the Clopper–Pearson interval would be about $[1.3\%, 4.2\%]$.

**Why the reverse-evidence label remains Tier 2 (not Tier 1).** Because the FN rate is non-zero, we rank the reverse-evidence Bilibili-no-tag reals at Tier 2 rather than treating them as equivalent to known-provenance labels. Section 4 bounds how much symmetric contamination is compatible with the observed temporal-slice AUROC (Eq. 2); Appendix C.1 reports how the headline gap responds to additional injected corruption.

### B.3 WATERMARK AND PROVENANCE-MARKER CONFOUND ANALYSIS

A natural concern with any in-the-wild AI-video benchmark is whether detectors are inadvertently learning watermark or provenance-marker fingerprints (visible vendor logos, burned-in "AI-generated" text overlays added by uploaders, container-level C2PA Content Credentials, or pixel-level neural watermarks such as SynthID-Video) rather than genuine generative artifacts. Two pieces of evidence bear on this confound.

**(i) Container-level provenance metadata is removed before evaluation.** C2PA-style credentials live in the media container / manifest layer rather than in the decoded frame tensor consumed by our detectors. Any such payload can be stripped by platform redistribution, and our own preprocessing then re-encodes every fake with `ffmpeg -c:v libx264 -crf 23 -maxrate 4M` (Section 3.1), dropping residual XMP / manifest metadata. Container-level credentials therefore do not reach the detector input.

**(ii) The Bilibili-only LP evaluation is a natural watermark-poor control.** Table 32 restricts the 1,999-video LP test set to its 1,721 Bilibili-sourced subset (same platform, same crawler, and overwhelmingly non-Veo / non-gallery content). The four backbones move within $\pm$0.8 pp of their full-test

AUROC (TSM 89.77 → 89.53, Swin 92.78 → 93.06, I3D 86.38 → 86.93, SlowFast 89.01 → 89.76). If the headline 90.5% LP average were watermark-driven, restricting to this watermark-poor population would collapse it; it does not.

A complementary positive control is the bench-trained-LP-on-GenVideo-Val transfer (Appendix E.5, 67.8–77.8% pooled AUROC): GenVideo's 2023–early-2024 generators (Crafter, Lavie, Gen2, ModelScope, etc.) predate modern video-provenance conventions, yet our heads still detect them, indicating that the learned features generalize to a watermark-poor population. We therefore treat visible provenance markers as one component of the in-the-wild distribution rather than as the driver of the benchmark outcome, on the same methodological footing as the platform re-encoding heterogeneity analyzed in Table 27.

# C Controls for the Alternative Explanations

## C.1 Label Noise: Derivations, Bounds, and Sensitivity

Section 4 distinguishes empirical audits from bounds under an explicit contamination model. This appendix gives the derivations, numerical checks, and sensitivity analysis.

**Symmetric contamination.** Suppose each observed class contains an $\varepsilon$ fraction of the other: the observed-fake pool is $(1-\varepsilon)$ true fakes and $\varepsilon$ true reals, and symmetrically for the observed-real pool. For a scorer that ranks true fakes above true reals, drawing one clip from each observed pool gives a correctly ordered pair with probability $(1-\varepsilon)^2$, an inverted pair with probability $\varepsilon^2$, and a tie-equivalent (same true class) with probability $2\varepsilon(1-\varepsilon)$, which contributes $\frac{1}{2}$ each. Hence $\mathrm{AUROC} = (1-\varepsilon)^2 + \varepsilon(1-\varepsilon) = 1-\varepsilon$.

**One-sided contamination.** If mislabelling only runs one way—undisclosed AI content sitting in the real pool—and $\alpha$ is the ratio of mislabelled to correct clips in the affected class, the ceiling is $(1+\alpha/2)/(1+\alpha)$.

Both forms assume score-independent contamination within each true class and that the observed class orientation is informative. The blind human audit provides a separate check on labels; it does not establish these modeling assumptions.

Table 19: **Noise ceilings implied by achieved accuracy.** Monte-Carlo simulation (20,000 clips, 40 repetitions) matches the closed forms to ≤0.0011 AUROC (symmetric) and ≤0.0004 (one-sided).

| Observed result | AUROC | symmetric $\Rightarrow \varepsilon \leq$ | one-sided $\Rightarrow \alpha \leq$ |
|---|---|---|---|
| post-M0 slice (crawled after M0 froze) | 94.8% | **5.2%** | **11.6%** |
| Best-of-LP/FT mean, M0-Test | 91.7% | 8.3% | 19.9% |

The post-M0 slice was crawled after the M0 split closed and deduplicated against it, so it is temporally held out with no clip-level overlap. It may still share platform- and content-level structure with M0, so we describe it as temporally held out rather than fully independent.

**Sensitivity to further corruption.** We inject *additional* symmetric corruption into the VidTide test labels and recompute the headline gap (Table 20). The corruption is synthetic and applied only to VidTide, so this measures robustness of the conclusion, not how much of the observed gap existing noise explains. Each row is 200 resamples of the corruption draw across 11 checkpoints.

Table 20: **Effect of injected label corruption on the headline gap.** The measured gap widens under this additional random corruption. This intervention does not exclude structured errors already present in the labels.

| Added corruption | Resulting gap | Increase vs. $\varepsilon = 0$ |
|---|---|---|
| 0 | −30.98 pp (reproduces the 31.0 headline) | — |
| 2.5% | −31.66 pp | **0.68 pp** |
| 5% | −32.31 pp | 1.33 pp |
| 20% | −36.37 pp | 5.39 pp |

### C.2 CONTROLLED BITRATE LADDER: FULL RESULTS

**Protocol.** 800 label-balanced clips from M0-TEST; each cut to an identical 16 s window starting 10% into the clip; each window re-encoded with libx264 at eight fixed target bitrates; `orig` is a stream copy of the same window. 6,400 re-encodes and 32,000 feature extractions in total. Detector heads are the VIDTIDE-trained linear probes of Table 6, unchanged across rungs. Intervals are paired bootstraps over clips (5,000 resamples, fixed seed) on the paired difference between rungs.

Table 21: **Per-backbone AUROC (%) across the bitrate ladder.** Rows are ordered by realised bitrate. The 48× span from b20000k to b250k costs 1.73 pp on the mean (CI [0.84, 2.66]); across the platform-relevant range (449–6,207 kbps) the spread is 0.71 pp and no backbone exceeds 1.08 pp.

| Rung | Realised | Swin-T | TSM | I3D | SlowFast | Mean |
|---|---|---|---|---|---|---|
| b20000k | 10,892 | 89.34 | 86.68 | 82.74 | 83.22 | **85.50** |
| b8000k | 6,207 | 89.36 | 86.68 | 82.73 | 83.24 | 85.50 |
| b4000k | 3,351 | 89.38 | 86.60 | 82.76 | 83.19 | 85.48 |
| b2000k | 1,728 | 89.37 | 86.61 | 82.91 | 83.25 | 85.54 |
| orig | 808 | 89.15 | 86.91 | 82.62 | 83.48 | 85.54 |
| b1000k | 883 | 89.07 | 86.35 | 82.75 | 83.31 | 85.37 |
| b500k | 449 | 88.63 | 85.60 | 82.36 | 82.76 | 84.83 |
| b250k | 228 | 87.82 | 82.80 | 82.60 | 81.85 | **83.77** |
| **Span (48× cut)** | | 1.52 | **4.11** | 0.26 | 1.63 | **1.73** |

**The semantic family is untouched too.** CLIP ViT-L/14 zero-shot over the identical eight rungs spans **60.40–60.95%**—a 0.55 pp range—with the most compressed rung (228 kbps) scoring highest of all. Compression moves neither the supervised artifact detectors nor the semantic baseline.

**Observational cross-check.** Binning the full 5K/5K pool by bitrate and computing AUROC within bins, the ten GenVideo-trained detectors are flat at $+0.09$ pp per bin. We report the ladder rather than this binning as the primary evidence, because only the ladder holds content fixed.

**Scope.** The fixed 16 s window places absolute values 2.9–5.5 pp below Table 6 (M0 has a 91 s median duration and the detectors consume 32 frames). The window is identical across rungs, so between-condition comparisons are unaffected; only the absolute level shifts.

### C.3 CONFOUND CONDITIONING: SCHEMES AND RESIDUAL SIGNAL

**Schemes.** All three operate on identical rows of M0-EVAL (1,999-clip test split scored by five backbones): (i) *stratified* by duration × bitrate deciles; (ii) *+platform*, adding source platform to the strata; (iii) *matched*, caliper matching each fake to a real on duration, bitrate, and platform, yielding 486 pairs. Each scheme re-scores the same cached predictions, so differences across columns are due to the conditioning alone. Intervals are paired bootstraps over clips (5,000 resamples).

**How strong the unconditioned confound is.** A gradient-boosted model given only metadata (platform, duration, file size, resolution, fps, bitrate) and no pixels reaches **83.85%** AUROC on M0-EVAL. In-the-wild AI videos genuinely differ in length and encoding; that is a property of the data, and the point of the controls is to separate it from the visual signal. Under matching it falls to 55.03% (CI [51.47, 58.58]) while the visual mean holds 88.29% (CI [86.61, 89.89]); the paired difference is 33.26 pp (CI [29.51, 37.02]).

**Residual metadata inside a single ladder rung.** The ladder provides a control in which the confounds are absent by construction rather than statistically adjusted: within one rung every clip has the same duration (16 s) and the same encoder settings, so neither can carry label information. The only metadata left is file size, which at fixed bitrate mostly tracks scene complexity. Its single-variable AUROC across the eight rungs is **45.6–52.9%**, a mean absolute deviation from chance of **3.04 pp**, while the visual probes on those same clips stay near 89%.

### C.4 ZERO-SHOT MLLM BASELINES

**Protocol.** A 1,000-clip stratified subset of M0-EVAL (500 fake / 500 real), the paper's own MLLM prompt, 8 frames per clip, greedy decoding, bf16 on a single GPU. No parse failures occurred in

any of the 3,000 evaluations. *Every* existing detector was re-scored on exactly these clips, so no comparison in Table 7 mixes pools. Intervals are paired bootstraps over clips (5,000 resamples, fixed seed); because all arms score identical clips, paired differences can be tested directly.

**Two properties worth recording.** The 2025-01 model emits only **8 distinct probability values** across 1,000 clips, which caps achievable AUROC—reading token logits instead of the verbalised probability would likely sharpen it. Its mean predicted probability is 0.811 (fakes 0.892, reals 0.731): the direction is right, but the model leans toward calling everything AI.

**On sample size.** Two numbers are comparable when they are computed on the same clips; equal sample sizes are not what makes them comparable. All existing detectors were re-scored on the identical 1,000-clip subset, so every entry in Table 7 comes from the same videos. The 5K/5K figures in Table 5 are separate measurements and are never placed alongside these. A smaller subset does widen the intervals, which is why we report them; even at $n = 1{,}000$ the generational differences stay well clear of zero.

**Coverage.** Of other recent methods, DeCoF and AIGVDBench do not currently release weights. The evaluated lineup covers 11 GenVideo-trained checkpoints across 5 detector families (Table 24), CLIP zero-shot, three MLLM generations, and 5 K400-pretrained backbones retrained on VIDTIDE, spanning 2019–2026.

### C.5 LINK DECAY AUDIT: FAILURE MODES AND REGIONAL ACCESS

**Protocol.** M0 froze in April 2026. In July 2026 we re-resolved a stratified sample of 1,200 manifest URLs, counting every outcome that is not a successful fetch as a failure—including 11 Reddit CDN links that return HTTP 403. Survival intervals are Wilson. Results are in Table 22.

Table 22: **Link survival three months after the M0 freeze** (stratified sample of 1,200 manifest URLs; any non-fetch counts as a failure). Kinetics-400 is reported separately rather than folded into the headline because it is the reference point—an existing URL-distributed community dataset—not VIDTIDE-crawled material.

| Source | Sampled | Live | Rate | Wilson 95% CI |
|---|---|---|---|---|
| Bilibili | 500 | 483 | 96.60% | [94.62, 97.87] |
| Reddit | 300 | 289 | 96.33% | [93.55, 97.94] |
| YouTube | 250 | 238 | 95.20% | [91.80, 97.23] |
| Showcase | 50 | 50 | 100.00% | [92.86, 100.00] |
| **VIDTIDE-crawled subtotal** | **1,100** | **1,060** | **96.36%** | **[95.09, 97.32]** |
| Kinetics-400 (borrowed real pool) | 100 | 88 | 88.00% | [80.19, 93.00] |

**Failure modes.** Failures separate into *recoverable* (Bilibili uploader-only 8, withdrawn 5) and *permanent* (Bilibili deleted 4, YouTube deleted 10, YouTube restricted 2, Reddit CDN 403 $\times$11). The manifest records which of the two a row hit, so a user can distinguish a slice that shrank from a slice that is temporarily unreachable.

Table 23: **Regional access.** Re-resolution of 20 URLs per source from two networks. Reproducing a full slice requires a network that reaches all sources.

| Source | China network | Unrestricted |
|---|---|---|
| Bilibili | 20/20 | reachable |
| YouTube | **0/20** | reachable |
| Reddit (incl. CDN) | **0/20** | reachable |

**Reference point.** Kinetics-400 (Kay et al., 2017) is reported separately from the headline rather than folded into it: it is an existing URL-distributed community-standard dataset that we borrow rows from, so it calibrates what link decay looks like for this class of release. Over the same three-month window VIDTIDE-crawled material failed at 3.64% against its 12.00%—under a third the rate. Link decay is a standing condition of URL-distributed video datasets rather than something introduced here.

Table 24: **Model zoo. Group A** models supply GenVideo-trained checkpoints evaluated zero-shot on VIDTIDE (Table 5); **Group B** models supply K400-pretrained backbones retrained on VIDTIDE (Table 6).

| # | Model | Family | Used in | Source |
|---|---|---|---|---|
| *Group A: Established AI-video detectors (zero-shot transfer to* VIDTIDE*)* | | | | |
| A1 | DeMamba (Chen et al., 2024) | Mamba | Tab. 5 | Official GitHub (Pika / Crafter ckpts) |
| A2 | NSG-VD (Zhang et al., 2025) | Diffusion-prior + MMD | Tab. 5 | Official GitHub (Pika ckpt only) |
| A3 | STIL (Gu et al., 2021) | Spatio-temporal CNN | Tab. 5 | Official GitHub (Pika / Crafter ckpts) |
| A4 | NPR (Tan et al., 2024) | Frequency-domain CNN | Tab. 5 | Official GitHub (Pika / Crafter ckpts) |
| A5 | TALL (Xu et al., 2023) | Thumbnail-layout Swin | Tab. 5 | Official GitHub (Pika / Crafter ckpts) |
| A6 | CLIP ViT-L/14 | Zero-shot VLM | Tab. 5 | OpenCLIP (no GenVideo training) |
| *Group B: K400-pretrained backbones (LP / FT on* VIDTIDE *train split)* | | | | |
| B1 | DeMamba (Chen et al., 2024) | Mamba | Tab. 6 | Official GitHub (K400 init) |
| B2 | VideoSwin-Tiny (Liu et al., 2022) | Transformer | Tab. 6 | MMAction2 (K400 init) |
| B3 | TSM (Lin et al., 2019) | CNN | Tab. 6 | MMAction2 (K400 init) |
| B4 | I3D | CNN | Tab. 6 | MMAction2 (K400 init) |
| B5 | SlowFast | CNN | Tab. 6 | MMAction2 (K400 init) |

Table 25: **Per-checkpoint bACC / F1 on VIDTIDE M0 5K slice.** bACC = balanced accuracy, F1 = standard F1 score, both at decision threshold 0.5. Same Pika and Crafter checkpoints as Table 5; "—" = no public Crafter checkpoint released by (Zhang et al., 2025) for NSG-VD.

| | Pika | | Crafter | |
|---|---|---|---|---|
| **Model** | bACC | F1 | bACC | F1 |
| DeMamba (Chen et al., 2024) | 55.8 | 23.4 | 54.1 | 20.3 |
| NSG-VD (Zhang et al., 2025) | 52.3 | 39.9 | 49.3‡ | 41.8‡ |
| STIL (Gu et al., 2021) | 50.4 | 2.4 | 51.6 | 10.3 |
| NPR (Tan et al., 2024) | **65.0** | **59.5** | 48.4 | **47.1** |
| TALL (Xu et al., 2023) | 52.4 | 12.3 | 51.4 | 12.2 |
| CLIP ViT-L/14 (zero-shot) | bACC 53.8 | F1 65.0 | (pooled, no Pika/Crafter distinction) | |
| **Avg (5 supervised)** | 55.2 | 27.5 | 51.0 | 26.3 |

‡NSG-VD Crafter checkpoint trained by us (Appendix D.2); all other cells use the publicly-released checkpoints.

**Versioning rules.** Each slice carries a version; each row carries a status. A row that stops resolving is marked `retired` instead of removed, so the manifest only grows. An evaluation names the manifest version it used; two results quoting the same version were computed on the same rows. A comparison spanning versions intersects them and reports the shared subset, so a shrinking pool widens intervals instead of quietly changing which videos are compared. The leaderboard scores all submissions against a single as-of snapshot for the same reason. Re-downloads are verified two ways and reported separately: SHA-256 for byte-identical files, frame fingerprints for content the platform has re-encoded since the freeze.

### C.6 FINE-TUNING: FULL PER-BACKBONE TABLE

The full adaptation results are reported in the main text, Table 6.

## D MODELS, CHECKPOINTS, AND TRAINING DETAILS

### D.1 MODEL ZOO

### D.2 TRAINING THE NSG-VD CRAFTER CHECKPOINT

The official NSG-VD release (Zhang et al., 2025) ships two MMD-discriminator checkpoints (Pika and SEINE) but not a Crafter one, leaving the Crafter column of Table 5 and Table 25 empty in the paper's first draft. To fill this cell we trained a Crafter-specific checkpoint ourselves, reusing the

Table 26: **Fine-tuning hyperparameters.** All models are fine-tuned on VIDTIDE M0 train split using identical protocols to ensure fair comparison.

| Hyperparameter | Linear Probe | Full Fine-Tune |
|---|---|---|
| Frozen backbone | Yes | No |
| Learning rate | 1e-3 | 1e-4 |
| LR scheduler | Cosine decay | Cosine decay |
| Epochs | 10 | 20 |
| Batch size | 32 | 16 |
| Optimizer | AdamW | AdamW |
| Weight decay | 1e-4 | 1e-4 |
| Warmup epochs | 1 | 2 |
| Input frames | 8 (uniform) | 8 (uniform) |
| Input resolution | 224 × 224 | 224 × 224 |
| Data augmentation | Random crop, horizontal flip | Random crop, horizontal flip |

authors' released training code (`deep_MMD` + `SingleSwinBlockDiscriminator`) without modification.

**Training data.** GenVideo's official split provides ∼1,259 Crafter val/test videos but *no* Crafter train split; we synthesized one by reusing the test split as training data (which is acceptable here because our Crafter test videos are disjoint from any VIDTIDE evaluation pool). As the real class we used MSR-VTT val+test (10K videos) since Kinetics-400 is not redistributable. All videos were resampled to 8 frames at $224\times224$ and converted to velocity features via the public $256\times256$ ImageNet diffusion prior ($T=5$ reverse steps).

**Training setup.** A100-40GB, 30 epochs, MMD-MP mode (`is_yy_zero=True`), single Swin-B discriminator block ($d=300$), Adam $lr=10^{-4}$, batch size 32 paired (16 fake + 16 real). Training converges in ∼10 minutes. Best held-out AUROC on the Crafter val split (vs. MSR-VTT reals) is **0.9284**, which is ∼3.3 pp below the authors' reported in-domain AUROC on Kinetics-400; the gap is consistent with the MSR-VTT-vs-K400 substitution (MSR-VTT has more compression artifacts, making real-vs-fake separation slightly easier on the real side and harder on the fake side).

**Evaluation on VIDTIDE.** We evaluated the resulting 1.7 MB checkpoint on the same 5K-real / 5K-fake VIDTIDE M0 slice used for every other row of Table 5, using the authors' published inference protocol unchanged (100 Bilibili reals as the MMD reference set, 8 frames at $224\times224$, $T=5$ diffusion-velocity reverse steps). The resulting AUROC places NSG-VD Crafter in the same "at-or-below chance" bucket as NPR Crafter (49.0), confirming that the Crafter-checkpoint gap is a *generic* cross-distribution-shift phenomenon, not an artifact of any single detector family.

### D.3 FINE-TUNING HYPERPARAMETERS

## E EXTENDED RESULTS

### E.1 PER-CHECKPOINT BACC AND F1 (DEFERRED FROM TABLE 5)

The headline gap table (Table 5) reports AUROC only, since ΔAUROC is the cleanest checkpoint-comparable measure of detector degradation. For completeness, Table 25 provides the per-checkpoint balanced accuracy (bACC) and F1 score on VIDTIDE M0. All numbers use the default decision threshold of 0.5 on the model's fake-probability output; we did *not* re-tune the threshold on VIDTIDE data (which would inflate F1 numbers misleadingly).

### E.2 PER-GENERATOR DETAILED RESULTS

### E.3 PER-SLICE LINEAR-PROBE DECOMPOSITION

Table 28 decomposes the LP results from Table 6 into 12 evaluation slices covering platforms, generators, and a cross-tab control.

Table 27: **Detector × generator AUROC on M0-EVAL.** Color encodes AUROC (deep red = near-chance, deep green = near-perfect). Within a single detector, AUROC varies by up to 30.2 pp across generators (DeMamba-Pika: Veo 3 0.952 vs. Dreamina 3 0.650).

| | Kling 2.1 | Veo 3 | Sora 2 | Dreamina 3 | Runway Gen-4 |
|---|---|---|---|---|---|
| CLIP (zero-shot) | 0.63 | 0.74 | 0.53 | 0.58 | 0.75 |
| DeMamba (Crafter) | 0.73 | 0.77 | 0.63 | 0.64 | 0.75 |
| DeMamba (Pika) | 0.83 | 0.95 | 0.71 | 0.65 | 0.92 |
| NPR (Crafter) | 0.55 | 0.61 | 0.54 | 0.52 | 0.54 |
| NPR (Pika) | 0.76 | 0.75 | 0.68 | 0.71 | 0.75 |
| NSG-VD (Crafter) | 0.46 | 0.59 | 0.45 | 0.49 | 0.43 |
| NSG-VD (Pika) | 0.51 | 0.55 | 0.53 | 0.55 | 0.53 |
| STIL (Crafter) | 0.69 | 0.69 | 0.61 | 0.65 | 0.59 |
| STIL (Pika) | 0.70 | 0.74 | 0.63 | 0.56 | 0.68 |
| TALL (Crafter) | 0.57 | 0.70 | 0.52 | 0.46 | 0.65 |
| TALL (Pika) | 0.72 | 0.84 | 0.62 | 0.57 | 0.83 |

Table 28: **Per-slice LP decomposition across backbones.** Each cell is "AUROC / TPR@5" (%); TPR@5 shares a slice-independent 5%-FPR threshold set on the full 1,000-real test pool. Slices: overall, source platforms, named generators (≥ 10 fakes), platform-controlled cross-tab, and a fresh-generator OOD slice. Best AUROC per row bolded.

| Slice | DeMamba | Swin-T | TSM | I3D | SlowFast | Avg |
|---|---|---|---|---|---|---|
| *Overall* | | | | | | |
| All test videos | **94.7 / 79** | 92.8 / 71 | 89.8 / 56 | 86.4 / 46 | 89.0 / 58 | **90.5 / 62** |
| *By source platform* | | | | | | |
| Bilibili | **94.0 / 77** | 93.2 / 71 | 89.7 / 55 | 86.8 / 45 | 89.7 / 58 | 90.7 / 61 |
| Reddit | **97.3 / 87** | 94.3 / 75 | 91.6 / 60 | 88.2 / 50 | 90.2 / 61 | 92.3 / 67 |
| YouTube | **96.2 / 78** | 89.3 / 50 | 87.9 / 47 | 80.8 / 41 | 80.8 / 31 | 87.0 / 49 |
| *By claimed generator* | | | | | | |
| Unlabeled (Bili) | **94.5 / 79** | 92.6 / 70 | 89.5 / 55 | 86.2 / 45 | 88.7 / 57 | 90.3 / 61 |
| Kling 2.1 | **95.4 / 83** | 94.0 / 77 | 93.4 / 57 | 91.6 / 55 | 93.3 / 76 | 93.5 / 70 |
| Veo 3 | **98.5 / 94** | 98.2 / 89 | 95.2 / 69 | 88.8 / 53 | 95.7 / 72 | 95.3 / 75 |
| Sora 2 | **93.9 / 68** | 91.6 / 61 | 85.4 / 50 | 86.0 / 36 | 87.6 / 46 | 88.9 / 52 |
| Dreamina 3 | **96.6 / 86** | 95.4 / 77 | 93.8 / 82 | 88.8 / 64 | 91.4 / 73 | 93.2 / 76 |
| Runway Gen-4 | **93.5 / 71** | 85.6 / 50 | 84.0 / 29 | 77.8 / 14 | 75.4 / 57 | 83.3 / 44 |
| Other (10 gens) | **96.0 / 85** | 94.2 / 79 | 93.5 / 67 | 89.7 / 58 | 89.5 / 58 | 92.6 / 69 |
| *Platform-controlled cross-tab* | | | | | | |
| Bilibili ∩ labeled | **94.6 / 77** | 93.5 / 76 | 91.4 / 63 | 89.0 / 48 | 91.0 / 64 | 91.9 / 66 |
| *Held-out OOD (50 fakes, 14 generators, 49 reals)* | | | | | | |
| OOD fresh* | N/A | **96.0** / – | 94.0 / – | 94.8 / – | 94.3 / – | **94.8** / – |

* Zero-shot on 50 fakes from 14 generators crawled after the train cut; scores are per-backbone logits. Full per-generator breakdown: Table 29.

### E.4 OUT-OF-DISTRIBUTION GENERALIZATION TO FRESH GENERATORS (M0 → POST-M0 MINI-M1)

This subsection serves as a temporal-generalization stress test of the LP heads, addressing the M0 → M1 transfer question that the full M1 slice will close in the camera-ready release. The held-out pool here is a *post-M0 mini-M1*: 50 fakes from 14 distinct generators and 49 Bilibili real clips, all crawled in the four-week window *strictly after* the M0 train/test split closure (i.e., these videos were uploaded after the temporal cutoff used to build the M0 training set, and so cannot have leaked into training). Table 29 audits whether the LP gains generalize beyond the 5 named generators dominating

Table 29: **Fresh-generator OOD generalization per backbone.** Fifty fake clips from 14 distinct generators and 49 Bilibili real clips, all crawled in the four-week window *after* the M0 train/test split closure. Per-row AUROC (%) uses each generator's fakes vs. the full 49-real pool. Best per row bolded.

| Generator | Swin-T | TSM | I3D | SlowFast | Avg. |
|---|---|---|---|---|---|
| Unknown-AI (tag) | **97.1** | 94.8 | 93.9 | 95.3 | 95.3 |
| Sora 2 | **98.0** | **98.0** | 97.1 | 97.6 | 97.7 |
| Kling 2.1 | **97.6** | 96.7 | 94.3 | 95.9 | 96.1 |
| Veo 3 | 85.2 | **91.3** | 83.2 | 87.2 | 86.7 |
| Dreamina 3 | **100.0** | 99.0 | 96.9 | 89.3 | 96.3 |
| Runway Gen-4 | **100.0** | 98.0 | **100.0** | 97.3 | 98.8 |
| Pika 2 | 87.8 | **95.2** | 87.1 | **95.2** | 91.3 |
| Hailuo | **100.0** | 97.3 | 98.6 | 97.3 | 98.3 |
| Vidu | **91.2** | 70.1 | **91.2** | 88.4 | 85.2 |
| Kling 2 | **100.0** | **100.0** | **100.0** | 98.6 | 99.7 |
| Luma | **94.6** | 81.0 | 91.8 | 89.8 | 89.3 |
| Wan 2.1 | 94.6 | **95.9** | 99.3 | 96.6 | 96.6 |
| **Overall (50/49)** | **96.0** | 94.0 | 94.8 | 94.3 | **94.8** |

the M0 training pool, and equivalently, whether the LP heads survive the temporal shift from M0 to a post-M0 slice.

### E.5 Cross-Dataset Evaluation: VidTide-trained LP Heads on GenVideo

To answer the symmetric question to Table 5 (*do* VidTide*-trained models retain signal when transported back to GenVideo's lab distribution?*), we transfer each VidTide-trained K400-pretrained LP head zero-shot onto GenVideo-Val. This is the cross-dataset complement to the OOD-fresh M0 $\rightarrow$ post-M0 temporal test (§E.4).

**Setup.** We use the four LP heads from Table 6 (Swin-T, TSM, I3D, SlowFast) without re-training. GenVideo-Val is the public release accompanying (Chen et al., 2024): 11 named fake generators (Crafter, Lavie, Gen2, HotShot-XL, ModelScope, MorphStudio, Show-1, MoonValley, WildScrape, Sora v1) plus MSR-VTT as the canonical real pool. We sample up to 200 videos per generator and 200 MSR-VTT reals (with the cap matching feature-extraction budget on a single A100 container). For each LP head we extract pre-classifier features in the head's native input format (32 frames at 224$\times$224 for Swin / I3D / SlowFast, 8 frames at 224$\times$224 for TSM) and apply the cached weight + bias to obtain a per-clip fake-probability. AUROC is computed per generator (each generator's fakes vs. the full MSR-VTT real pool) and pooled across all generators.

**Result.** Table 30 reports the per-(backbone, generator) AUROC and the pooled cross-dataset AUROC. VidTide-trained LP heads achieve a **4-backbone-mean pooled AUROC of 71.7%** on GenVideo-Val (range 67.8%–77.8%), and a per-generator-mean ranging from 61.6% (HotShot-XL) to 85.9% (MoonValley). The reverse-transfer results show useful but incomplete generalization. The four tested LP heads average 89.5% on M0-Test and 71.7% on GenVideo-Val. The forward comparison in Table 5 uses a different detector group and reported source results, so comparing the two gap sizes does not isolate training-distribution coverage.

**Interpretation.** Three structural points frame how to read this table.

*(a) Transfer is incomplete in both directions.* The paired LP results are 92.8 to 77.8 for Swin, 89.8 to 72.2 for TSM, 86.4 to 67.8 for I3D, and 89.0 to 69.2 for SlowFast (all AUROC percentages). These are drops of 15.0–19.8 percentage points. They describe these four probes; the historical forward comparison uses other models and protocols and does not support a controlled ranking of the two training distributions.

*(b) The two benchmarks share nothing that a dataset-specific shortcut could exploit* (Table 31). VidTide is centered on 1080p, 24–30 fps, 10–15 s clips from current-frontier generators (Sora 2, Veo 3, Kling 2.1, Dreamina 3, etc., released 2024Q4–2026Q1). GenVideo-Val is centered on 256–512 px, 8–16 fps, 3-second clips from 2023–early-2024 generators (ModelScope, Lavie, Gen2, Crafter, etc.) that predate the modern T2V wave. The fact that VidTide's LP heads transfer to a

Table 30: **Zero-shot cross-dataset evaluation: VIDTIDE-trained LP heads on GenVideo-Val.** Per-cell AUROC (%) of VIDTIDE-trained linear-probe heads (no retraining) applied to GenVideo-Val with MSR-VTT as the real pool. "Pooled" is all-fakes-vs-all-reals AUROC across the 9 named GenVideo fake generators per backbone. Sora v1 is computed on its 56-clip native split (GenVideo-Val ships fewer than 200); all other rows use 200 clips per generator and 200 MSR-VTT reals. WildScrape is shown as a separate row (also AUROC 73.9% mean): GenVideo's documentation positions it as a wild-scraped real-domain class, but our heads consistently score it above MSR-VTT, suggesting the slice is closer to fake than real in the in-the-wild sense—we leave it out of the pooled AUROC for clean comparability.

| Generator | Swin-T | TSM | I3D | SlowFast | Avg. |
|---|---|---|---|---|---|
| Crafter | 83.0 | 76.7 | 73.0 | 71.9 | 76.1 |
| Lavie | 67.7 | 62.2 | 60.8 | 60.0 | 62.7 |
| Gen2 | 89.5 | 83.5 | 81.4 | 77.6 | 83.0 |
| HotShot-XL | 66.8 | 63.2 | 55.5 | 60.7 | 61.6 |
| ModelScope | 76.0 | 65.8 | 62.6 | 69.8 | 68.6 |
| MorphStudio | 74.9 | 68.5 | 62.8 | 61.1 | 66.8 |
| Show-1 | 74.5 | 68.6 | 62.2 | 56.0 | 65.3 |
| MoonValley | 88.3 | 85.4 | 81.5 | 88.5 | 85.9 |
| Sora v1 | 81.7 | 83.0 | 69.4 | 78.5 | 78.2 |
| **Pooled (all 9 fakes vs. MSR-VTT)** | **77.8** | **72.2** | **67.8** | **69.2** | **71.7** |
| WildScrape (separate, see caption) | 78.0 | 73.2 | 69.6 | 74.8 | 73.9 |

Table 31: **Differences between the training and transfer datasets.** The LP heads score 71.7% mean AUROC on GenVideo-Val. Transfer across these differences does not by itself identify which visual cues the models use.

| | **VIDTIDE (training)** | **GenVideo-Val (test)** |
|---|---|---|
| Origin | public posts on live platforms | generated in a lab |
| Generator vintage | Sora 2, Veo 3, Kling 2.1, Dreamina 3 (2024Q4–2026Q1) | ModelScope, Lavie, Gen2, Crafter, Show-1 (2023–early 2024) |
| Res. / fps / length | 1080p, 24–30 fps, 91 s median | 256–512 px, 8–16 fps, ∼3 s |
| Platform processing | re-encoded, transcoded, redistributed | none |
| Real pool | platform-native footage | MSR-VTT |
| Collection | platform disclosure tags, LLM screening, dedup | controlled generation |

totally different vintage of generators *at all* (mean per-generator AUROC range 62–86%) suggests that they have learned distribution-invariant artifacts shared across generator generations, rather than generator-specific fingerprints—consistent with the conclusion of Section 5.2 that *distributional coverage matters more than architecture choice for standard video backbones*.

*(c) Per-generator hardness ranking on GenVideo-Val is informative on its own.* The hardest GenVideo generators (HotShot-XL 61.6%, Lavie 62.7%, Show-1 65.3%) all share short-form animation aesthetics that lack the documentary-realism cues present in most M0 training fakes; the easiest (MoonValley 85.9%, Gen2 83.0%, Sora v1 78.2%) are higher-fidelity / more cinematic, exactly the regime our M0 training distribution emphasizes. The rank-ordering inversion of Sora v1 against Sora 2 (Sora v1 here at 78.2% vs. Sora 2 in M0 at 0.609 mean = 60.9% AUROC, Section 5.5) is consistent with the rapid frontier shift: Sora v1's outputs are already easier to detect than Sora 2's, even from the perspective of a VIDTIDE-trained detector.

Table 32: **LP AUROC decomposition: full test vs. Bilibili-only vs. cross-platform.** "Full" is the 1,999-video test set from Table 6. "Bili-only" restricts to the 1,721 Bilibili-sourced videos (960 real / 761 fake, same platform, no platform→label shortcut). "Cross-platform" restricts to the 278 non-Bilibili videos (40 Kinetics reals + 197 Reddit fakes + 34 YouTube fakes + 7 showcase fakes). All numbers are AUROC (%). $\Delta$(bili−full) is within $\pm 0.8$ pp for every backbone, confirming that the LP numbers are not driven by platform identity.

| Backbone | Full (n=1,999) | Bili-only (n=1,721) | Cross-platform (n=278) | $\Delta$ (bili−full) |
|---|---|---|---|---|
| TSM | 89.77 | 89.53 | 86.79 | $-0.24$ |
| VideoSwin-T | 92.78 | 93.06 | 81.52 | $+0.28$ |
| I3D | 86.38 | 86.93 | 69.24 | $+0.55$ |
| SlowFast | 89.01 | 89.76 | 72.29 | $+0.75$ |

Table 33: **Data scaling ablation.** Linear-probe test AUROC (%) on the VIDTIDE M0 test set (1,999 videos) as a function of training-set fraction (stratified by label × generator) for three K400-pretrained backbones. The 100% row matches the corresponding LP entries in Table 6.

| Model | 10% (∼800 vids) | 25% (∼2K) | 50% (∼4K) | 100% (∼8K) |
|---|---|---|---|---|
| DeMamba (Chen et al., 2024) | 90.2 | 92.9 | 93.5 | **94.8** |
| TSM (Lin et al., 2019) | 83.8 | 86.1 | 88.3 | **89.9** |
| VideoSwin-Tiny (Liu et al., 2022) | 87.6 | 89.6 | 91.5 | **92.8** |

### E.6 WITHIN-BILIBILI ROBUSTNESS OF LINEAR-PROBE RESULTS

A natural concern with the 90.5% LP average reported in Table 6 is whether it is partially driven by a *platform→label* shortcut: in the M0 active-eval pool, Kinetics-400 real videos are 100% real and showcase / non-Bilibili fake videos are 100% fake, so a detector that simply learns platform identity would inherit free AUROC. We control for this by re-evaluating each cached LP head on the Bilibili-only subset of the test split (1,721 videos; 960 real + 761 fake from the same platform and thus with no platform→label signal). Results are in Table 32.

The Bili-only AUROC is within $\pm 0.8$ pp of the full-test AUROC for all four backbones, so the headline LP numbers in Table 6 are *not* attributable to the platform→label shortcut. The cross-platform column is 3–17 pp below the full test set. This subset contains 40 Kinetics-400 reals and 238 Reddit, YouTube, and showcase fakes.

### E.7 DATA SCALING ABLATION

To understand how much VIDTIDE data is needed, we vary the training set fraction and measure AUROC on the M0 test set.

All three backbones improve monotonically with more VIDTIDE training data; gains are steepest in the 10→50% range (+3.3 pp DeMamba, +4.5 pp TSM, +3.9 pp VideoSwin) and partially saturate beyond (+1.3 / +1.6 / +1.3 pp from 50% to 100%). Notably, even at the 10% slice (≈800 videos), DeMamba already reaches 90.2 % AUROC—comfortably exceeding any GenVideo-trained baseline in Table 5. This implies that the dominant cost of in-the-wild detection is *distributional coverage* rather than raw scale: the first ∼800 platform-curated videos buy more accuracy than any algorithmic change tested in the same regime.

### E.8 HARD CASE ANALYSIS

We identify *hard cases* as videos where the mean detection confidence across all 7 models in our zoo falls below 0.6 (near chance). Table 9 in the main text reports the per-generator breakdown over the 7,973 fake videos in the M0 active eval pool; per-platform hardness numbers are discussed in the main text (Section 5.5).

The hardness ranking is consistent with subjective assessment of the corresponding generator galleries: Dreamina 3, Sora 2, and Pika 2 produce the most photorealistic clips and dominate the hard pool,

while the more stylized Veo 3 and Kling 2.1 leave more obvious artifacts. The single largest bucket—"other / unidentified AI" (78% of all M0 fakes)—reflects the long tail of platform-disclosed AI content that lacks an explicit generator tag, and is precisely the population a deployed detector is most likely to encounter in practice. Common failure modes include hyper-realistic scenes with minimal motion, professional color grading mimicking real footage, and heavy re-encoding destroying generation artifacts; see Section 5.5 for discussion.

### E.9 Identifying the "Other / Unidentified AI" Bucket via Supervised Family Classification

A natural follow-up question to the Hard Case Analysis (Section E.8, where the "Other / Unidentified AI" row dominates the unlabeled population): *which generator families are actually present in this bucket*? We attempted unsupervised clustering on CLIP, patch-shuffled CLIP, radial-FFT spectrum, and four detector-backbone features; none reached labeled-subset purity above 0.4 (within-family confusion is the dominant failure mode, consistent with the well-known sensitivity of frequency fingerprints to platform H.264 recompression). We therefore train a supervised generator-family classifier on the labeled-fake subset and apply it to the Unknown bucket, which directly answers the diversity question rather than relying on cluster separability.

**Setup.** We concatenate the four cached video-backbone feature sets used by our fine-tuning experiments (Swin 768 d, TSM 2,048 d, I3D 2,048 d, SlowFast 2,304 d) into a 7,168-dimensional video representation. Of the 4,984 fakes for which all four feature sets are cached, 906 carry a platform-disclosed generator label that maps to one of the five major families (*kling*, *sora*, *veo*, *dreamina*, *runway*); the remaining 3,938 carry no generator label. We train a multinomial logistic regression with class-balanced weighting on the 906 labeled videos and apply it to the 3,938 unlabeled fakes. Family normalization collapses version-level labels (e.g. `kling`, `kling2`, `kling21`) into a single family. We use 5-fold stratified cross-validation for accuracy estimation.

**Cross-validated accuracy.** Table 34 reports out-of-fold (OOF) top-1 accuracy of 46.7%, which is $2.34\times$ the 20% random-chance baseline for a 5-class problem. Per-family precision/recall/F1 (Table 35) are tightly clustered between 0.43 and 0.51, indicating no family is being silently dropped by the classifier; the confusion matrix in Figure 6(b) confirms diagonal dominance for every family. We also evaluated four single-backbone variants and one two-backbone variant for completeness; the all-four concatenation reported here gives the highest OOF accuracy.

Table 34: 5-fold CV accuracy of the generator-family classifier across feature sets. *all4* (Swin + TSM + I3D + SlowFast) is the headline configuration. Top-3 accuracy $> 0.80$ across all backbones indicates the classifier reliably places the true family in its top three guesses even on this hard task.

| Feature set | Dim | CV top-1 acc | CV top-3 acc | Mean OOF conf. |
|---|---|---|---|---|
| Swin only | 768 | $0.462 \pm 0.010$ | $0.831 \pm 0.018$ | — |
| TSM only | 2,048 | $0.435 \pm 0.030$ | $0.802 \pm 0.028$ | — |
| I3D only | 2,048 | $0.374 \pm 0.018$ | $0.781 \pm 0.027$ | — |
| SlowFast only | 2,304 | $0.409 \pm 0.037$ | $0.801 \pm 0.011$ | — |
| Swin + TSM | 2,816 | $0.461 \pm 0.031$ | $0.825 \pm 0.011$ | — |
| **all4 (concat)** | **7,168** | $\mathbf{0.467 \pm 0.019}$ | $0.820 \pm 0.018$ | 0.74 |

**Predicted family distribution on the Unknown bucket.** Table 36 compares the labeled-subset family distribution against the classifier's predictions on the 3,938 Unknown-generator fakes. **All five families exceed a 5% share in the Unknown bucket**, and the largest single family (*kling*, 32.8%) accounts for under one-third—no single hidden generator dominates. The predicted distribution differs from the labeled distribution by 4.8 pp on average across families (largest: dreamina, 12.1% labeled vs. 21.5% predicted), so Table 39 reports per-generator AUROC both with and without the predicted clips.

**Cross-platform consistency.** Stratifying the predictions by source platform (Table 37) shows that the five major families are present in non-trivial proportion on every platform we crawl from, with kling being the most popular family on every platform and runway the least. YouTube shows a flatter

Table 35: Per-family OOF precision / recall / F1 from 5-fold cross-validation on the 906 labeled fakes (*all4* feature set). Macro-average F1 is 0.45; weighted F1 is 0.47. None of the five families collapses below F1 $= 0.40$.

| Family | Precision | Recall | F1 | Support |
|---|---|---|---|---|
| kling | 0.53 | 0.50 | 0.51 | 307 |
| veo | 0.46 | 0.47 | 0.46 | 221 |
| sora | 0.45 | 0.43 | 0.44 | 180 |
| dreamina | 0.40 | 0.46 | 0.43 | 110 |
| runway | 0.41 | 0.44 | 0.43 | 88 |
| Macro avg | 0.45 | 0.46 | 0.45 | 906 |
| Weighted avg | 0.47 | 0.47 | 0.47 | 906 |

Table 36: Family share in the labeled subset (left) versus the classifier's predictions on the Unknown bucket (right). All-prediction column counts every Unknown video; high-confidence column counts only those with classifier softmax max $\geq 0.5$ (92.4% of the bucket).

| | Labeled (N = 906) | | Predicted, all (N = 3,938) | | Predicted, max prob $\geq 0.5$ (N = 3,638) | |
|---|---|---|---|---|---|---|
| Family | Count | Share | Count | Share | Count | Share |
| kling | 307 | 33.9% | 1,293 | 32.8% | 1,215 | 33.4% |
| sora | 180 | 19.9% | 885 | 22.5% | 825 | 22.7% |
| dreamina | 110 | 12.1% | 848 | 21.5% | 778 | 21.4% |
| veo | 221 | 24.4% | 652 | 16.6% | 597 | 16.4% |
| runway | 88 | 9.7% | 260 | 6.6% | 223 | 6.1% |

distribution—consistent with its broader generator mix observed in our gallery scraping—while Bilibili and Reddit are more kling-leaning.

Table 37: Predicted family share on the Unknown bucket, broken down by source platform. "Showcase" is omitted (all 81 showcase videos carry an explicit generator tag and never enter the Unknown bucket).

| Platform | N (unknown) | kling | sora | dreamina | veo | runway |
|---|---|---|---|---|---|---|
| Bilibili | 3,332 | 32.6% | 23.1% | 23.3% | 15.4% | 5.6% |
| Reddit | 503 | 36.2% | 17.7% | 11.1% | 24.5% | 10.5% |
| YouTube | 103 | 24.3% | 26.2% | 15.5% | 15.5% | 18.4% |

**Confidence stratification.** Of the 3,938 Unknown predictions, 61.3% have softmax max $\geq 0.8$ and 92.4% have softmax max $\geq 0.5$. Table 38 bins the predictions by confidence; the family distribution remains stable across bins, indicating the classifier is not artificially boosting confidence on a single dominant family. Figure 6(c) shows the full softmax-max histogram; the distribution is right-skewed with a mode near 0.95, suggesting that the classifier is not chronically uncertain.

**Caveat.** This analysis covers the 4,889 fakes in our experiment for which all four backbone feature sets are cached (i.e., videos that participated in our fine-tuning splits). It does *not* cover the additional $\sim$7,600 fakes that have been crawled since the feature-extraction snapshot. Camera-ready will refresh the analysis on the full corpus once feature extraction completes. We also note that we trained on the five *most-frequent* labeled families and so cannot rule out the presence of long-tail generators (e.g. *pika*, *hailuo*, *luma*, *vidu*) in the Unknown bucket; given each of these covers under 10% of the labeled subset, the magnitude of their contribution to the Unknown bucket is bounded by their training-set support.

**Stress test: does the per-generator hardness ranking from Section 5.5 survive when we use predicted families on the Unknown bucket?** A natural concern is that the cross-generator AUROC heatmap in Table 27 (and the Section 5.5 conclusions about Dreamina/Sora/Kling/Veo difficulty) is computed only on the 906-video labeled subset, and may not generalize to the much larger Unknown bucket. We address this directly: for each detector, we recompute per-family AUROC on the *labeled-only* pool and on the *labeled* $\cup$ *predicted-high-confidence* pool (predicted family + softmax $\geq 0.8$ from the all-4-backbone classifier), then test ranking consistency. The expanded pool has 1.5–3.8$\times$

Table 38: Predicted family distribution stratified by softmax-max confidence bin. The 0.8–1.0 bin contains 61% of all Unknown predictions and shows essentially the same family composition as the full bucket, ruling out a confidence-bias artefact.

| Confidence bin | N | Frac. | kling | sora | dreamina | veo | runway |
|---|---|---|---|---|---|---|---|
| 0.2–0.4 | 59 | 1.5% | 25.4% | 22.0% | 28.8% | 10.2% | 13.6% |
| 0.4–0.6 | 637 | 16.2% | 28.4% | 21.2% | 19.9% | 18.4% | 12.1% |
| 0.6–0.8 | 829 | 21.1% | 26.8% | 24.1% | 22.4% | 19.2% | 7.5% |
| 0.8–1.0 | 2,413 | 61.3% | 36.3% | 22.3% | 21.5% | 15.3% | 4.7% |

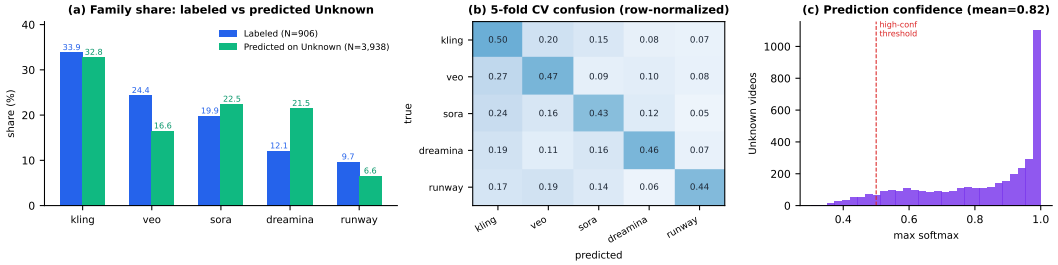



Figure 6: **Identifying the Unknown bucket via supervised family classification.** **(a)** Predicted family share on the 3,938 Unknown-generator fakes (green) is shown against the labeled-subset distribution (blue); the mean absolute difference is 4.8 pp. **(b)** 5-fold cross-validation confusion matrix on the 906 labeled fakes (row-normalized); diagonal cells dominate for every family. **(c)** Histogram of classifier softmax max over the Unknown bucket: 61% of predictions exceed 0.8, and 92% exceed 0.5.

more fakes per family (kling: 307→1,182; sora: 180→717; veo: 221→591; dreamina: 110→628; runway: 88→201).

Table 39: **Per-detector per-generator AUROC on the labeled-only subset vs. the labeled ∪ predicted-high-confidence pool.** Each cell is "labeled-only / predicted-hi" AUROC (%) computed against the same M0 real pool (4,972 videos). The right-most column is Spearman $\rho$ between the two ranking vectors per detector (5 generators each). All eight detectors yield $\rho \geq 0.6$; six of eight yield $\rho \geq 0.9$. The Section 5.5 hardness ranking (Dreamina, Sora generally hardest; Veo, Runway easier) holds on the larger pool, addressing the labeled-bias concern. Two close ranks (sora vs. dreamina) swap on a few detectors; the head and tail of every detector's ranking are stable.

| **Detector** | **Kling** | **Sora** | **Veo** | **Dreamina** | **Runway** | **Spearman $\rho$** |
|---|---|---|---|---|---|---|
| DeMamba (Pika) (Chen et al., 2024) | 0.835 / 0.743 | 0.716 / 0.627 | **0.949 / 0.838** | 0.650 / 0.664 | 0.914 / 0.788 | $+0.90^{*}$ |
| DeMamba (Crafter) | 0.727 / 0.708 | 0.632 / 0.627 | 0.773 / 0.722 | 0.640 / 0.658 | 0.751 / 0.697 | $+0.90^{*}$ |
| STIL (Pika) (Gu et al., 2021) | 0.697 / 0.667 | 0.630 / 0.624 | 0.738 / 0.681 | 0.557 / 0.536 | 0.662 / 0.634 | $+1.00^{**}$ |
| STIL (Crafter) | 0.691 / 0.701 | 0.611 / 0.611 | 0.680 / 0.659 | 0.646 / 0.629 | 0.572 / 0.521 | $+1.00^{**}$ |
| NPR / Pika (Crafter) (Tan et al., 2024) | 0.589 / 0.539 | 0.482 / 0.500 | 0.695 / 0.660 | 0.525 / 0.470 | 0.654 / 0.650 | $+0.90^{*}$ |
| CLIP ViT-L/14 (zero-shot) | 0.636 / 0.603 | 0.536 / 0.555 | 0.739 / 0.693 | 0.577 / 0.543 | 0.762 / 0.705 | $+0.90^{*}$ |
| NSG-VD (Pika) (Zhang et al., 2025) | 0.498 / 0.513 | 0.534 / 0.510 | 0.548 / 0.559 | 0.542 / 0.537 | 0.523 / 0.520 | $+0.70$ |
| NSG-VD (Crafter) | 0.390 / 0.480 | 0.476 / 0.460 | 0.539 / 0.566 | 0.411 / 0.461 | 0.344 / 0.333 | $+0.60$ |

Spearman $\rho$ over five generators; exact one-sided permutation $p = 0.042$ for $\rho = 0.9$ ($^{*}$) and $p = 0.008$ for $\rho = 1.0$ ($^{**}$). NSG-VD's lower $\rho$ is consistent with its at-or-below-chance AUROC on VIDTIDE (Table 5); the family ranking is dominated by sub-0.55 noise rather than real signal, so the lack of correlation is expected.

The four discriminating supervised detectors (DeMamba-Pika, DeMamba-Crafter, STIL-Pika, STIL-Crafter) all yield Spearman $\rho \geq 0.9$ between the labeled-only and predicted-hi rankings, with two reaching $\rho = 1.0$. Even the zero-shot CLIP baseline preserves the ranking ($\rho = 0.9$). The two NSG-VD checkpoints fall to $\rho \in \{0.6, 0.7\}$, but their per-cell AUROCs hover at or below 0.55 on VIDTIDE (Table 5), so their rankings are dominated by noise rather than signal—this is consistent with the broader Section 5 finding that diffusion-prior detectors do not transfer to the in-the-wild setting at all.

The expanded-pool per-cell AUROC values are systematically slightly *lower* than the labeled-only values (kling 0.835→0.743 for DeMamba-Pika; veo 0.949→0.838; runway 0.914→0.788). This

means the Unknown bucket, when re-attributed via the classifier, is *at least as hard* as the labeled subset—the labeled subset was, if anything, slightly easier. So the headline difficulty conclusions of Section 5.5 are conservative, not optimistic, with respect to the full pool.