BENCHMARKS LIE, SIGNALS REVERSE: A FORENSIC AUDIT OF AI-GENERATED VIDEO DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-generated video detection has run an implicit measurement program for a decade: assemble real and generated video, train a separator, report benchmark accuracy. We audit the program and report its failure mode. Without precedent, to our knowledge, the field's strongest prior, motion amount, reverses sign across generator generations, fake-to-real flow ratio 0.362 on a 2024 generator against 3.342 on Seedance-2.5, and the 2026 wall is a detector-reference pairing: on Seedance every confounded-corpus system fails or inverts, and the standardized corpus's own system reads 0.4478 below chance on one reference family, above on another. The infrastructure failure is published, VidAudit's: of ten surveyed families four fall to zero-feature container rules that decode no frames, and on GenVidBench a duration threshold reaches a deployable train-fit 79.87 official while a constant FAKE prediction scores 80.00, above ten of eleven published systems. The surveyed signals fail our falsification battery: eight gates, each with a demonstrated kill, leave 0 of the 17 tested frozen-estimator families standing as a universal detector; the battery and family list co-evolved, so the zero reads as an exploratory falsification ledger, not a closed census. The systems layer survives only under a container-standardized protocol: 0.865 pooled AUC zero-shot on the one standardized benchmark, a hundred in-family labels reading 0.996 on the wall pool, near-duplicates unexcluded, off-family content ranking 0.744 with 8% detected. The limits are stated, not softened: single-seed rows, a small wild track, a frozen-estimator survey. The contribution is the protocol, the falsifications, and a label-priced adaptation ladder.

1 INTRODUCTION

A program run without a protocol. Detection of AI-generated video has spent a decade inside one measurement program: collect real video, generate fake video, train a separator on the pair, and report accuracy on a benchmark. The program presumes the two halves differ by generation alone, so whatever the separator learns must be generation; that presumption has failed in public for years, one shortcut at a time, and the field has patched instances rather than audit the program. This paper is the audit.

The failure mode. Every time a control closes one channel between the two halves of a benchmark, another remains open. The sequence runs in this paper's own order: the file container (duration, frame rate, resolution, codec, readable before a single frame is decoded), then the padded border of the frame, then the content of the clips, then their motion statistics, and finally the generation stage, which hands the detector a channel even when everything else is held fixed. None of these channels is exotic; each is a place where collected and generated video differ for reasons that have nothing to do with being generated.

The limiting observation. The sequence has a limit, and stating it clarifies what a benchmark can honestly promise. Suppose a construction controlled every channel perfectly: real and generated video would then be identically distributed in every channel available to measurement, and the property "AI-generated" would live only in the generation history of the file, which no measurement of the file reads. Under that limit every detector retreats to chance, however good it is: perfect control dissolves the task rather than defeating the detector. We state this as an observation about measurement, in words,

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.

Claim	Decisive control	Licensed reading
Shortcuts decide the surveyed benchmarks	zero-feature container rules (§2.1)	container columns beside accuracy
The sampler re-imports duration	scalar-duration learner vs. bank (§2.3)	pipelines inherit container fields
Direction reverses across generations	cross-generation gate (§3.3)	recalibrate direction per generation
The 2026 wall follows the training corpus	corpus swap at fixed recipe (§4.3)	standardize the corpus, not one axis
The clean control is a floor	round-trip, border, content bounds (§2.4)	GenBuster is the minimum, not the goal
The instrument is the contribution	eight gates plus self-falsification (§3.1)	readings are earned by surviving gates
Not claimed		no state of the art, architecture, or solved generalization; no universal detector failure, direction law, or cross-reference robustness

Table 1: The paper's audit map, the contract the text keeps: each load-bearing claim with the control that decides it and the reading it licenses. The final row lists what the evidence does not support.

and claim no impossibility theorem from it. Its practical content is humbler: no benchmark in this field is unbiased in the absolute sense the program implicitly wants, because the limit itself forbids it.

From unbiasedness to falsifiability. Our response is to stop pursuing the unreachable and start demanding the checkable: a detection claim becomes trustworthy when it has survived named controls that would have killed it if it were riding a shortcut, whatever its benchmark looks like. We built those controls while failing to defend two method claims of our own; assembled into a battery, they are the instrument this paper contributes.

Published anchors. The forensic core has published precedents: VidAudit's clip-length classifier already reaches a leave-one-generator-out AUC of 0.998 on GenVidBench (Cakiroglu et al., 2026), and two concurrent audits document the GenVideo family's duration and frame-rate asymmetries (Ben Hayun et al., 2026; Michels et al., 2026). Our increment is specific: we execute container rules under a benchmark's own official metric, header-only on the as-released files, and we trace the route by which a learned detector inherits the container field through its sampler.

Three findings. *First* (§2), the infrastructure fails before detection begins: of 10 surveyed benchmark families, 4 fall to zero-feature container rules and 5 more confess the confound in their own documentation. *Second* (§3), the signal space fails an eight-gate battery: 17 families tested, zero universal survivors, the motion prior that anchors several published detectors inverting across generator generations. *Third* (§4), detection survives only on the one container-standardized benchmark, at 0.865 pooled AUC zero-shot, and is recovered at the 2026 wall only by a ladder priced in labels.

What a benchmark score measures. Every accuracy in this field estimates an operational quantity: how well a separator predicts the labels of one fixed construction of two video populations. Without further argument it is not an estimate of generator-invariant visual detection; the two coincide only under a construction that has closed every shortcut channel, and no surveyed family of constructions has fully done so. This estimand gap, more than any detector's weakness, is the paper's subject.

Our own claims, falsified by our own controls. Two of our prior unpublished submissions claimed detection gains from physics-consistency and coherence probes; controls built to defend them falsified them instead, and generalized became this paper's instrument. The battery's validation set is our own falsified claims.

2 THE FORENSIC FINDING: BENCHMARKS FALL BEFORE DETECTION BEGINS

The short form of this section: in most of the field's benchmarks the two halves differ in the file wrapper that surrounds the pixels, so a rule that never decodes a frame already separates them, and the official metric pays for it.

2.1 CONTAINER RULES END THE TASK BEFORE IT BEGINS

The container. A video file arrives wrapped in a container whose header carries its duration, frame rate, resolution, and codec, all readable before a single frame is decoded; a zero-feature rule is a threshold on one of these fields. Every rule in this section is one line of ffprobe over per-video audit rows committed with the paper, re-executable by any reader on the as-released files. One exhibit decodes pixels, the overlay rule of §2.3, flagged where it appears.

The flagship case. GenVidBench (Ni et al., 2025) scores detectors over four generator columns (CogVideo, SVD, MuseV, Mora) against web-collected reals, averaged into one Mean Top-1. Its Pair1 training split uses Vript reals (Yang et al., 2024), its Pair2 test split HD-VG reals (Wang et al., 2023a), and our rules read Pair2. The real half runs long, median duration 11.0 seconds; the fake half runs short by construction, the longest Pair1 fake exactly 3.0 seconds and every Pair2 fake column pinned at a fixed duration up to 4.125.

One field suffices. Call FAKE any clip shorter than a threshold. An oracle threshold of 4.15 seconds, selected on the test files and therefore not deployable, scores an official 99.83 Mean Top-1; a threshold trained only on Pair1, deployable in the only sense that matters, still scores 79.87, among the benchmark's published systems. The two classes separate on this single field almost completely. The phenomenon is published and the credit is VidAudit's (Cakiroglu et al., 2026); we add the official-metric scoring and the header-only execution on the as-released files.

The same rule, a second family. GenVideo-Val, the official test split behind DeMamba (Chen et al., 2024), pairs collected reals against generator output, and the GenVidBench duration threshold transferred across benchmarks reaches 98.96 accuracy. That is an accuracy, not the F1 and AP the protocol reports, so an existence proof rather than a leaderboard entry. Frame rate and resolution separate the two sides alone as well; the confound was documented first by others (Ben Hayun et al., 2026; Michels et al., 2026), our addition again execution and official-protocol scoring.

The survey. We audited 10 benchmark families under a convenience-sample rule, those we hold locally plus prominent benchmarks whose papers document their own container columns. Four have fallen, meaning a zero-feature rule was executed locally on an as-released copy or the family's own paper admits the split: GenVidBench and both GenVideo splits above, GenWorld by its own table (Chen et al., 2025), and RA-Bench (Liang et al., 2026), audited on every file of its full release, where frame rate alone separates the sides. Five more are confessed in their own documentation (Song et al., 2024; Bai et al., 2024; Ma et al., 2024; 2026; Hu et al., 2025): the codec mismatch that excluded DVF, GVD's frame-format patch, GVF's admitted resolution and length inconsistency, AIGVDBench's wide frame-rate band, and the mismatched durations MVAD's own table prints. One family, GenBuster-200K (Wen et al., 2025), standardizes all four axes and survives (§2.4); the full ledger is Figure 2. On nine of the ten, a high score is not evidence of detection.

2.2 A METRIC THAT PAYS THE SHORTCUT

A four-to-one floor. The official GenVidBench metric averages five columns, four fake and one real, so its prevalence is four to one and a classifier that leans fake is paid on four columns for every one it is fined on. A constant FAKE prediction reads 80.00 under the official metric, above ten of the eleven published systems on the matched protocol, and on the benchmark's other published protocol the constant sits within a fraction of a point of the best published headline. Constant-predictor and prevalence floors are classical measurement points (Provost et al., 1998; Schaeffer et al., 2023), and prevalence-aware rescoring moves deepfake leaderboards (Nguyen et al., 2026); we claim only this family's instantiation, unprinted before.

Our own system as the meter. To exhibit the metric from the inside we trained a single-seed Pair1 classifier over our frozen bank of seventeen motion features, the motion-statistics expert of §4 wearing its instrument hat. It scores an official 84.29, above the matched-protocol published best and above the constant, while recalling fewer than half of the real videos. The real column is an operating point set by prevalence, not a measurement of anything, and under balanced rescoring the same scores fall from first to tenth of twelve. A metric exhibit, then, not a detector we offer for use.

2.3 How a learned detector inherits the container

The sampler is the door. A learned detector that never reads a duration can still carry one. The evaluation samples 8 frames centered in a 3-second window, spacing them evenly over the shorter of the clip's duration or the window. On the short fake columns the spacing scales with duration; on the long real columns the window caps it. Only the fake side's spacing carries the duration field, and at matched scene speed a shorter spacing mechanically measures less motion: the field enters the features through the sampler's arithmetic.

A scalar outruns the bank. Train a logistic regression on the scalar duration alone, nothing else. It reaches a pooled AUC of 0.9968 on the official pool, against 0.7656 for the seventeen-feature bank on the same pool. One container field carries more benchmark information than an entire learned feature bank. That is the inheritance route made quantitative: only a pipeline that samples a duration-capped window can see the field, so only it inherits the field.

What remains inside the band. Restrict both sides to the duration band where they overlap, upper edge the longest pinned fake duration 4.125 seconds, and the comparison loses its real half: 106 reals survive against tens of thousands of fakes. Inside it the bank's pooled AUC falls to 0.6947, with motion-amount matching pushing it further toward chance: a motion-quantity prior, not structure.

Moving the operating point is not moving the detector. Retraining on a duration-balanced mix leaves pooled discrimination where it was while the fixed operating point slides into the nearly-all-fake corner the four-to-one prevalence pays for: the intervention moves threshold behavior, not what the classifier detects (Appendix C.11).

Three layers. The decomposition has a container layer that separates the classes outright, a content layer that survives inside the band at a much lower ceiling, and a threshold layer the official four-to-one metric converts into points at zero detection cost. Published leaderboard numbers on this family are explained by the three together, without any detector detecting anything.

A pixel-level sibling. The container is not the only pre-detection channel: a visual label pass finds persistent screen graphics, station logos and news banners, on 88 percent of GenVideo-100K reals, read lower at 69.5 by a hardened rerun labeler, against 0.5 percent of GenBuster fakes, read at 2.5 by the rerun. A zero-training static-overlay rule, sixteen frames and within-clip quantiles with no learned constant, separates scraped reals from text-to-video fakes at pooled AUC 0.7366 to 0.890 across nine pool pairings (Appendix C.12, Figure 1). Two fences hold: video-driven generators inherit the channel from their source frames, so it marks construction rather than generation, and it does not decompose any trained detector's number. On the four-axis standardized benchmark only a small prevalence difference remains.

2.4 The one container-standardized benchmark

One standard, verified on disk. GenBuster-200K (Wen et al., 2025) ships every file on both sides of its splits at one duration, one frame rate, one resolution, and one codec; we verified this on the local copy rather than trusting the paper, probing every file, and on the mini split the single-axis duration, frame-rate, and resolution rules all read exactly 0.5. The container door is closed here, which is what makes the benchmark usable as a control.

Three bounds on what its cleanliness buys. Closing the container does not close the task. First, a pure round trip through a video autoencoder, no generator, no sampler, no content change, already separates from its own source for every deep detector in our roster, in both directions; any clean-protocol reading therefore passes through a shared encode lineage rather than codec-free signal. Second, cropping the padded border band alone removes up to 0.298 of the standardized-corpus detector's wall reading, so the border stays open. Third, a content-matched construction, first-frame-conditioned

generation from the reals themselves, lowers the same detector from 0.928 to 0.8539, so content carries part of what the container was blamed for.

A floor, not a destination. The honest reading of GenBuster is as the floor the field must clear before its numbers begin to mean detection. The channels above it, border, content, motion, and the generation stage itself, are each measurable and each still open somewhere in every construction we know, and at the far end of that sequence sits the limiting observation of §1: a construction that closed every channel would leave nothing to detect.

3 The Instrument: Eight Gates, and What They Falsify

This section turns the same discipline on the signals themselves: one battery of controls, applied without modification to every computable signal family, every falsification recorded whether the claim was ours or the literature's.

3.1 The battery

Eight gates, one line each. Each gate is a control with a pass condition and at least one demonstrated kill, its specimens tabulated before and after in Table 12.

- **g1, compression.** Recompress heavily; a reader whose discrimination moves materially is reading the codec. It killed our latent-branch detector, whose fragility was recompression-specific, not a general high-frequency dependence.
- **g2, motion-amount matching.** Re-weight the reals to the fakes' motion distribution and retest; a statistic that dies was a motion-quantity proxy. It falsified the motion-smoothness family and took our own bank's headline column to exactly chance.
- **g3, cross-generation direction.** The discriminative sign must hold across generator generations; the flow-cycle family reads fakes as more regular than reals on 2024 generators and less regular on Seedance, which inverts it.
- **g4, zero-shot transfer.** A bank member must transfer to an unseen generator without labels; it killed our band-limited spectrum claim, jointly with g6.
- **g5, structural null.** A mechanism claimed at a scale must peak at that scale and vanish under its specificity nulls; the token-grid family failed both halves.
- **g6, resolution equalization.** Re-render both sides to a common short side. Five of the six image-forensics classics flipped sign on Seedance-2.0.
- **g7, constant floor.** Score a constant predictor under the metric in question; on GenVidBench the all-fake constant outscores the published mid-table (§2.2).
- **g8, prevalence inversion.** Re-score under balanced accuracy and pooled AUC; our flow bank, officially first of twelve, falls to tenth (§2.2).

Two audit modes complete the instrument, labeled as audits rather than gates because no control in the battery covers them: a leakage audit for channels that ride the plumbing and a protocol audit for pre-registration and label access, which together falsified two more claims of ours (§3.4).

A proposed community checklist. The battery is cheap to run, its gates being controls rather than trainings, and applicable to any claimed AIGV detection signal or any benchmark that proposes to measure one. A claim earns its reading by passing the gates that apply to it; a benchmark earns its scores when a zero-feature rule cannot take them.

3.2 Seventeen families, zero universal survivors

Scope and count. Under the survey's scope rule, computable from decoded frames with frozen estimators, 19 families entered, 17 were tested, one remains queued and one is data-blocked rather than excluded. Of the seventeen, 0 survive every applicable gate in every domain. The per-family ledger, each family with the one-line verdict that killed it, is Appendix B.

A bounded zero. The zero is a verdict about this list under these gates, not about the signal space. The battery and family list co-evolved with our own probe line, stated rather than hidden; a missed family can be added and the battery applied unchanged, and a single survivor would overturn the count. The fence excludes learned, augmentation-trained detectors, whose certification elsewhere (Michels et al., 2026) does not conflict with our verdict.

Two levels, not one. The universal level is empty while the bank level is populated: a feature clears the bank bar by reaching a moderate margin from chance in any single domain after its applicable gates, and several do. The flow-magnitude prior, dead as a universal claim, separates its best 2024 domain at 0.955. The bank is a per-domain toolbox, not a detector; its directions must be re-earned per domain.

Multiplicity, accounted. The admission surface is fixed before the null: 182 feature-by-domain comparisons, of which 138 met the bank bar and are the ones the permutation null and Benjamini–Hochberg control adjudicate. All admitted cells clear both; the cells below the bar enter no verdict. The survey cannot be read as a fishing expedition that happened to reject everything, because the rejections are counted against a committed surface.

3.3 Two regularities behind the zero

The iron law of single-domain fits. The survey's domains share one real pool, 97 percent measured overlap, so a feature's learned direction is defined against one pool's capture signature as much as against its generators. On top of that, a joint model trained on one domain transfers below chance across generators even under leakage control: 0.198 signed AUC from the strongest 2024 domain to Seedance. Single-domain joint fits do not merely weaken out of domain; they invert.

Direction, inverted. The second regularity is the inversion itself: every falsified family has at least one domain where the sign flip alone accounts for the kill, so the universal level closes by mechanism rather than accumulated near-misses. The sharpest instance is motion amount, the prior under several published detectors. The fake-to-real flow-magnitude ratio reads 0.362 on CogVideoX and 0.139 on Wan-2.2, where 2024 fakes move less than reals, and 1.217 on Seedance and 3.342 on Seedance-2.5, where fakes move more; every interval excludes one, the Seedance-2.0 value is corrected for a frame-duplication artifact in the appendix, and the signs replicate fully against an independent-family reference while the magnitudes move. A 2024-trained motion expert inherits the prior and lands below chance on Seedance. This contradicts the unconditional under-motion prediction printed elsewhere (Cakiroglu et al., 2026) and realizes, as a measurement, the obsolescence Michels et al. (2026) forecast for the prior. Decay under generator progress is a published genre (Wang et al., 2020; Cozzolino et al., 2023a; Thakur et al., 2026); sign inversion, to our knowledge, is not. We call it a generation shift, not a law.

3.4 The validation set: five claims of our own

The battery's credibility rests on what it did to us: five claims of ours, committed before evaluation, were falsified by the gates or the audit modes, and Table 5 records each with its honest replacement. Two examples carry the structure. The leakage audit found a five-class attribution margin riding a missing-value mask in the feature cache; the mask alone reproduced most of the apparent discrimination, and re-extracting every missing cell from video at source dissolved the margin to its honest value. The g1 gate found our latent-branch detector's accuracy recompression-specific, a codec reader rather than a video reader. None of the five received a post-hoc reanalysis; each was retracted or restated at the number the control left standing. An instrument that kills its builders' claims first is the only kind worth trusting on anyone else's.

4 The Systems Layer: Detection Under Standardization

With the infrastructure audited and the instrument built, this section reports the systems layer under the discipline the first two sections motivate. Three declarations govern what follows: every system row is single-seed except where a replicate seed is stated, every rate and ranking carries its protocol and sample size, and every row scored on a confounded benchmark is a metric exhibit rather than a performance claim.

4.1 Benchmark-trained systems ride the container

Below the constant. On GenVidBench's official protocol our own systems sit under the floor of §2.2: the physics backbone, a ViT classifier with a coherence-prediction auxiliary, scores 51.29 against 45.50 for the plain baseline, a VideoMAE on the identical split and loader. The auxiliary's name describes the recipe, not a certified mechanism: the matched control of §3.4 found

a rotation-prediction auxiliary with no physics content scoring higher. Both sit far below the all-fake constant, which is the point of printing them.

Move the container, watch the score fall. The container claim survives intervention, not just rules. We retrained the benchmark's own baseline recipe, the TSM row of its published table (Lin et al., 2019), published Mean Top-1 74.13, our run's divergences declared in the appendix. The retrained run reads 78.29 percent real-class accuracy natively; re-encoding only the real side of the test pool to each fake column's measured container profile, one column at a time, drops it to 30.78 at the deepest column, every paired interval excluding zero. The fakes are untouched; nothing changes but the wrapper around the reals. A released checkpoint with lineage disjoint from ours agrees: BusterX++ (Wen et al., 2025) loses 9.10 points of pooled real-class accuracy under the identical intervention, interval 7.42 to 10.75.

4.2 Zero-shot on the clean container

The tracks. GenBuster scores in-domain training (id), held-out generators (ood), and a wild mix of recent commercial generators, 150 clips at ten to twenty per generator whose per-generator cells certify little and whose pooled readings carry the weight; the ood and wild reference AUCs borrow the id track's 1000 reals, an asymmetry that is a mini-instance of §2.

The clean anchor. A vanilla VideoMAE (Tong et al., 2022) trained on the benchmark's own clips, the clean anchor, reaches 93.2 id and 86.3 ood accuracy, cross-recipe against the published row, excluded from zero-shot comparisons; on the wild mix its ranking decays less than its accuracy, reference AUC 0.8156.

Zero-shot transfer, honestly bounded. Trained on the confounded Pair1 corpus, scored zero-shot on the clean container, the physics backbone pools 0.865 on the mini split (single training draw). Its paired advantage over the plain baseline reads 15.59 points there and 5.05 on the official id track; the preregistered second seed reproduces the ordering only, at 14.11 and 5.38, both intervals excluding zero; a disjoint real draw moves nothing (Appendix C.3). On the id track the GenVideo backbone reads ahead by a paired -5.32. On the wild mix every system decays or inverts, the physics backbone to 0.702, except the GenVideo backbone, which ranks highest there at 0.860 (n = 150) while the motion-statistics expert falls to 0.471 from 0.769 on the previous track: ranking and operating point are different questions with different leaders; only the ranking transfers.

4.3 The 2026 wall

The phenomenon. On the 2026 Seedance generations every zero-shot system trained on the field's confounded corpora fails or inverts against the OpenVid-family reference, the web-collected reals the benchmark's id track uses. The strongest confounded reading on the container-normalized pools is the plain baseline's 0.588 on Seedance-2.0 and 0.574 on Seedance-2.5, both weak; the physics backbone reads 0.153 and 0.154, both inverted; the motion-statistics expert sits inverted at 0.348. The wild mix locates the same decay but cannot certify direction at its handful of clips per generator.

Per-generation scale. We re-measured the wall on its own pools: 1000 self-collected Seedance-2.0 fakes from the generator's public showcase prompts, 995 decodable, plus 183 Seedance-2.5 clips, every clip re-encoded to the GenBuster container so the zero-feature rules read exactly chance and the gate is closed by construction; a bitrate-matched control moves the inversions by at most two thousandths, so rate is not the carrier (Appendix C.4).

The cross-source matrix. Whether the wall belongs to the detectors or to the detector-reference pairing was preregistered over two independent reference families before scoring: the same Seedance-2.0 fakes re-scored against HD-VG and Vript reals in place of OpenVid, the matrix of Table 2. The inversions do not survive: every inverted lineage, and the physics backbone at both seeds, reads above chance against both families, the clean-room DeMamba recovering to 0.703 and 0.7351. The holdings are not exempt: the corpus-retrained DeMamba's OpenVid holding drops below chance against HD-VG at all three preregistered seeds, 0.4478, 0.4324, deepest at the third seed's 0.3802, while Vript reads chance at two seeds and below it at the third (0.419). A motion-matched re-read collapses the motion-magnitude prior to 0.5038 under the same protocol, so the pairings are not readable off motion amount. The multireference, multiseed retest (Appendix C.6), its decision rule frozen before any run, returns the paired verdict on Seedance-2.0: the holding itself is an OpenVid-family pairing,

System (Seedance-2.0 fakes)	OpenVid	HD-VG	Vript
physics backbone (seed 42)	0.153	0.5495	0.5664
plain baseline (seed 42)	0.588	0.6099	0.5846
retrained TSM	0.1386	0.6212	0.6002
clean-room DeMamba	0.192	0.703	0.7351
corpus-retrained DeMamba	0.928	0.4478	0.4981
clean anchor (seed 43)	0.856	0.6364	0.6314

Table 2: The cross-source wall matrix: reference AUC of container-normalized Seedance-2.0 fakes against three real-reference families. OpenVid is the in-family reference the wall was first read against; HD-VG and Vript are independent collection lineages. Inversions against OpenVid wash out against the independent families, while the corpus-retrained holding inverts against HD-VG; both directions of the wall are detector-reference pairings. Multiseed verdicts and intervals: Appendix C.6.

while on Seedance-2.5 the same system holds all three reference families, the grid's only such reading, so the wall is generation-specific as well as reference-specific. The clean anchor is paired on both generations at its two seeds, holding 0.856 and 0.886 in family while both independent families sit above chance and below the robustness bar at every seed. No tested recipe is reference-stable on Seedance-2.0; both wall readings, failures and holdings, are detector-reference pairings.

The wall is not ours alone. The clean-room DeMamba, the published architecture retrained on Pair1, divergences declared, inverts on both generations, 0.192 and 0.205; the retrained TSM likewise at 0.1386 and 0.1123, both seeds below chance. The same DeMamba, scored zero-shot on the GenBuster wild mix, ranks it at 0.748 including the 15 Seedance-2.0 clips it inverts elsewhere. The corpus decides: the identical architecture and recipe on the container-standardized GenBuster corpus holds the same pools at 0.928 and 0.976 with the wild mix at 0.876, while its fixed-threshold detection collapses to 0.042: the ranking crosses the corpus change and the operating point does not.

Fresh generators reproduce the split. Four further single-generator pools of 500 clips, from open-weight generators released after every audited detector's training data, re-run the protocol (Appendix C.5). The split reproduces completely: the standardized architectures hold 0.9366 to 0.9963, the confounded-corpus systems span 0.018 to 0.678, 25 of 28 cells deviate from chance after Benjamini–Hochberg control, and a disjoint real draw changes no conclusion.

The causal complement. If the corpus is the carrier, normalizing one axis of it should not suffice, and does not: the plain baseline retrained at the same recipe and seed with only the training manifest's container normalized inverts the same pool at 0.056, deeper than its native reading and reproduced at a second seed with every interval below chance: rate matching does not remove the inversion, and container normalization alone produces one. The preregistered corner is starker: the clean-room DeMamba trained on the container-normalized Pair1 manifest reads 0.01432 against its native corner's 0.192. The corpus carries content as well as containers, and swapping the fake pool's content alone moves the physics backbone by a paired -0.456. The wall follows the training-distribution bundle, and only the standardized corpus, not any single axis of repair, moves a system across it.

4.4 What survives, and what it costs

Triage, not a detector. One system-level capability survives the audits usefully: generator attribution as triage. After the leakage fix of §3.4 the five-class classifier reads macro-F1 0.889, and its operational role is to tell an operator which generator generation a suspect clip belongs to, which selects the direction-calibrated bank and, at the wall, which family's labels to buy; fed the wrong family it is harmful, and it is not a gate-certified detector.

No fusion claim survives. On the clean container three-system fusion trails the solo physics backbone by a paired -2.70: adding the weaker systems subtracts, with no decision rule to blame, a retraction of the fusion rows our earlier drafts carried.

The priced ladder. Adaptation to the wall is a ladder priced in labels: calibrating the frozen bank's directions with labeled clips recovers almost all of them while fused AUC stalls at 0.712 against the 0.7203 label-peeking oracle: the ceiling lives in the features, and no amount of signing fixes it. Whether fine-tuning breaches that ceiling was preregistered, and it does, upward, on the pool's

895-clip held-out split. Ten in-family labels take the plain baseline past the oracle at 0.831; a hundred take the physics backbone to 0.996; the frozen motion-statistics expert stays below chance at the same budget: deep representations recover with the generation's own labels, the frozen bank does not. These controlled in-family results are not a deployment procedure (same-prompt near-duplicates cannot be excluded here); the transfer rows and saturation bookkeeping are Appendix C.10. The negative rung matters as much: wrong-family labels invert the model past its zero-shot reading by a paired -0.342, a fingerprint of family rather than clip memory, and the spillover is priced too: the fine-tuned plain baseline ranks the wild mix at 0.744 while detecting 8% of it.

Two constructive endings. Both constructions we preregistered localize signal. In the first (Appendix C.7), an image-to-video generator conditioned on the first frame of each of 300 source reals locks content; the motion match fell short of its preregistered threshold, but the corpus split survives: every confounded-corpus detector inverts while both GenBuster-trained systems detect, the corpus-retrained DeMamba at 0.963 paired. In the second (Appendix C.9), the regeneration-strength dial regenerates the same 300 sources at a controlled mixing strength with content, motion, and container held, zero strength a pure autoencoder round trip that every deep detector reads: the retrained TSM scores it more fake than its own source at 0.867, three primary detectors invert with the GenVideo backbone at 0.207, and the curve peaks at 0.922 for the GenVideo backbone at the pure-noise end. Lock the container, border, content, and motion, and the generation stage still hands the detector a channel: clean is a direction, not a destination, and the audited systems layer says which channels remain open.

5 Related Work

Benchmark audits and bias layers. Three 2026 audits overlap the forensic finding: VidAudit (Cakiroglu et al., 2026) re-scores two benchmarks under six controls, STALL (Ben Hayun et al., 2026) documents the GenVideo asymmetries and standardizes the protocol, and Michels et al. (2026) decompose four motion detectors into preprocessing and prior bias. None scores a container rule under an official metric, executes zero-decode header rules, or traces the sampler route; that is our increment. Container-adjacent bias layers are established image-domain prior art (Grommelt et al., 2024; Choi et al., 2025; Guillaro et al., 2025), and Corvi et al. (2025) acted on the codec confound earliest, codec-matched retraining reversing a detector ranking well before our audit.

Decay and its genre. That detectors weaken as generators improve is a named genre (Wang et al., 2020; Cozzolino et al., 2023a; Thakur et al., 2026); our addition is the sign inversion, unreported by the genre.

Measurement science. The paper's lineage is shortcut learning and dataset bias: datasets recognizable from their samples alone (Torralba & Efros, 2011), a deployed recognizer riding camera-trap location (Beery et al., 2018), Clever Hans predictors unmasked by attribution (Lapuschkin et al., 2019), networks using whatever correlated feature is cheapest (Geirhos et al., 2020); constant floors and metric dependence are classical evaluation points (Provost et al., 1998; Schaeffer et al., 2023; Nguyen et al., 2026) we instantiate on AIGV leaderboards. The surveyed frozen-estimator families descend from the field's lineages, flow and depth consistency (Chang et al., 2025; Zheng et al., 2025), temporal regularity (Liu et al., 2024; Internò et al., 2025; Ma et al., 2024), latent reconstruction (Wang et al., 2023b; Ojha et al., 2023); the field's own map of that space (Hou et al., 2026) runs no falsification battery, the gap our survey occupies, and generator attribution (Kundu et al., 2025) and content-level shortcuts (Fang et al., 2026) are complements, not competitors.

6 Limitations and Conclusion

The instrument measures itself. The bank is both the audit instrument behind §2.2's metric exhibit and the audited motion-statistics expert of §4, a disclosed circularity. The survey's domains share one real pool at 97 percent overlap, and the full boundary list is Appendix D.7.

Established. Four of the ten surveyed families fall before detection begins, to container shortcuts, constant floors, and executable rules on as-released files.

Observed. Seventeen tested families yield zero universal survivors under eight gates, the motion prior reversing across generations, and a 2026 wall following the training-distribution bundle: confounded-corpus systems fail or invert in-family yet recover on independent references, both sides pairings, the three-seed verdict.

Hypothesis, and close. A benchmark that prints no container columns, states no constant floor, and survives no zero-feature rule is not yet measuring detection. The program is executable tomorrow: report container columns beside every accuracy, state the constant floor, run a zero-feature rule first, calibrate direction per generation. Clean is a direction, not a destination.

REFERENCES

Jianfa Bai, Man Lin, and Gang Cao. Ai-generated video detection via spatio-temporal anomaly learning. *arXiv preprint arXiv:2403.16638*, 2024. PRCV 2024; constructs the GVD benchmark.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *European Conference on Computer Vision (ECCV)*, pp. 456–473, 2018. arXiv:1807.04975; camera-trap domain shift.

Ben Hayun, Betser, Levi, Kassel, and Gilboa. Training-free detection of generated videos via spatial-temporal likelihoods. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. arXiv:2603.15026; introduces STALL.

Mert Onur Cakiroglu, Zhihe Lu, Mehmet Dalkilic, and Hasan Kurban. Auditing generalization in ai-generated video detection: A six-control protocol and the VidAudit toolkit. *arXiv preprint arXiv:2606.31004*, 2026.

Chirui Chang, Jiahui Liu, Zhengzhe Liu, Xiaoyang Lyu, Yi-Hua Huang, Xin Tao, Pengfei Wan, Di Zhang, and Xiaojuan Qi. How far are AI-generated videos from simulating the 3D visual world: A learned 3D evaluation approach. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. arXiv:2406.19568; introduces L3DE.

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.

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

Hyeongjun Choi, Inho Jung, and Simon S. Woo. Combating dataset misalignment for robust AI-generated image detection in the real world. In *ACM Asia Conference on Computer and Communications Security (ASIA CCS)*, 2025. doi:10.1145/3709022.3736541.

Riccardo Corvi, Davide Cozzolino, Ekta Prashnani, Shalini De Mello, Koki Nagano, and Luisa Verdoliva. Seeing what matters: Generalizable AI-generated video detection with forensic-oriented augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 26418–26446, 2025. arXiv:2506.16802.

Davide Cozzolino, Andrea Pianese, Matthias Nießner, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. *arXiv preprint arXiv:2211.00680*, 2023a.

Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of AI-generated image detection with CLIP. *arXiv preprint arXiv:2312.00195*, 2023b. CVPR 2024 Workshop.

Ziyuan Fang, Tianyi Wei, Guanjie Wang, Weiming Zhang, Nenghai Yu, and Wenbo Zhou. Real data lies: Unveiling and closing the quality shortcut in generalizable ai-generated video detection. In *International Conference on Machine Learning (ICML)*, 2026.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. arXiv:2004.07780.

Patrick Grommelt, Louis Weiss, Franz-Josef Pfreundt, and Janis Keuper. Fake or JPEG? revealing common biases in generated image detection datasets. In *European Conference on Computer Vision (ECCV) Workshops*, 2024. arXiv:2403.17608; doi:10.1007/978-3-031-92089-9_6.

Fabrizio Guillaro et al. A bias-free training paradigm for more general AI-generated image detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2412.17671; introduces B-Free.

Hou et al. Detecting AI-generated video: A vision-language dual-view survey. In *Findings of the Association for Computational Linguistics: ACL*, 2026. arXiv:2607.10787.

Mengxue Hu, Yunfeng Diao, Changtao Miao, Jianshu Li, Zhe Li, and Joey Tianyi Zhou. MVAD: A comprehensive multimodal video-audio dataset for AIGC detection. *arXiv preprint arXiv:2512.00336*, 2025.

Christian Internò, Robert Geirhos, Markus Olhofer, Sunny Liu, Barbara Hammer, and David Klindt. AI-generated video detection via perceptual straightening. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2507.00583; introduces ReStraV.

Rohit Kundu, Vishal Mohanty, Hao Xiong, Shan Jia, Athula Balachandran, and Amit K. Roy-Chowdhury. SAGA: Source attribution of generative AI videos. *arXiv preprint arXiv:2511.12834*, 2025.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10:1096, 2019. arXiv:1902.10178.

Shuo Liang, Yixing Ma, Pengfei Zhou, Zhenglin Wan, Xingyan Chen, Zihan Mei, Manting Li, Feihan Chen, Zhiwen Wang, Bin Xu, Haotian Zhang, Jiajun Song, Shiya Su, Run Liu, Zhenghang Ni, Yifa Yu, Jintao Hong, Bolong Feng, Yifei Liu, Zirui Zhang, Jingxuan Zhang, Songlin Zhao, Yifan Bai, Kang Tan, Yizhe Liu, Junhao Du, Yongtao Ge, Zhaopan Xv, Xinyuan Zhang, Mengru Ma, Chunhua Shen, Wei Wang, Yang You, Zheng Zhu, Kaipeng Zhang, and Wangbo Zhao. Can we defend against AI-generated video attacks on real-world crisis events? A systematic evaluation of detectors, generators and social dissemination. *arXiv preprint arXiv:2608.14391*, 2026. Introduces RA-Bench.

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

Qingyuan Liu, Pengyuan Shi, Yun-Yun Tsai, Chengzhi Mao, and Junfeng Yang. Turns out I'm not real: Towards robust detection of AI-generated videos. *arXiv preprint arXiv:2406.09601*, 2024. Introduces DIVID.

Lianghui Ma et al. DeCoF: Detecting ai-generated video via frame consistency. *arXiv preprint arXiv:2402.02085*, 2024. ICME 2025; introduces the GVF benchmark.

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

Michels, Jorissen, and Michiels. Dataset biases and shortcut learning in motion-based AI-generated video detection. *arXiv preprint arXiv:2607.00948*, 2026.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. OpenVid-1M: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024. ICLR 2025; source of the GenBuster real videos.

Nguyen, Radoi, Hermary, Astrid, Mejri, Ghorbel, and Aouada. When AUC misleads: Polarization-aware evaluation of deepfake detectors under domain shift. *arXiv preprint arXiv:2606.19184*, 2026.

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. *arXiv preprint arXiv:2501.11340*, 2025. AAAI 2026.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. arXiv:2302.10174.

Foster Provost, Tom Fawcett, and Ron Kohavi. The case against accuracy estimation for comparing induction algorithms. In *International Conference on Machine Learning (ICML)*, pp. 445–453, 1998.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2304.15004; Outstanding Paper Award.

Xiufeng Song, Xiao Guo, Jiache Zhang, Qirui Li, Lei Bai, Xiaoming Liu, Guangtao Zhai, and Xiaohong Liu. On learning multi-modal forgery representation for diffusion generated video detection. *arXiv preprint arXiv:2410.23623*, 2024. NeurIPS 2024 Datasets and Benchmarks; introduces DVF.

Thakur, Li, and Stamp. Robustness of AI-art detectors under generator shift. *arXiv preprint arXiv:2608.11643*, 2026.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Antonio Torralba and Alexei A. Efros. Unbiased look at the dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1521–1528, 2011.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-generated images are surprisingly easy to spot... for now. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. arXiv:1912.11070.

Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv preprint arXiv:2305.10874*, 2023a. Introduces the HD-VG-130M dataset; GenVidBench Pair2 real corpus.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22445–22455, 2023b. arXiv:2303.09295.

Haiquan Wen, Yiwei He, Zhenglin Huang, Tianxiao Li, Zihan Yu, Xingru Huang, Lu Qi, Baoyuan Wu, Xiangtai Li, and Guangliang Cheng. BusterX: MLLM-powered ai-generated video forgery detection and explanation. *arXiv preprint arXiv:2505.12620*, 2025. Introduces the GenBuster-200K dataset and GenBuster-Bench.

Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. arXiv:2406.06040; GenVidBench Pair1 real corpus.

Chende Zheng, Ruiqi Suo, Chenhao Lin, Zhengyu Zhao, Le Yang, Shuai Liu, Minghui Yang, Cong Wang, and Chao Shen. D3: Training-free AI-generated video detection using second-order features. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. arXiv:2508.00701.

AI Assistance Disclosure

In accordance with the ICLR policy on LLM use, we state the extent of AI assistance in this work in full.

Scope of AI use. AI coding and analysis agents (Claude Code, running GLM and Kimi models; Codex, running GPT models) were used, under author direction, to implement the training and evaluation code, execute experiments, process and audit data, interpret results, and draft the manuscript. No video in the study was synthesized by these agents: every generated clip was produced by a named released or commercial video generator, driven by our experiment scripts. The same agents were also used for retrieval and discovery of related work, surfacing candidate papers; the author read and selected every source the paper cites.

Human contribution. The author defined the research direction, designed or vetoed key experiments, and set the standards for verification and reporting, and takes full responsibility for the work.

Verification and responsibility. Every quantitative claim in this paper is recomputed at compile time from committed artifacts by the verification gate (the Reproducibility Statement below gives the mechanics). The author takes full responsibility for the results and for all content of this submission, including text, claims, and artifacts produced with AI aid. All experiments are reproducible from the released materials: the code, the per-video prediction files, and every artifact needed to re-derive each printed number are attached as supplementary material, and the training pipelines are documented in the released runbook.

REPRODUCIBILITY STATEMENT

Every quantitative claim in this paper is recomputed at compile time from committed result artifacts by a verification gate: each number is written inline with the JSON path and key that produces it, and the build fails on any mismatch, so no digit in the PDF can drift from its source (the gate's counts are printed in the verification appendix from a machine-written include the gate itself regenerates each run; hand-written sources carry no self-referential counts). The released repository carries the per-video prediction or measurement files behind every table row, so each reported AUC, detection rate, prevalence, and ratio can be recomputed from raw scores without re-running any model: an audited automated classification of every inline claim finds all 1031 unique numbers fully recomputable from committed per-video files (the audit tooling and its dependency maps ship with the repository). This is numeric re-computability, deliberately weaker than full rerun reproducibility: re-executing a training pipeline is not bit-exact for every family (seed handling is declared per trainer), so a rerun bounds variance rather than reproducing digits, and the gate guarantees only the former. Training and evaluation pipelines are documented in the released runbook (`RUN.md`), and the exact environment the paper was built in is pinned in `requirements-paper.txt`. Artifacts that depend on benchmark redistribution terms (benchmark video files and model checkpoints above a size threshold) are re-derivable from the committed manifests and scripts; the provenance of every external artifact is documented in the release. Code and artifacts are released with the repository upon publication.

A VERIFICATION APPENDIX

Every numeric claim in this paper is written inline as `\Cnum{printed}{repo artifact:key}` (ours) or `\Cext{printed}{registry key}` (external, checked against a hand-transcribed registry carrying page, table, and URL). At build time `scripts/paper/verify_numbers.py` recomputes each from the committed artifact and fails the build on any mismatch (no tolerance branch: the printed digits must be exactly a rounding of the computed value). The gate's guarantee is numeric re-computability from committed artifacts, not full rerun reproducibility of training runs, whose single-seed schedules and declared seed handling it does not bind. The build that produced this PDF verified 1820 claims with 0 failures and 0 warnings; the per-file counts are Table 3 (per source file; most abstract and introduction numbers print through shared macros defined in later sections, so they count there, not at the use site). The command is `python3 scripts/paper/verify_numbers.py --strict`, followed by `make compile`; both are wired into the repository Makefile's `verify` and `compile` targets. Because this appendix quotes the gate's own accounting, its counts are not hand-written: they are generated into a machine-written include (`generated/verified_macros.tex`) by the same gate run that writes the ledger artifact, and every run regenerates them; the gate never edits hand-written sources, and any drift between the two fails the build with an instruction to rerun the gate.

B THE SIGNAL-FAMILY SURVEY LEDGER

The survey verdict of Section 3.2 is a count over signal families, and a count that cannot be inspected is exactly the kind of claim this paper exists to audit. Table 4 is therefore the full ledger: every family under the survey's scope rules, its representative features, its status, and the one-line verdict that falsified it. The counts the body quotes are recomputed from this ledger by the same build-time gate as every other number in the paper: of the 19 families entering the survey, 17 were tested, 1 remains queued (feature-series structure and ballistic-trajectory fits), and 1 is data-blocked rather than excluded post hoc (audio-visual synchrony: no paired-audio real pool locally). Of the seventeen

Source file	Verified claims
00_main	153
08_conclusion	0
09_appendix_verification	0
10_appendix_survey	12
11_ai_disclosure	1
12_appendix_moved	1644
verified_macros (this table)	10

Table 3: Per-file counts of provenance-verified numeric claims in the build that produced this PDF, as reported by the verification gate itself; the counts are generated into `generated/verified_macros.tex` from the ledger artifact by the same gate run and are not hand-written.

tested, 0 survive every applicable gate in every domain: 7 were falsified by cross-generation sign inversion (g3), 2 by motion-amount matching (g2), 2 by zero-shot or controls retest (g4), 1 by its structural null (g5), 2 never carried signal at bank magnitude, 2 are in-domain discriminants that never made a cross-generation claim, and 1 (the deep RGB backbone) decays without inverting. The fence of Section 3.2 applies to the whole table.

C Extended Detail Relocated from the Main Text

This appendix holds analyses whose full detail the main text compresses for the page limit. Nothing here is new; each subsection is the complete form of a result the main text states in residual form, and every number keeps its inline provenance and build-time verification.

C.1 Metric-dependence of system ordering (extended)

This subsection is the complete form of the PM-A-versus-PM-B tri-metric comparison that Section 4.4 summarizes: the metric lesson of Section 2.2 arriving one layer down, into a fully matched comparison.

Under the official Mean Top-1, PM-A leads by $+5.79$ points (paired video-level cluster bootstrap 95% CI $[5.49, 6.08]$ over the five Pair2 columns, $n = 67822$ paired videos). Re-scoring the same committed scores under two prevalence-free views of the same benchmark reverses that ordering. Under balanced accuracy the delta shrinks to $+2.48$ (CI $[2.24, 2.73]$); under pooled AUC it reverses to -2.16 (CI $[-2.59, -1.68]$: PM-B wins the pooled ranking on identical predictions). "Balanced" itself has two conventions here, and we print both rather than pick one silently: balancing classes (equal mean over the four fake columns' recalls and the real column) gives $+2.48$; balancing columns one-to-one (the same five per-column accuracies the official metric averages, equally weighted) gives $+5.79$ (CI $[5.49, 6.08]$). The two conventions change the margin (+2.48 versus +5.79 points) though not the sign here, which is its own lesson: metric definitions remain load-bearing even after the 4:1 column pathology is removed. And the official margin is carried by a single column: the per-column deltas are $+52.43$ on Mora, $+12.30$ on MuseV, and $+2.91$ on SVD, against *negative* -35.68 on CogVideo and -3.03 on HD-VG; excluding Mora alone flips the official mean to -5.87 (CI $[-6.16, -5.59]$). The official metric rewards exactly the column where PM-A wins; the ordering on GenBuster's clean splits, where the comparison is interpretable, is stated in Section 4.2.

C.2 The instrument's self-validation record

This is the validation-set table of Section 3.4: the five claims of ours the battery and its controls falsified, each with the instrument that falsified it and the honest number that replaced it. The fourth row shows the gates localizing a failure mechanistically, identifying the channel a claimed number was riding.

Family (representative features)	Status	Verdict
A. Motion-magnitude prior (te_mean, mag_mean)	tested	direction flips across generator generations; bank member per domain
B. Motion smoothness, jerk (j_mean/std/max)	tested	motion-amount prior in disguise; dies under matched control
C. Flow invariants (div/curl statistics)	tested	scale-coupled; matched control reversed; variance members bank per domain
D. Depth-motion consistency (P2 LSQ residual)	tested	zero-shot transfer reversed cross-generator
E. Temporal periodicity (lag-k autocorrelation)	tested	joint denoising erases the periodicity; below the bank bar
F. Temporal band-limitation (HF_ratio)	tested	headline zero-shot was an fps/resolution confound; clean retest at chance
G. Token-grid noise (token_contrast)	tested	grid-specificity and phase nulls fail; symmetric-30fps court sign-flips it
H. Global motion coherence (affine explained energy)	tested	fakes more coherent than reals on 2024 generators; reverses on Seedance
I. Flow multimodality (dir_entropy, bimodality)	tested	no signal under motion-matched control
J. First-transition priors (frame0_zscore)	tested	real-camera onset prior, in-domain only; never a cross-generation claim
K. Spatial-spectral classics (noise residual, HF energy)	tested	compression-robust but direction flips across generators; resolution equalization flips Seedance
L. Latent-probe, VAE recon (v5/v6 latent branch)	tested	harmful as co-training; in-domain scope only
M. RGB appearance, deep (backbone logits)	tested	decays across generations without inverting; carries the surviving rows
N. Audio-visual synchrony (onset correlation)	blocked	no paired-audio real pool locally; blocker recorded, not excluded post hoc
O. FB flow-cycle error (fb_err_mean/std)	tested	fakes lower cycle error on 2024 generators, higher on Seedance
P. SSIM flicker (ssim_std/fluct)	tested	same-sign in three domains but fails under motion matching on Wan
Q. VFI interpolability (vfi_norm_mean/std)	tested	holds for 2024 CogVideoX, reverses on Wan, null on Seedance
R. Static-region noise floor (static_noise_mean/std)	tested	fakes too clean on CogVideoX, noisier on Seedance; strong per-domain, flips direction
S. Feature-series / ballistic (DINO series, trajectory fits)	queued	not yet tested; counted as queued, not as falsified

Table 4: The signal-family survey ledger: every family under the survey's scope rules (computable from decoded frames with frozen estimators), its representative features, its status, and the one-line verdict that falsified it. The counts quoted in the paragraph above are gate-verified at build time against the committed per-family summary artifact. Bank-level membership is a separate, weaker verdict (a feature clearing $|\mathrm{AUC} - 0.5| \geq 0.10$ in any single domain after its applicable gates): several rows above (A, C, R) are dead as universal claims and alive as per-domain bank members, which is the two-level reading of Section 3.2. Per-family provenance: each row's primary artifact is recorded in the committed ledger script and existence-checked at summary-generation time.

C.3 Second-real-pool replication

This subsection is the complete form of the real-draw sensitivity check summarized in Section 4.2. The real half of the mini split is one 4,800-video draw; swapping it for a disjoint, container-verified draw of identical size from the benchmark's full real pool moves nothing that matters: PM-A 0.8645 to 0.8603, v7e1 0.7791 to 0.7701, PM-C 0.7235 to 0.7154, and PM-B least of all; every shift under one AUC point, the system ranking invariant, and only one of the four paired intervals excluding zero (v7e1, by -0.0005). One systematic wrinkle is honest to keep: all 16 system-by-generator cells shift negative (sign-test $p = 1/65536$), so an independent draw from the same corpus is very slightly harder, for every system and every generator, a real-draw effect an order of magnitude smaller than every gap Section 4 rests on, and bounded as such. Both real draws come from the same upstream corpus (OpenVid-1M via the benchmark's uniform re-encode), so the swap answers

Claim (as first made)	Instr.	What it found	Honest number
Band-limited temporal spectrum transfers zero-shot (0.907)	g4, g6	the seedance real pool was an fps/resolution staircase; on the clean matched-fps domain the feature does not separate	0.489
5-class generator attribution 0.990	leak audit	the NaN pattern of the feature cache carried the class (mask alone 0.893); after re-extracting every missing cell from video at source	0.889 (mask-only 0.096; CogVideoX-vs-Wan-2.2 at uniform container 0.98)
Rank fusion beats VideoMAE on all three domains (0.9917)	protocol audit	the passing cell was a 7%-coverage subset swapped in after the pre-registered full-set run failed (full-set delta CI $[-0.0074, -0.0013]$, significantly worse); the 0.9944 cell was in-domain supervised adaptation, cross-domain 0.964, -0.0121 vs. the bare backbone	retracted; at matched 10-label access comparable, CI $[-0.027, 0.069]$
Explicit physics features survive zero-shot where learned features die	g2	on the MuseV column the bank's 0.776 inverted to 0.400 under motion matching (n=1653, 9 buckets; 8k-subset checkpoint; the deployed full-data checkpoint of the same bank runs 0.766 to 0.4999, same verdict), and its best single feature collapsed (0.698 to 0.118); "physics survives" was, exactly inverted, "the container channel survives in the only pipeline that observes it"	no structural residue
Coherence-branch co-training drives the detection gain (0.9651 AUC on GenVideo-Val)	control audit	replacing the coherence auxiliary with a rotation-prediction auxiliary (no physics content, identical protocol and split) scores higher; the gain is generic auxiliary-task regularization, not coherence	0.9733 [0.9714, 0.9754] against the claim's [0.9621, 0.9682], the marginal intervals disjoint over the shared n=18286 videos (schedule asymmetry declared: control schedule ends at epoch 3, coherence peak is epoch 3 of 5)

Table 5: The instrument's validation set: five claims of ours the battery and its controls falsified. Each row is a committed-before-evaluation claim, the instrument that falsified it, and the honest number that replaced it. Rows 2 and 3 were falsified by audit modes rather than gates: no gate in the battery of Section 3.1 covers cache-missingness leakage or pre-registration/label-access accounting, so we label them as what they are instead of dressing process audits as gate falsifications. Row 5 was falsified by a matched-control audit (an auxiliary-task substitution under an otherwise identical protocol) that predates the battery's formalization; it is included under the same discipline, with its declared schedule asymmetry, and no paired test is claimed (the two marginal intervals are disjoint). The fourth row's mechanism is the one Section 2.1 documents at the benchmark level: the pipeline that samples 8 frames in a clamped window is the only one that can see the duration field, so it is the only one that inherits the field.

sampling sensitivity, not capture diversity; the cross-corpus question stays open and is stated as a limitation.

C.4 THE 2026 WALL: WILD-TRACK PER-GENERATOR DETAIL

This subsection is the complete form of the wild-track localization that Section 4.3 summarizes and Table 6 tabulates. The per-generation pools' container gate, in full: every rule scores AUC 0.5 on both generations. Seedance-2.0 is the weakest shared cell for the two strongest systems in the suite (it is not the weakest cell for the other two: PM-B's lowest cells are Wan-2.6 and Sora-2 at 0.461, and PM-C's is Veo-3.1 at 0.259). On its 15 Seedance-2.0 videos PM-A scores reference AUC 0.401 (CI [0.227, 0.572], including 0.5), v7e1 0.548, and PM-B 0.571; PM-C's Seedance-2.0 cell is the one whose interval excludes 0.5 at this cell size: 0.279 (CI [0.145, 0.417]). No single generator carries the pooled ranking: recomputing the pooled wild AUC with Seedance-2.0 excluded moves the two strongest systems by about three points and the weaker systems by less: PM-A $+3.3$ (0.702 to 0.7354), v7e1 $+3.5$ (0.86 to 0.8944), PM-C $+2.1$, PM-B $+0.6$. The 12–16-point gap between v7e1 and the rest persists across the remaining ten generators, and the decay of the weaker systems is broad-based, not concentrated: v7e1 stays at or above 0.79 on every generator except Seedance-2.0, while PM-A holds seven generators below 0.77. On the wild mix itself the clean-room DeMamba of Section 4.3 detects 77.3% of the pool at the fixed threshold while ranking it at the 0.748 of the main text. The neighboring weak points for PM-A are Veo-3.1 (0.584, CI [0.427, 0.762]) and Sora-2 (0.555, CI [0.396, 0.693]); 3 of its eleven wild generators have point estimates below 0.6, and 2 of those three have intervals spanning it, so the count separates weak cells from fine ones only weakly at $n = 10$–15. Wan-2.5 is the clean exception, where the deep backbones hold (0.855 and 0.952). The decay is the one our local measurements had already documented, on a different model and reference class than the rows above: our locally evaluated transfer numbers put the v4 release backbone at 0.780 on local Seedance-2.0 and 0.731 on local Seedance-2.5, against 0.964 in-domain on CogVideoX; the external 2026-era wild numbers sit below even those decaying local values, in the direction the survey's direction-shift trend had documented on independent data. One further frontier-family cell is exploratory and labeled as such (post hoc, in no preregistered family): 200 clips of a 2026 open-weight frontier generator (MiniMax H3, local ComfyUI generation; 75 prompts, full-decode container audit clean), scored zero-shot in the container-matched condition, every clip re-encoded to the GenBuster container and read against the same id-reals reference. The pool reads in the wall's direction, every interval excluding chance: the retrained TSM at 0.245 [0.214, 0.275], the clean-room DeMamba at 0.225 [0.200, 0.250], and the physics backbone at 0.322 [0.288, 0.354], while the two container-standardized readers hold at 0.892 and 0.930; the GenVideo backbone, inverted at the Seedance wall, here resolves just above chance (0.563), another instance of its reference-pairing sensitivity. H3 and Seedance are both 2026 frontier-generation models, so the replication is consistent with a frontier-generation pattern, a one-cell observation and not a claim about the generation at large. Per-video scores and the pool manifest are committed under `results/exploratory/`.

C.5 THE OPEN-WEIGHT GENERATOR SWEEP

This subsection is the complete form of the corpus-not-era quantification that Section 4.3 states in residual form: four further single-generator pools of 500 clips each, from three open-weight generators released after every audited detector's training data (LongCat-Video, contributing a matched and a demonstration pool; Helios-Distilled, demonstration; LTX-2.5, matched; matched pools share their fixed prompt set with a controlled real collection, demonstration pools use the generator's published demonstration prompts). The generator-native containers differ by design; every clip was re-encoded to the wall container, after which every container axis is constant on both sides and the post-encode standard check reported 0 mismatches on all 2000 clips, so the zero-feature rules are degenerate by construction here. The seven detectors of the wall suite plus the corpus-standardized DeMamba retrain scored every pool zero-shot against the same committed reference real scores the wall protocol used (same checkpoints, same pipeline, declared reuse), and every cell carries a stratified bootstrap interval plus a label-permutation test for deviation from chance. The evaluation's preregistration record survives only as the protocol documentation embedded in the released evaluation script, together with its ambiguity log of implementation decisions the protocol text left open; we cite it in that status rather than as a standalone document. One recorded ambiguity is material and stated here: the preregistration text named an eight-detector family while declaring seven detectors, and was resolved as the 28 committed cells with no eighth detector invented.

The multiplicity statement behind the residual text: 25 of the 28 cells deviate from chance after Benjamini–Hochberg control at q=0.05 (raw and corrected calls coincide), and the family-level guard

System	n	Seedance-2.0	Seedance-2.5 (n=183)	wild pooled (n=150)
PM-A (physics backbone)	995	0.153 [0.137, 0.169]	0.154 [0.127, 0.182]	0.702
v7e1 (GenVideo backbone)	995	0.453 [0.428, 0.479]	0.472 [0.426, 0.519]	0.860
PM-C (motion-stat expert)	995	0.348 [0.324, 0.373]	0.358 [0.312, 0.405]	n/a
PM-B (plain baseline)	995	0.588 [0.561, 0.615]	0.574 [0.544, 0.605]	n/a
DeMamba (clean-room, Pair1)	895	0.192 [0.172, 0.214]	0.205 [0.182, 0.231]	0.748 [0.701, 0.789]
TSM (retrained, Sec. 4.1)	895	0.1386 [0.1223, 0.1554]	0.1123 [0.0844, 0.1421]	n/a
TSM (retrained, seed 43)	895	0.129 [0.112, 0.146]	0.086 [0.0614, 0.1138]	n/a
DeMamba (clean-room, seed 43)	895	0.256 [0.232, 0.280]	0.276 [0.250, 0.303]	n/a
DeMamba (GenBuster-200K corpus)	895	0.928 [0.916, 0.939]	0.976 [0.961, 0.988]	0.876 [0.841, 0.911]
VideoMAE-GB anchor (clean anchor, seed 43)	995	0.856 [0.839, 0.873]	0.886 [0.858, 0.913]	n/a
WaveRep G1 (published weights)	995	0.369 [0.346, 0.393]	0.722 [0.682, 0.761]	0.825 [0.788, 0.861]
WaveRep G4 (published weights)	995	0.547 [0.520, 0.572]	0.801 [0.765, 0.835]	0.899 [0.871, 0.924]
PM-B (container-normalized training)	995	0.056 [0.047, 0.066]	0.016 [0.0095, 0.0239]	n/a
PM-B (container-normalized, seed 43)	995	0.070	0.022	n/a

Table 6: The 2026 wall at per-generation scale: reference AUC against the id reals after container normalization to the GenBuster standard (container rules return exactly chance, as stated in the text). The first two columns are the per-generation wall pools, single-generator and container-normalized; the third is the GenBuster wild mix, a different and heterogeneous pool: eleven families pooled, including the 15 Seedance-2.0 clips, and it is the pooled reading the per-generation rows decompose. Claims about the Seedance generations refer to the wall columns, not the wild mix. The first four rows are our suite (single-seed); DeMamba clean-room and the retrained TSM of Section 4.1 are external architectures trained on GenVidBench Pair1 and scored zero-shot under the identical wall protocol (divergences declared in the released report of that run); the GenBuster-200K DeMamba row is the same architecture and recipe retrained on the container-standardized corpus, the anchor row is the clean anchor's seed-43 replicate, the two seed-43 rows replicate the inverted TSM and DeMamba lineages at a second training seed (identical recipe and pools; the trainers of these rows and of the clean anchor reseed the global generators, so the seed moves initialization and data order together, while the physics backbone's trainer does not set the global seed, so its replicate reflects an additional independent initialization draw and same-seed reruns of that family are not bit-reproducible; every replicate measures full run-to-run variance, which is the quantity a robustness claim wants), the WaveRep rows are published weights zero-shot, and the last row is the PM-B recipe retrained with only its training container normalized. The n column prints each row's Seedance-2.0 fake count, verified per row against its own artifact: the suite, anchor, container-normalized retrain, and WaveRep rows score all 995 decoded clips, while the retrained external lineages (the clean-room DeMamba and retrained TSM at both seeds, and the GenBuster-200K-corpus DeMamba) score the 895-clip s2test subpool, the fine-tune study's held-out split; Seedance-2.5 is the same 183-clip pool for every row. The bitrate-matched control re-scores the PM-A and PM-C rows on the Seedance-2.0 pool only (both move by at most two thousandths); the corpus-not-era detail is Appendix D.5; the per-generator wild detail is the surrounding subsection.

is a max-statistic permutation null: within-cell label permutation with fixed margins, the statistic the largest $|\mathrm{AUC} - 0.5|$ across cells, over 2000 permutations (seed 42). The null's largest deviation is 0.0731 (95th percentile 0.0485) against the strongest observed 0.4824; the null is conservative under the shared reference reals' positive cross-cell dependence, and every below-chance cell clears it.

The frozen bank's calibration ladder on the same pools (the wall-ladder procedure verbatim: per-feature signed directions from k clips plus k reals, 20 draws) shows the 0.72 ceiling of Section 4.3 is a pool property rather than a universal bound. On the matched LongCat pool the ladder sits at the Seedance-2.0 level at every budget: default directions 0.7245, k=20 0.7188 [0.6096, 0.7539], k=100 0.7261 [0.7177, 0.7383]. The other pools read higher: LTX-m 0.7569 at k=20, LongCat-s 0.8983, and Helios-s 0.9096 [0.9026, 0.9122]; on Helios-s the default directions anti-detect (0.1333 at k=0) and twenty labels restore the same 0.91 the pool's calibrated ceiling sits at, the direction-flip risk quantified in Appendix D.7 in action rather than a new phenomenon; at k=100 Helios-s reads 0.911.

Reference-pool robustness, all cells. The standardized band, in full: 0.903 to 0.9935 across anchor, seed, and corpus-retrain cells. Re-scoring the same committed predictions against the disjoint second real draw (n=4800 reals, subsampled to 1,000; the pool of Appendix C.3) moves no conclusion: the standardized side reads between 0.9026 (anchor, GenBuster-trained, seed 43, Seedance-2.0 face) and 0.9935 (anchor, Helios-s face), the confounded side stays inverted or weak (PM-A 0.2445 [0.2244,

System (all zero-shot, fixed 0.5)	LongCat-m	LTX-m	LongCat-s	Helios-s
PM-A (Pair1)	0.282 [0.257, 0.309]	0.264 [0.239, 0.289]	0.056 [0.045, 0.069]	0.215 [0.194, 0.238]
PM-B (plain baseline)	0.585 [0.557, 0.611]	0.590 [0.562, 0.618]	0.524 [0.494, 0.554]	0.663 [0.635, 0.689]
PM-C (motion-stat expert)	0.559 [0.529, 0.589]	0.678 [0.651, 0.705]	0.654 [0.626, 0.682]	0.018 [0.011, 0.026]
v7e1 (GenVideo backbone)	0.522 [0.486, 0.553]	0.526 [0.495, 0.557]	0.157 [0.131, 0.183]	0.604 [0.574, 0.634]
DeMamba (clean-room, Pair1)	0.251 [0.227, 0.275]	0.210 [0.188, 0.233]	0.158 [0.140, 0.178]	0.272 [0.248, 0.297]
TSM (retrained)	0.209 [0.187, 0.232]	0.141 [0.123, 0.161]	0.218 [0.195, 0.241]	0.036 [0.028, 0.045]
VideoMAE-GB anchor (GenBuster-200K)	0.9495 [0.939, 0.959]	0.9438 [0.933, 0.954]	0.9366 [0.923, 0.949]	0.9788 [0.973, 0.984]
DeMamba (GenBuster-200K corpus)	0.9648 [0.956, 0.972]	0.9788 [0.973, 0.984]	0.9665 [0.959, 0.974]	0.9963 [0.994, 0.998]

Table 7: The open-weight generator sweep: reference AUC against the id reals after container normalization, per pool (pool suffix -m = matched prompt set, -s = the generator's demonstration prompt set; 500 fakes each against 1,000 reference reals; stratified bootstrap 2000/seed 42; one-sided label permutation in the observed direction, 2000/seed 42, per cell). The top block is the confounded-corpus suite of Sections 4.2 and 4.3; the bottom block is the two container-standardized architectures. Three top-block cells sit within 0.026 of chance (PM-B on LongCat-s, v7e1 on LongCat-m and LTX-m); every other cell deviates from chance in its observed direction at family-controlled q=0.05, and every inverted cell's interval lies entirely below 0.5.

0.2657] on Seedance-2.0 and 0.2448 on 2.5; PM-B 0.5725 and 0.5593; v7e1 0.3529 [0.328, 0.3783] and 0.3764; clean-room DeMamba 0.2785 [0.2534, 0.3048] on LongCat-m; the retrained TSM 0.1453 [0.1284, 0.1638] on Seedance-2.0, 0.1182 [0.0902, 0.149] on 2.5, 0.2189 [0.1957, 0.2419] on LongCat-m, and 0.0395 [0.0307, 0.0487] on Helios-s), while the standardized cells hold theirs (anchor seed 42: 0.9452 [0.936, 0.9541] and 0.9615; seed 43: 0.9026 [0.8889, 0.9162] and 0.9273; sweep faces: 0.9796 [0.9741, 0.9847] and 0.9935; corpus-standardized DeMamba 0.9481 [0.938, 0.9575] and 0.9819 [0.969, 0.9924]). The swap stays within the OpenVid real family, so it answers sampling robustness, not reference-source robustness; the independent-family boundary itself is measured by the Vript re-scoring that follows.

Independent-family reference: the Vript re-scoring. A preregistration committed before the run (`docs/prereg/vript_referenced_wall.md`) fixed the decision rule: a lineage replicates iff its Vript interval excludes 0.5 on the side of its committed cell, a flip whose interval excludes the committed sign licenses a per-lineage qualifier in this appendix, and the full matrix is reported either way, no selective reporting. The reference is a Vript draw of 300 reals, a collection lineage different from OpenVid and HD-VG (the seed-42 alphabetical draw already committed for the flow-magnitude replication), each detector scoring its own reference under the wall convention, primary variant gbstd (the wall reference convention; the native container scored as a declared sensitivity arm; per the preregistration's scope guard, magnitudes across references are not interpreted). Convention parity was verified first: all fourteen committed wall cells recompute from the committed prediction files to under two thousandths. The preregistered prediction, that the sign pattern replicates, fails, and the five licensed qualifier cells follow, each a lineage that does not replicate against the independent-family reference. PM-A on Seedance-2.0: committed 0.153 below chance, Vript 0.5425 [0.5085, 0.5787] above, the interval excluding the committed sign. The GenVideo backbone on the same pool: 0.4527 below to 0.6811 [0.6463, 0.716] above. The retrained TSM on the same pool: 0.1386 to 0.5723 [0.538, 0.6078]. The clean-room DeMamba: 0.192 to 0.6942 [0.6597, 0.7309] on Seedance-2.0 and 0.2053 to 0.7184 [0.6705, 0.7645] on 2.5, the qualifier holding on both cells. Four cells replicate in sign: PM-B on Seedance-2.0 at 0.5976, the clean anchor at 0.6034 and 0.6912, and the corpus-retrained DeMamba on 2.5 at 0.7886, the one holder above 0.7. Five cells do not resolve, their intervals spanning 0.5: PM-A on 2.5, PM-B on 2.5, the retrained TSM on 2.5, the corpus-retrained DeMamba on 2.0, and the GenVideo backbone on 2.5, whose committed cell was itself ambiguous, so the rule does not apply there. Read beside the in-family swap above: every inversion survives a second OpenVid draw and washes out against Vript, so each wall inversion is a property of the detector-plus-reference pairing, the reference family and the container lineage it shares, rather than of the detector alone, while the holding signs persist at compressed magnitudes. The full matrix, the per-cell decision ledger, and the native sensitivity arm: `results/paper/vript_referenced_wall.json`.

C.6 THE MULTI-REFERENCE, MULTI-SEED WALL RETEST, IN FULL

This subsection is the complete form of the retest that Section 4.3 states in residual form. The preregistration (`docs/prereg/multiref_multiseed_wall.md`, frozen 2026-09-25 before any new scoring or training launch) fixed the panel, pools, statistics, and decision rule; nothing below was selected after the fact. The panel: the two corpus-standardized systems, the GenBuster-200K-corpus DeMamba at seeds 42, 43, and 44, and the VideoMAE-GB anchor at seeds 42 and 43 (the anchor's declared two-seed limit; the DeMamba's third seed is governed by the completion clause below, which it satisfied). The pools: the fixed 895-clip Seedance-2.0 test split and the 183-clip Seedance-2.5 pool, joined exactly by clip id against the committed wall predictions (every system-seed join exact on both pools; decode failures zero on every new scoring arm). The reference families: the in-family OpenVid bench-id reals (the wall's own reference, 1000) and the HD-VG and Vript gbstd draws of 500 already gated by the 9/24 cross-source preregistration. The statistics follow that preregistration's frozen conventions: reference AUC per cell with stratified video-level bootstrap intervals (2,000 resamples, seed 42), two-sided label permutation (5,000, seed 43), and Benjamini–Hochberg over the independent cells. The decision rule, verbatim: a family robust-holds iff every tested seed has ci95_lo > 0.7; a family collapses iff a majority of tested seeds has ci95_lo ≤ 0.7; verdict (a), at least two of the three families robust-hold; verdict (b), OpenVid robust-holds and at least one independent family collapses, the paper-rewrite clause; verdict (c), anything else; all cells committed under every outcome, no re-rolls, no seed substitution, no pool edits after the freeze.

The verdicts, final with seed 44 appended under the completion clause: the corpus-retrained DeMamba is (b) on Seedance-2.0, OpenVid robust-holding at all three seeds while both independent families collapse, HD-VG below chance at all three (permutation p 0.0020, 0.0004, and 0.0000, the 5,000-draw floor; BH q 0.0022, 0.0005, and 0.0000; every interval entirely below 0.5) and Vript collapsing with seeds 42 and 43 at chance (0.4981 and 0.5144, permutation p 0.9088 and 0.3736, intervals spanning 0.5) while seed 44 drops below chance (0.419, permutation p 0.0000, interval entirely below 0.5: the collapse bar met by chance readings at two seeds and an inversion at the third); on Seedance-2.5 the same system is (a), OpenVid and Vript robust-holding at all three seeds and HD-VG above chance at all three while under the 0.7 bar at seed 42, the one family the rule classes neither robust nor collapsed. The anchor is (b) on both generations at its two declared seeds: OpenVid robust-holds while both independent families sit in the 0.56–0.69 band at every seed, above chance and below the robustness bar. The third DeMamba seed independently reproduces both of its verdicts, and on the (b) cell it reads deeper than either earlier seed: HD-VG 0.3802 against the snapshot's 0.4478 and 0.4324, with Vript below chance where the snapshot sat at it. Read against Table 2: the seed-42 column reproduces the committed cross-source cells exactly, the seed-43 column is the replication the single-seed reading lacked, and seed 44 is the preregistration's own out-of-sample check; the HD-VG collapse, the Vript non-holding, and the OpenVid holding all repeat. What the grid adds: the holding side's collapse is not a seed-42 accident, the Seedance-2.5 holding is genuinely cross-reference (openvid and Vript robust together at three seeds, HD-VG above chance at all three), and no tested recipe is reference-stable on Seedance-2.0 under the frozen bar.

The completion clauses, executed. Seed 44 (DeMamba, byte-identical recipe) completed its 4-epoch checkpoint before the preregistered deadline and finished its wall evals before the submission freeze, so its cells are appended and the verdict recomputed over all three seeds, printed above, is final; it reproduces both DeMamba verdicts unchanged from the seeds-$\{42, 43\}$ morning snapshot, with every system-seed join exact and zero decode failures on its arms. The optional DAVIS arm, a fourth professional-footage reference family, was closed by its preregistered frame-count rule (DAVIS-2017 trainval Full-Resolution carries zero sequences with at least 121 frames, so the frame-count-matched gbstd pool is impossible without fabricating frames), declaring the three-family tested limit with no substitute pool sought. Deviations logged under the freeze, design untouched: seed 44's first launch OOMed on GPU0 (foreign memory resident, no checkpoint written) and was relaunched on GPU2 with dataloader workers 12 instead of 8, decode parallelism only; and a join-check bookkeeping fix (the anchor's 995-row Seedance-2.0 score files are subset to the 895 pool inside the seed-join check; cell computation was already correct).

C.7 THE MOTION-MATCHED CONSTRUCTION PILOT

This subsection is the complete form of the construction experiment that Section 4.3 states in residual form. The preregistration was committed before generation began with its readings fixed in advance,

System	Gen	Family	seed 42	seed 43	seed 44	median	all-seed int.	reading
DeMamba (GB-200K corpus)	S-2.0	OpenVid (in-family)	0.9282 [0.916, 0.9388]	0.8545 [0.8368, 0.8711]	0.8834 [0.8683, 0.8981]	0.8834	0.8545–0.9282	robust-hold
DeMamba (GB-200K corpus)	S-2.0	HD-VG	0.4478 [0.4164, 0.4809]	0.4324 [0.4028, 0.4647]	0.3802 [0.3495, 0.4123]	0.4324	0.3802–0.4478	collapse
DeMamba (GB-200K corpus)	S-2.0	Vript	0.4981 [0.4664, 0.5313]	0.5144 [0.4814, 0.5459]	0.419 [0.3876, 0.4511]	0.4981	0.419–0.5144	collapse
DeMamba (GB-200K corpus)	S-2.5	OpenVid (in-family)	0.9764 [0.9613, 0.9884]	0.9644 [0.9462, 0.9797]	0.9728 [0.9578, 0.9852]	0.9728	0.9644–0.9764	robust-hold
DeMamba (GB-200K corpus)	S-2.5	HD-VG	0.7409 [0.6963, 0.7838]	0.7756 [0.731, 0.8162]	0.7447 [0.7002, 0.7863]	0.7447	0.7409–0.7756	neither
DeMamba (GB-200K corpus)	S-2.5	Vript	0.7712 [0.7277, 0.8125]	0.8165 [0.7771, 0.8544]	0.7734 [0.7322, 0.8159]	0.7734	0.7712–0.8165	robust-hold
VideoMAE-GB anchor	S-2.0	OpenVid (in-family)	0.8898 [0.8756, 0.9042]	0.8578 [0.8405, 0.8742]	—	0.8738	0.8578–0.8898	robust-hold
VideoMAE-GB anchor	S-2.0	HD-VG	0.5647 [0.535, 0.5973]	0.6364 [0.6082, 0.6662]	—	0.6005	0.5647–0.6364	collapse
VideoMAE-GB anchor	S-2.0	Vript	0.6066 [0.5749, 0.6373]	0.6314 [0.6017, 0.6614]	—	0.619	0.6066–0.6314	collapse
VideoMAE-GB anchor	S-2.5	OpenVid (in-family)	0.9208 [0.8984, 0.9424]	0.8858 [0.857, 0.9128]	—	0.9033	0.8858–0.9208	robust-hold
VideoMAE-GB anchor	S-2.5	HD-VG	0.6502 [0.6032, 0.6958]	0.6782 [0.6301, 0.723]	—	0.6642	0.6502–0.6782	collapse
VideoMAE-GB anchor	S-2.5	Vript	0.6911 [0.6454, 0.7358]	0.6725 [0.6281, 0.7163]	—	0.6818	0.6725–0.6911	collapse

Table 8: The multiref multiseed grid: reference AUC with bootstrap intervals for both corpus-standardized systems at all preregistered seeds (three for the corpus-retrained DeMamba; two for the anchor, its declared limit, so its seed-44 column is empty), per generation and reference family; "reading" applies the frozen decision rule per family (robust-hold: every seed ci95_lo > 0.7; collapse: majority of seeds ci95_lo ≤ 0.7; Vript at Seedance-2.0 collapses by the rule, seeds 42 and 43 at chance and seed 44 below chance). The median and all-seed interval are the preregistration's per-family summaries over tested seeds. OpenVid is the in-family control (the wall's own reference); HD-VG and Vript are the independent families. Every value recomputes at build time from the committed artifact; per-video reference scores for the September-25 scoring arms under `results/multiref_multiseed_wall/preds/`.

and one amendment was adopted after an adversarial design review and before any evaluation-stage artifact existed; the amendment restricts the primary roster to the six detectors disjoint from the selection bank and demotes the two systems that share its features to a stress test flagged for circularity, excluded from the primary inference family and from the readings. The design: 300 source reals drawn from the GenBuster id pool; an open-weight image-to-video generator (LTX-2.5, the distilled variant) conditioned on each source's first frame under a neutral prompt, 4 candidates per source (1200 clips, seeds fixed by declared rule); selection by source-whitened Mahalanobis distance over the 17-feature motion bank, computed on the post-normalization files of both classes; both classes decoded and re-encoded through the identical wall-container recipe, the real side of every paired metric being each clip's own source through the same path, so the paired metric is codec-symmetric by construction; the unpaired reference a disjoint set of 700 id reals (path intersection with the sources empty, per-clip reference scores reused from the wall protocol). The container gate passes exactly: every single-feature rule reads AUC 0.5 on both classes. 5 of the 1,200 raw candidates carry container-valid headers over corrupt streams; the exclusion count, frozen before any detector ran, leaves 1195 normalized candidates with no post-hoc replacements, and every later stage is missingness-aware.

The pre-registered validity condition on matching fails, and the numbers are exact: the observed median selected distance is 2.701 against a randomized-assignment null median of 2.853 (null 95% interval [2.723, 2.984]), a reduction of 5.3% against the required quarter, and 15 of the 17 per-feature standardized deltas carry clustered intervals excluding zero (largest 0.58 standard deviations): at four candidates per source, selecting the minimum-distance candidate has almost no purchase on the motion distance, the motion constraint was not achieved at this k, the reading is downgraded by the preregistration's own condition, and the pilot licenses no statement that matching suppressed detection. The pilot's own did-matching-matter contrast is correspondingly null: the four per-candidate-slot pools and the selected pool are within noise for every detector (the detecting DeMamba spans 0.853 to 0.856 across slot pools against its 0.854 selected).

Table 9 is the full roster. Relocated verbatim from Section 4.3: selecting on a class-correlated feature can itself bias the pool, the risk the randomized-assignment null guards and this pilot fails; every confounded-corpus detector in the roster inverts (the retrained TSM 0.131, the clean-room DeMamba 0.147, the GenVideo backbone 0.298; the co-trained backbone 0.181 as a stress row flagged for circularity) while both GenBuster-trained systems detect (0.854 and 0.715). The pre-registered reading that holds is the intermediate split, downgraded: the confounded-corpus systems separate inverted (across the four pools the six span 0.018 [0.011, 0.026] to 0.678 [0.651, 0.705]), both GenBuster-trained systems separate in the detecting direction, and the baseline backbone sits at chance on the paired metric, the one cell where the weaker statement "no detection at this n" is licensed. Equivalence at the pre-registered margin (0.05 in AUC) is met by no primary detector on either metric, so the constructive-close reading, that corpus confounds fully explain detection, is not

Detector	refAUC [CI]	paired [CI]	det@0.5	re-enc.
TSM (retrained, Sec. 4.1)	0.1305 [0.109, 0.154]	0.060 [0.033, 0.090]	0.247	0.400^{i}
DeMamba (clean-room, Pair1)	0.1473 [0.124, 0.1715]	0.110 [0.077, 0.147]	0	0.537
v7e1 (GenVideo backbone)	0.2983 [0.265, 0.334]	0.200 [0.153, 0.247]	0.01	0.180^{i}
PM-B (plain baseline)	0.5695 [0.537, 0.603]	0.543 [0.487, 0.600]	0	0.643^{i}
DeMamba (GenBuster-200K corpus)	0.8539 [0.830, 0.876]	0.963 [0.940, 0.983]	0	0.645^{i}
VideoMAE-GB anchor (GenBuster-200K)	0.7150 [0.682, 0.746]	0.830 [0.790, 0.873]	0.057	0.353^{i}
PM-C (motion-stat expert; stress*)	0.5455 [0.510, 0.579]	0.540 [0.487, 0.597]	0.820	0.510
PM-A (physics backbone; stress*)	0.1812 [0.157, 0.208]	0.037 [0.017, 0.060]	0.007	0.857^{i}

Table 9: The motion-matched construction pilot, all cells: 300 first-frame-conditioned candidates (one per source after selection, content locked by construction, both classes through the identical container path) scored against a disjoint 700-real reference (refAUC) and against each clip's own re-encoded source (paired, the codec-symmetric metric); det@0.5 at the fixed threshold; re-enc. is the control side's P(score(re-encoded source) $>$ score(original)), whose interval excluding 0.5 (marked ^{i}) invalidates codec-lineage readings for that detector. All intervals and tests source-clustered (300 clusters), 2000 resamples or permutations, seed 42, two-sided; Benjamini–Hochberg over the twelve primary cells leaves every separating and inverted primary cell below $q = 0.0016$ and the baseline's paired metric unseparated, which is its at-chance reading. Rows marked stress* share features with the selection bank and are flagged for circularity: excluded from the primary family, the joint max-statistic, and the readings.

claimed. The joint max-statistic over the primary family (one shared source-level permutation per replicate, preserving the cross-detector dependence) rejects chance at $T = 0.382$ ($p = 0.0005$), the stress family at 0.358. The re-encode control, each re-encoded source scored against its original, invalidates codec-lineage readings for 6 of the eight detectors and is clean for the clean-room DeMamba and the stress PM-C; wherever an invalidated detector's number appears the caveat stands, and the paired metric, both sides through the identical path, carries the reading regardless. Detection at the fixed 0.5 threshold collapses for every primary detector (the detecting DeMamba, ranking 0.854, flags 0 of the fakes there): ranking crosses corpora, the operating point does not, the wall's standing split at pilot scale.

C.8 Axis-by-axis decomposition of the container bundle

The container intervention of Section 4.1 moves every container axis at once, and the honest question to ask of it is which axis carries how much of the effect. The controls that answer this exist across the paper; this subsection assembles them into one table so that the isolation achieved, and its boundary, can be read at a glance. Every row varies one axis, or one bundle declared as such, while holding the container fixed wherever the axis under test allows it, and every number re-verifies at build time as elsewhere. Two boundaries the table states rather than hides: the corpus row is itself a bundle (content, collection quality, and generator mixture move together with the training corpus), and no row above touches the generation stage itself; the preregistered regeneration-strength experiment (each real source regenerated by the same generator at a controlled noise strength, from a pure autoencoder round trip at strength zero to full denoising at strength one; protocol `docs/prereg/v2v_strength_dial.md`) closes that last axis, and its row carries the curve's evaluated endpoints (Appendix C.9).

C.9 The regeneration-strength experiment

This subsection is the complete form of the construction experiment that closes the generation-stage row of Table 10; Section 4.3 states its residual form. The preregistration was committed before any generation for it began, with its readings fixed in advance, and one amendment, adopted after an adversarial design review and before any evaluation-stage artifact existed, reimplemented the strength mechanism so that the endpoints differ only in a single declared quantity: every endpoint runs the identical generator graph with the identical sampler schedule, and the strength multiplies that schedule, making the initial latent mixing coefficient exactly the endpoint's label. The design: the same 300 source reals as the construction pilot of Appendix C.7 (the same committed draw), each regenerated by the same open-weight generator (the distilled LTX-2.5 image-to-video pipeline

Axis	Control	Headline result	Where
Rate	wall fakes re-encoded at the reals' measured mean bitrate; detector, reference pool, and every other axis unchanged	both wall inversions stand: PM-C 0.3484 [0.3251, 0.3726], PM-A 0.1517 [0.136, 0.1681], point movements of at most two thousandths: rate alone is not the carrier	4.3; App. D.5
Border	letterbox control: center-crop both classes of the corpus-trained wall readers to 1024×576, padding removed, nothing else changed	removes up to 0.298 of the wall reading (0.967 falling to 0.669 on the text-to-video corpus): the padded border band carries part of the channel	4.3; App. D.5
Resolution	re-render both sides to a common short side (gate g6)	sign flips on Seedance: the noise residual reads 0.169 natively and 0.733 equalized, the HF energy ratio 0.142 to 0.934 (five of six image-forensics classics): native-resolution contrast, not signal	Table 12
Temporal sampling	sampler spacing is window-capped for the reals and duration-proportional for the short fake columns (measured, not intervened)	real spacing fixed at 0.4286 s against fake 0.155 to 0.343 s, and a learner reading only the scalar duration reaches pooled AUC 0.9968: the spacing axis carries the duration field into features that never read a duration	2.3
Content	fake pool's content swapped (caption-matched against collected clips); container and detector fixed	PM-A moves 0.749 [0.718, 0.779] to 0.293 [0.270, 0.315], paired -0.456 [-0.488, -0.425]: content carries signal the container does not	4.3; App. D.5
Motion	construction pilot: content locked by first-frame conditioning, motion selected among candidates (the matching itself downgraded: a 5.3% reduction against the randomized-assignment null, under the pre-registered quarter)	the corpus split survives with content and container held: the confounded-corpus readers invert (0.131 to 0.298), both GenBuster-trained systems detect (0.715, 0.854), the container gate exact at 0.5	4.3; App. C.7
Corpus	retrain at fixed recipe and seed; only the training corpus changes	the confounded Pair1 corpus inverts deeper (0.056 [0.047, 0.066] under container-normalized training), the container-standardized corpus holds (0.928 [0.916, 0.939], 0.976): the residual bundle is the training corpus, its content, collection quality, and generator mixture still moving together	4.3; App. D.5
Generation stage	regeneration-strength experiment, preregistered: each of 300 real sources regenerated by the same generator at a controlled initial-latent mixing strength, content, motion, and container held fixed; strength zero is a pure round trip of the source's own frames through the generator's autoencoder	the round trip alone separates for every deep detector in the roster in the content-matched paired metric (hardest reader 0.867 [0.827, 0.903], three detectors inverted, the flow-stat expert exactly chance), the full generation graph at zero strength exactly transparent, the first correct-direction reading created by the sampler arriving at 0.638 at strength 0.4, and the readers crossing at strength 0.8 (anchor 0.813, round-trip reader 0.580): the generation stage itself carries detector-readable signal at both ends of the strength curve, with the localization and monotonicity readings decided on it	4.3; App. C.9

Table 10: The container bundle, decomposed axis by axis. Each row restates a control the paper reports in its own place, gathered here beside the axis it isolates; every number is re-verified at build time, and intervals are the source subsections' bootstrap 95% CIs. The decomposition has a stated boundary: the corpus row is itself a bundle (content, collection quality, and generator mixture move together with the corpus), and no row intervenes on the generation stage, which is the preregistered regeneration-strength experiment's axis (Appendix C.9). The rate row's scope is the rate axis only: an encode always perturbs codec internals as well (Appendix D.7).

run as video-to-video) under a neutral prompt and a fixed per-source seed shared across endpoints, so across endpoints a given source differs only in the strength. A separate no-DiT-graph endpoint holds the strength-zero reading as a pure round trip through the generator's own video autoencoder. Both classes pass through the identical re-encode into the wall container, and the container gate passes exactly on every evaluated endpoint (every single-feature rule at AUC 0.5); the paired metric scores each regenerated clip against its own source under the same container re-encode alone, so the container codec is symmetric by construction and the autoencoder pass is the one asymmetric step, and the unpaired reference is the pilot's disjoint 700-real pool with per-clip reference scores reused.

Two structural checks precede every reading. First, the strength-zero endpoint through the full generation graph (mixing coefficient exactly zero) is numerically identical to the pure round trip: the raw decodes match at infinite peak signal-to-noise ratio on all scorable clips, and the formal paired contrast between the two endpoints is exactly 0.0000 for all eight detectors, so the generator's graph adds nothing at zero strength and every round-trip statement carries over verbatim. Second, the pre-registered motion-validity check holds at every mixing endpoint: the source-whitened 17-feature distance between a regenerated clip and its own source sits at median 0.142 at the round trip, 0.403 at strength 0.6, and 0.534 at 0.8, against the pilot's selected-candidate median of 2.701, so the selection

mismatch that downgraded the pilot's reading does not apply here. The pure-noise endpoint is the one exception, its median 3.733 sitting above the pilot's own median: regeneration from pure noise destroys the source's motion statistics, which is a further reason the amendment restricts that endpoint to unpaired readings.

The round-trip endpoint decides the first pre-registered reading, and it holds: an autoencoder-layer artifact exists. Every bank-disjoint deep detector in the roster separates the pure round trip from its own source in the paired metric, in both directions; the flow-stat expert sits at exactly chance (paired 0.500, $p = 1.0$), which is the outcome the bank's own earlier falsification under recompression predicted, and the branch that would have reversed that falsification is not triggered. Three primary detectors read the round trip as more real than its own source, extending the inversion lineage to an operation class with no sampler, no dynamics, and no content change; the GenBuster-corpus DeMamba is the striking cell, inverted at paired 0.430 while ranking the same endpoint clearly fake against the reference pool (reference AUC 0.782), a split between the two metrics the re-encode control explains: six of the eight detectors are codec-lineage-sensitive, so their unpaired numbers carry that caveat wherever they appear, and the paired column carries the reading regardless.

Table 11 is the curve. The trend across the evaluated endpoints is reported descriptively, as the preregistration's multiplicity clause licenses, with the strong statements decided by the full-family analyses that follow it: at strength 0.2 the three inverted detectors move toward chance (two of them wash out entirely) while nothing new becomes detectable, so the codec-layer signal the inverted lineage was reading dissolves under mild mixing faster than any sampler signal appears; at strength 0.4 the clean-trained anchor crosses decisively above chance (paired 0.638, interval [0.581, 0.691]), the first reading created by the sampler side of the experiment rather than inherited from the round trip; at strength 0.6 every significant paired movement is downward (the baseline backbone washes to exactly chance at 0.503, paired contrast -0.124 against strength 0.4), the GenBuster-corpus DeMamba's five-point decline ends below the significance bar, and the same DeMamba that washed out at 0.2 re-inverts to its deepest level (0.355, interval excluding 0.5) under a reference AUC flat at 0.782; at strength 0.8 the readers cross, the anchor's sampler-born reading reaching 0.813 and overtaking the retrained TSM for the first time in the experiment, the TSM's largest single-step fall (paired contrast -0.190) leaving it at 0.580, the GenVideo backbone's inversion finally washing to chance (0.457), and the GenBuster-corpus DeMamba sitting at its deepest inversion (0.305, interval excluding 0.5) under the same flat reference band (0.777). The pure-noise endpoint closes the curve from its other end and is excluded from every paired metric by the amendment: at strength 1.0 the source latent has zero weight in the mix, so the content-matched estimand would compare a clip to content it no longer contains, and only the unpaired ranking is defined. Read descriptively, under the codec-lineage caveat that applies to every number there: the GenVideo backbone posts the curve's highest value (0.922, interval [0.904, 0.939]), the retrained TSM ranks the endpoint below chance (0.259), the GenBuster-corpus DeMamba holds its flat band (0.786), and the flow-stat expert's stress row moves below chance for the only time in the experiment (0.418). Pixel-space context for the trend: median reconstruction quality falls monotonically from 34.99 dB at the round trip to 26.46 dB at strength 0.6 and 22.62 dB at 0.8 (12.06 at the pure-noise anchor), and more mixing makes the regenerated clip less like what the paired detectors call fake, for every detector except the anchor.

What the preregistration deferred to the full endpoint family is decided on it. One shared source-level permutation across the primary family at each endpoint (two thousand replicates) rejects the joint null at every endpoint ($p = 0.0005$ at each, the permutation floor), so the per-endpoint separations above are not artifacts of the twelve-cell within-arm family, while the circularity-flagged pair clears the same null at the three highest mixing strengths and rejects it at the round trip and the two lowest ($p = 0.0015$ there against $p = 0.1079$ and above at the three highest); the pure-noise endpoint's unpaired family rejects it as well ($T = 0.4224$, $p = 0.0005$). The localization reading, whose threshold rule was fixed before generation and is defined only over strengths at most 0.8, places the first separating endpoint at strength 0.0 for all six bank-disjoint detectors, three above chance and three inverted, so no primary detector is clean at low strength; the only correct-direction reading first created by the sampler side of the curve is the clean-trained anchor's, at strength 0.4. The non-monotone reading holds: the GenBuster-corpus DeMamba's paired direction flips twice under a reference band flat near 0.78, the retrained TSM declines from its round-trip maximum and then ranks the pure-noise endpoint below chance in the unpaired metric, under the codec caveat that attaches to every unpaired number, and the two detectors inverted at the round trip carry the two remaining shapes, one crossing to the top correct-direction reader, the other washing to chance. Each arm was

Detector	round trip	c=0.0	c=0.2	c=0.4	c=0.6	c=0.8
TSM (retrained, Sec. 4.1)	0.867 [0.827, 0.903]	0.866 [0.826, 0.903]	0.815 [0.772, 0.859]	0.832 [0.789, 0.873]	0.770 [0.723, 0.817]	0.580 [0.527, 0.637]
DeMamba (clean-room, Pair1)	0.682 [0.632, 0.733]	0.681 [0.629, 0.731]	0.658 [0.604, 0.710]	0.626 [0.571, 0.681]	0.557 [0.498, 0.613]	0.542 [0.488, 0.597]
DeMamba (GenBuster-200K corpus)	0.430 [0.375, 0.485]	0.431 [0.376, 0.487]	0.527 [0.471, 0.582]	0.451 [0.396, 0.510]	0.355 [0.302, 0.408]	0.305 [0.253, 0.357]
VideoMAE-GB anchor (GenBuster-200K)	0.417 [0.360, 0.473]	0.418 [0.364, 0.475]	0.544 [0.483, 0.594]	0.638 [0.581, 0.691]	0.660 [0.607, 0.710]	0.813 [0.770, 0.857]
v7e1 (GenVideo backbone)	0.207 [0.160, 0.253]	0.207 [0.161, 0.254]	0.322 [0.269, 0.376]	0.336 [0.279, 0.386]	0.370 [0.313, 0.423]	0.457 [0.400, 0.513]
PM-B (plain baseline)	0.637 [0.583, 0.693]	0.639 [0.585, 0.696]	0.661 [0.607, 0.715]	0.631 [0.577, 0.681]	0.503 [0.447, 0.560]	0.453 [0.400, 0.510]
PM-C (motion-stat expert; stress*)	0.500	0.498	0.503	0.517	0.513	0.507
PM-A (physics backbone; stress*)	0.640 [0.587, 0.693]	0.642 [0.589, 0.696]	0.661 [0.607, 0.715]	0.571 [0.513, 0.624]	0.510 [0.453, 0.563]	0.460 [0.407, 0.517]

Table 11: The regeneration-strength curve, paired metric: P(regenerated clip scored more fake than its own re-encoded source), source-clustered bootstrap intervals (2000 resamples, seed 42), per endpoint 300 sources minus declared corrupt-output exclusions (n = 299, 298, 298, 300, 300; the round trip at the full 300). The round-trip column is the separate no-DiT-graph autoencoder endpoint; the c=0.0 column is the same endpoint through the full generation graph at zero mixing, identical to the round trip by the zero-contrast check in the text; the pure-noise endpoint (c=1.0) enters no paired metric by the amendment and is absent from this table, its unpaired readings stated in the text. Multiplicity: Benjamini–Hochberg at q=0.05 within each endpoint over the twelve primary cells (six bank-disjoint detectors, both metrics), two-sided label permutation (2000, seed 42); the family is stated here because the experiment's strong claims are localization statements over the full curve, decided in the text above. Rows marked stress* share features with the selection bank, are flagged for circularity, and enter no primary family; the PM-C stress row prints without intervals because it is exactly chance at every evaluated mixing endpoint (permutation $p \geq 0.57$). Unpaired reference AUCs are stated in the text only where they carry a reading, each under the codec-lineage caveat the re-encode control imposes.

preceded by a twelve-clip zero-strength control batch through the identical graph; all six decode identical to the strength-zero arm on every clip (72 of 72) under byte-identical frozen generation state, so generation drift across the multi-day run is zero. The pre-registered seed replicate of the midpoint (the first fifty sources under the second seed rule) moves every bank-disjoint detector's paired reading by at most 0.08 (0.12 across the full roster), every interval covering zero; at fifty sources the clustered intervals are wider than the equivalence margin, so the bound is the stated deltas and intervals, not a cross-seed equivalence claim.

Deviations and scope. Five generation outputs arrived corrupt (container-intact headers over broken streams) and are excluded per the preregistered missingness rule with no post-hoc replacements, the counts above; their regeneration remains deferred, and the counts carry the arms as evaluated. The endpoints were evaluated sequentially as each finished generating, never pooled, and the strength-zero round-trip endpoint was evaluated ahead of the generator endpoints, each recorded as a deviation before the stage ran; the full decode scan that catches the corrupt outputs found none on the three cleanest endpoints and none on the seed replicate. The experiment runs one generator family at one resolution with a fixed neutral prompt, so its readings localize layers of that stack, not the field's generators at large; equivalence at the pre-registered margin is claimed nowhere, so no endpoint is declared indistinguishable from its source, and the descriptive trend sentences above sit beside the full-family joint max-statistic and the decided readings stated above them.

C.10 THE K=895 EXTENSION BOOKKEEPING

This subsection is the complete form of the label-saturation bookkeeping that Section 4.3 summarizes. Extending the fine-tune pool to the full split, 895 fine-tune fakes, with the split rotated so that the 100 former fine-tune clips become the held-out test (clips the k=895 model never trained or selected on; their only prior use was as the k≤100 steps' label pool and in direction diagnostics), leaves the level where it was. On the rotated pool PM-B holds 0.994 (CI [0.986, 1.000]; the interval is wide because the rotation caps the held-out pool at n=100) and PM-A 0.998; on the fixed 183-clip Seedance-2.5 pool, the one test set shared by every k, PM-B moves 0.968 at k=100 through 0.989 at k=200 to 0.998 at k=895, and PM-A 0.995 to 0.999. Nearly nine times the label budget buys no new level: every step from k=100 upward already sits between 0.96 and 1.00, far past the 0.72 ceiling the frozen bank cannot cross, and the only movement whose intervals separate is PM-B's on the 2.5 pool. The two reference classes agree (0.994 against the id reals, 0.995 against the training real pool), so the number is not an artifact of which reals score it.

C.11 DURATION-BALANCED RETRAINING: E2 BOOKKEEPING

This subsection is the complete form of the E2 experiment that Section 2.3 summarizes. The duration-balanced retrain used 55556 training rows with 1571 reals surviving the match. The scale-up on the identical features agrees with its conclusion: training on $9\times$ more data moved the official metric by $+10.79$ (CI [10.49, 11.09]) while balanced accuracy moved -0.21. Nine times the data produced no detectable improvement in detection, only a harder lean toward the 76%-fake training prevalence.

C.12 THE OVERLAY-CHANNEL AUDIT

This subsection is the complete form of the pixel-visible channel that Section 2.3 states in residual form. The audit instrument is a zero-training static-overlay rule: sixteen frames decoded per clip, a per-pixel temporal standard-deviation map over those frames (the "static" map), a Sobel gradient map of the temporal median frame (the "edgy" map), and an overlay score defined as the fraction of pixels that are both static (below the within-clip 25th percentile of the static map) and edgy (above its 90th percentile of the gradient map), computed over the full frame and over the outer 18% border band. Every threshold is a within-clip quantile, so the rule has no learned parameters, no benchmark labels, and no fitted constants; a persistent high-contrast logo or banner scores high, static low-gradient background and moving edges score low. Unlike the container rules of Section 2.1 it decodes frames, so it is a pixel-level exhibit rather than a zero-decode rule.

The pools: three real pools (Vript, GenVidBench Pair1; HD-VG, Pair2; GenVideo-100K reals) against three fake pools (MorphStudio, Pair1; CogVideo, Pair2; Sora, GenVideo-100K's Sora training pool), 100 clips per pool, with a 200-frame visual label pass per pool classifying overlay-bearing frames by kind (station logos, news banners, and other overlays). The full prevalence table behind Section 2.3's headline cells: 88% of GenVideo-100K reals and 41 to 47.5% of the GenVidBench reals carry such an overlay, against 7.5% of GenBuster's reals and 0.5% of its fakes; video-driven generators can inherit the channel from their source frames, and 55% of GenVidBench's MorphStudio clips do. All nine real-versus-fake rule cells separate the two classes in the reals-higher direction at pooled AUC 0.7366 to 0.890; the weakest pairing is HD-VG against MorphStudio (0.7366 [0.663, 0.809]), the two headline cells quoted in Section 2.3: 0.753 [0.684, 0.825] (Vript reals against MorphStudio fakes) and 0.890 [0.840, 0.933] (GenVideo-100K reals against CogVideo fakes), and the Sora pairings sit between (0.7933 to 0.840 across the three real pools).

The label pass and its independent check. The prevalences above are the original label pass's values, a single large vision-language model labeling one center frame per video, and the per-kind splits are quoted from that pass. The original pass's model outputs were lost in a disk failure and rerun on the same frames: the original 320-pixel extractions survived, so the rerun's comparability is frame-exact, and it used a smaller model of the same vendor under a hardened prompt at 640 pixels. Binary agreement with the recovered original labels is 70.9% over 1199 frames, with the heavily overlaid scraped-real pools read lower (69.5% on GenVideo-100K against the original's 88) and the container-clean reals higher (22.0% against 7.5), the labeler dependence the original audit itself disclosed. The load-bearing contrast holds under both labelers, and the fakes side tightens into an interval: 2.5% [0.5, 5.0] of GenBuster fakes carry an overlay under the rerun labeler. The original labels' independent check is now cross-vendor: a different vendor's vision model adjudicated 100 stratified frames at native resolution and agrees with the original labels on 88.0% of binary rulings, the rate the original pass's own human audit reached (87.5%), confirming 44 of the 50 frames where both machine labelers agreed, so the printed prevalences survive an independent-vendor check rather than only a same-vendor one.

What the channel does and does not show. It is a collection-side artifact of the real pools: the confounded benchmarks drew their reals from broadcast and web video that carries logos and banners, and their text-to-video fakes do not reproduce them, so a rule that reads nothing but overlay presence separates the two classes at the same order as the container rules. Two boundaries fence the reading. First, video-driven generators inherit the channel from their source frames, so the channel is a property of the construction, not of generation as such: 55% of MorphStudio clips carry an overlay, which is why the weakest pairing is the one against MorphStudio. Second, the channel does not decompose any learned detector's number; establishing what share of a trained system's accuracy rides overlay presence would need the overlay-matched retraining controls of Section 4.1, here run for the container but not for this channel. The learned companion components of the original audit (an overlay-only

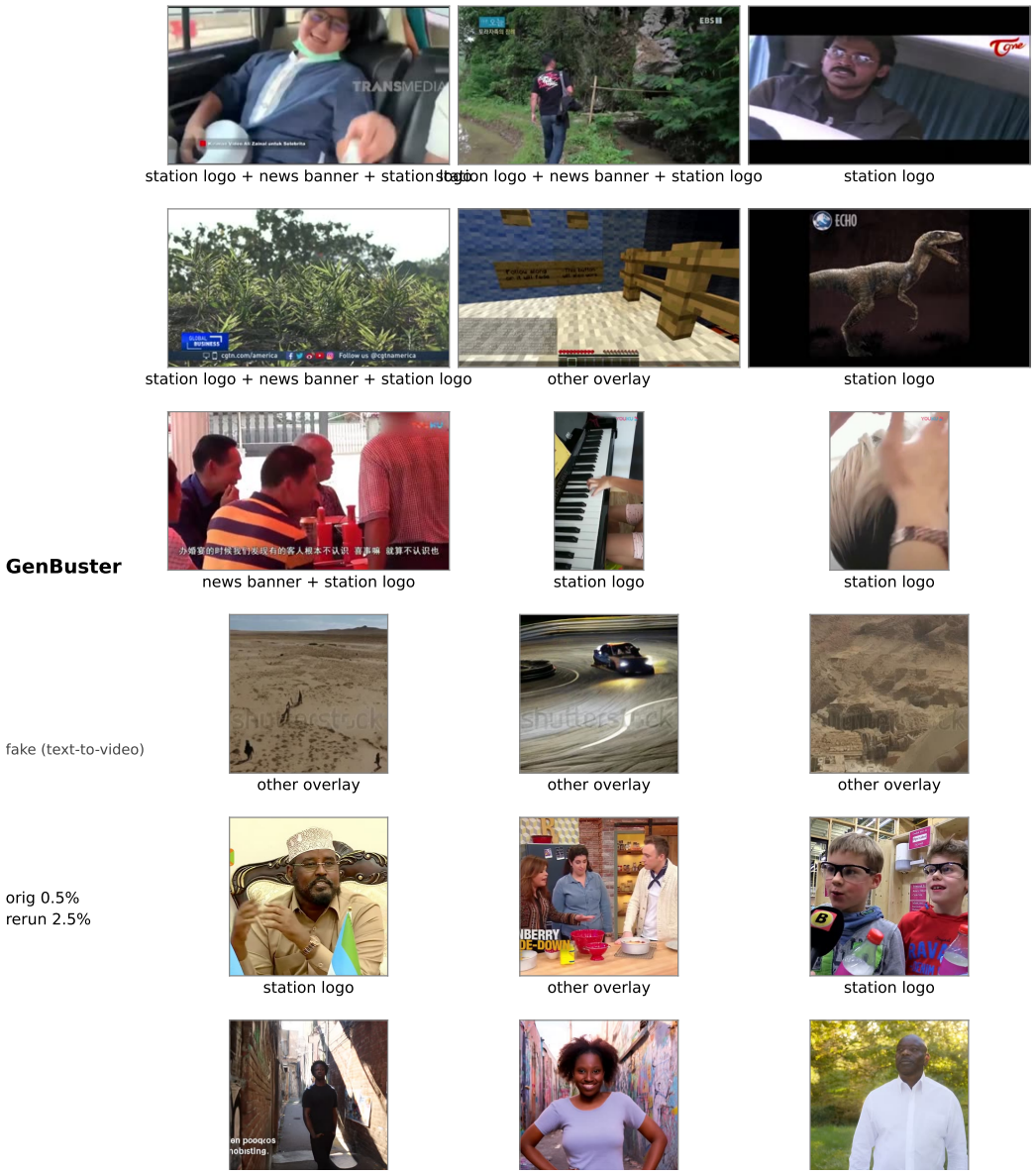

Figure 1: The overlay channel as pixel content. Each row is one pool of the label pass: the three scraped real pools, the video-driven MorphStudio fakes, and the two GenBuster pools. Cells show, per pool, the first three frames in index order that the recovered original label pass marked overlay-bearing (the GenBuster fakes have a single positive, so the remaining cells show the first frames labeled clean); the text under each frame is that pass's ruling, and the left column gives the pool's any-overlay prevalence under the original labeler and under the rerun labeler of the previous paragraph. Logos and banners are visible pixel content of the scraped reals and of MorphStudio, which inherits the channel from its source frames, and are absent from the text-to-video fakes. Frames are the surviving original 320-pixel center-frame extractions; the figure is re-cut by `scripts/paper/make_overlay_examples.py` from the committed label manifest and the two prevalence files, with no hand-placed element.

detector set and a spoofing intervention) are not part of this release; the rule readouts and label prevalences above are recomputed from the released per-clip rule scores and the label manifest, and the exhibit figure can be re-cut from the released frame draws; Figure 1 is that re-cut.

C.13 Clean-control provenance detail

This subsection holds the two provenance facts that Section 2.4 records for the local copies behind the clean-control verification. The control's boundary, in full: its four declared container axes are equalized, while bitrate, source quality, and re-encoding still differ. First, the mini split's files were recovered from a truncated archive with a missing central directory by parsing local file headers (1,780 of 1,781 entries raw-inflated; integrity artifacts committed), which cannot affect a constant-container result. Second, the GenVideo-100K local copy carries 1,079 zero-byte files among 189,998 plus two byte-identical duplicate generator directories, so its rule was scored on a committed stratified sample with 71 NA-probed rows excluded.

D Extended Evidence from the Second Cut

The main text states each finding once, with its headline numbers and intervals; this section holds the extended form of the compressed passages, in the order the sections appear. Nothing here is new, and every number keeps its inline provenance and build-time verification.

D.1 The abstract's extended form

The abstract states each claim once; the sentences below are its extended form, with the values and citations the one-page statement omits. The clip-length phenomenon is published: VidAudit's clip-length classifier reaches leave-one-generator-out AUC 0.998 on GenVidBench (Cakiroglu et al., 2026), and the GenVidBench duration rule's train-fit score sits below the matched published best (83.27), with the two sides separable at AUC 0.9968 from duration alone, the first score we know of a container rule under any AIGV benchmark's own leaderboard metric. On the TSM normalization the full per-column record: the intervention takes the MuseV column from AUC 0.856 to 0.552, a fall that leaves weak but statistically nonzero residual discrimination at this sample size, and real-class accuracy falling to 58.32% at the CogVideo-container end of the range (the MuseV floor appears in Section 4.1), and the official mean Top-1 from the native value to 70.92–76.43: the retrained detector's number rides the container gap. The sign inversion's full four-ratio instance: on Wan-2.2 the fake/real flow-magnitude ratio is 0.139, and Seedance-2.5 shows 3.342 on a duplicate-free real pool. The wall was re-measured at per-generation scale (995 container-normalized Seedance-2.0 clips; 183 Seedance-2.5), where the co-trained backbone is inverted to AUC 0.154 on the 2.5 generation (the 2.0 value appears in Section 4.3), and k-shot calibration recovers feature directions (truth-agreement 17.6% to 91.5%) without recovering detection. The priced-ladder sentence, in numbers: ten in-family labels fine-tune a deep backbone past it (0.831), one hundred hold 0.995 on the next generation, the same labels leave the frozen bank flat (0.339), and wrong-family fine-tuning inverts the model (0.245), while the step's spillover ranks the wild mix at 0.744 at 8% detection. The label curve saturates: PM-A reaches 0.998 from 895 clips against 0.996 from one hundred.

D.2 Introduction: extended statements

The failure history, in full: two of our prior submissions, unpublished method papers on physics- and coherence-based detection, claimed gains from spatiotemporal coherence probes and explicit physical-consistency features; both sets of claims were falsified under controls we built to defend them (motion-amount matching, symmetric frame-rate protocols, container equalization), and the instrument that falsified them is the contribution of this paper. That sequence of claim, control, falsification, and sharper control is the validation history of the measurement instrument, and the paper reports it in those terms. The same battery is applied without modification to seventeen signal families, drawn from our own probe-bank lineage and from literature classics, under the survey's scope rules rather than under any claim of novelty for the set.

The first finding's extended form: three 2026 audits overlap the container finding: VidAudit's six-control audit (Cakiroglu et al., 2026), STALL's standardized GenVideo protocol (Ben Hayun et al., 2026), and the Michels et al. bias decomposition (Michels et al., 2026); none of them scores a container rule under a benchmark's official metric, executes zero-decode header rules on the as-released files, or traces the sampler path by which a learned detector inherits the field. The flagship table does not reconstruct from its own per-source rows under any stated aggregation (standard

reconstructions differ by 0.95 to 20.0 points), and the three-layer decomposition into container (AUC 0.9968, reaching the features through the sampler's window-capped duration/7 spacing; the introductory rule is FAKE below 4.15 s), in-band content (0.6947), and the threshold layer explains published leaderboard numbers without any detector detecting anything. The TSM result in full: retraining the benchmark's baseline recipe under the same protocol and changing the real side's container as a unit moves the official mean Top-1 from 80.42 to 70.92–76.43, so the retrained baseline's number rides the container gap it was never supposed to see.

The second finding's extended form: the survey is fenced to the frozen-estimator population (learned augmentation-trained detectors certified elsewhere (Michels et al., 2026) sit outside it), and the sign-inversion statement carries its full four-ratio instance (0.139 on Wan-2.2 and 3.342 on Seedance-2.5, alongside the two quoted in the main text) and its full citation load: VidAudit reads the prior's direction as fixed, predicting generated video under real unconditionally, while Michels et al. (2026) find under-motion on four of five datasets and predict that the prior's discriminative power "may quickly become obsolete" as generators improve, a forecast our reversal realizes on the measurement side; the reversal carries a 2024-trained expert below chance (0.946 to 0.160), the repeated instance of Section 3.3. The decay half of the finding is a designed reading rather than a discovery: on the 2026-era wild track the flow-statistics expert scores 0.471 after 0.769 and decays from 0.769 to 0.471 (an interval that includes 0.5 at $n = 150$, so we claim the decay, not an inversion), and GenBuster-Bench's tracks are era-defined precisely so as to exhibit such decay (Wen et al., 2025).

The first finding's residual numbers: the all-fake constant scores 80.00, and the TSM normalization drops real-class accuracy from 78.29% to as low as 30.78% (values also in Sections 2.2 and 4.1).

The second finding's residual numbers: the iron-law estimate's interval is [0.177, 0.222], the four-ratio instance reads 0.362 on a 2024-era generator against 1.217 on Seedance, the clean-domain estimate is 0.198, and the below-chance expert runs 0.946 to 0.160.

The third finding's residual numbers: the survey leaves 0 universal survivors, PM-A transfers at pooled AUC 0.865, and the ladder runs to 0.996 at a hundred labels (0.998 at 895) with off-family spillover 0.744 at 8% detection, all stated with intervals in Sections 3.2 and 4.3.

The third finding's extended form: the identical PM-A-versus-PM-B predictions order one way under the official metric ($+5.79$ points), shrink under balanced accuracy ($+2.48$), and reverse under pooled AUC (-2.16; intervals, both balanced conventions, and the per-column decomposition are Appendix C.1); the official margin is carried by a single column worth $+52.43$ points. What survives is deliberately narrow: no fusion claim (fusion loses to the solo backbone on the clean container), and an adaptation protocol rather than a deployment artifact: verify the container, calibrate direction against roughly ten labeled clips from the generation actually in the feed, stopping when the calibration ladder shows the bank has nothing left to recover, which at the 2026 wall it does, and monitor the operating point rather than the ranking.

D.3 Forensics: extended exhibits

GenVidBench container statistics. Real pools: HD-VG at a median duration of 11.0 s at 30 fps; Vript at a median duration of 6.98 s and 23.98–30 fps (median 29.97; at least a tenth of Vript reals run at roughly 24 fps). Recomputed on the full committed cache, duration-only AUC is at worst 0.9917 (CogVideo) and at best 0.9996 (MuseV), with per-column values in Figure 2. The train-fit rule's mechanics: a duration classifier trained only on Pair1 (where the longest fake is 3.0 s) places its decision boundary near 3 s, which classifies every CogVideo test clip (all 4.125 s) as real. The oracle threshold reaches perfect accuracy on all four fake columns and 99.14% on the real column, and it is not doing delicate work: any threshold in $[4.13, 4.7]$ s scores at least 99.8.

GenVideo-Val execution detail. The threshold used is GenVidBench's own oracle threshold (4.15 s), run on all 18286 test files, where the rule's AUC is 0.9922 (accuracy in the main text). DeMamba's published F1 / AP on the Val split are 90.2 / 97.1. The duration threshold was transferred across benchmarks (with 0 false positives and 191 false negatives on the Val files), while Val's own oracle threshold (10.667 s) gives 99.56%. The triple confound, quantified: in a stratified probe of 200 real files, all 200 run at 3 fps and 298×224, while the fake files run 8–50 fps at 320p–1080p; a duration-only rule, an fps-only rule, and a resolution-only rule each separate the two sides alone.

Per-family survey prose. GenWorld's own Table 1 admits the split outright: every one of its 10 generator sets at 8 fps against reals at 10–30 fps. GVD/AIGVDet (Bai et al., 2024) patches a JPEG-versus-generated frame-format bias by re-compressing only the fake frames. DVF (Song et al., 2024) was excluded by Corvi et al. (2025) for pairing MPEG-4-Part-2 reals with H.264 fakes, and removing that codec split inverted their detector ranking. GVF/DeCoF (Ma et al., 2024) admits resolution and length inconsistency in its appendix. AIGVDBench (Ma et al., 2026) unifies the codec axis (its own mixed-versus-unified control moves results by at most 0.033) but leaves fps spanning 5–24 across generator sets. And MVAD (Hu et al., 2025) prints no container columns at all, but its own Table 2 documents mismatched durations between the two sides (reals at 3–60 s, fakes at 2–60 s with different per-generator profiles), which is affirmative confound evidence on the one axis it does report; we classify it confounded rather than unclassifiable on that basis, with fps, resolution, and codec remaining unstated. On GenVidBench the matched 143k published best (83.27) sits between the train-fit rule's 79.87% and the oracle rule's 99.83%, inside the range a threshold on one container field traverses by moving a single scalar, and far below the 0.9968 AUC at which the two sides separate from duration alone without any threshold sweep.

RA-Bench: the tenth family, audited on its full release. RA-Bench (Liang et al., 2026) pairs 1319 real crisis-event anchor videos against 16056 clips from 9 generators, and it is the surveyed family built with anti-shortcut defenses: its release standardizes the codec, states an audio-removal policy, and ships an evaluator whose constant-scorer floor is exact. Our audit is zero-decode over the full public main-track media of the current v2 release, every 17375 files streamed, probed from the container header alone, and deleted (per-file rows committed). The distributional rules fall: frame rate alone separates the two sides at AUC 0.9237 [0.9124, 0.9348], duration at 0.8841, file size at 0.8146, and audio-track presence at 0.9252 (each permutation $p < 5 \times 10^{-4}$, two-sided). The defenses hold where they were built: every file in both classes carries the same codec tag (17375 of 17,375), and a constant scorer scores exactly 50.0 AUC under the release's own evaluator on all 9 sources, so this family's metric floor is honest where GenVidBench's is degenerate (Section 2.2). The audio channel carries a construction violation of the release's own stated policy: one generator's entire fake set ships audio (1787 of 1787 clips), and 5 of 1524 clips of the Seedance-2.0 set do likewise; 1269 of 1319 reals carry audio, and 154 reals sit within half a frame of 24 fps (11.68%). One number diverges from the audit of this benchmark we had previously submitted, and we disclose it rather than reconcile silently: the native-resolution header rule reads balanced accuracy 0.8416 on the current release against the previously submitted 0.989 (the frame-rate value above sits inside the previously submitted interval [0.919, 0.940]); the release's README documents a post-release rights review that moved 511 real anchors to URL-only listing, the simplest consistent explanation being that the two audits ran on different snapshots of a changing release. A benchmark that changes under its audit is itself a measurement finding: the instrument reads the release it is handed, not the paper's description of it.

The prior-audit delta list, in full. VidAudit (Cakiroglu et al., 2026) audits GenVidBench and AIGVDBench under a six-control protocol (canonical codec re-encoding, a leakage audit, a real-vs-real floor probe, matched-harness retraining, seed-stable confidence intervals, and cross-dataset replication) and ships an audited 14-detector toolkit; its clip-length existence proof (LOGO AUC 0.998; arXiv June 2026) is the phenomenon our per-video rows translate into official metric points, but its number is a LOGO AUC on canonically re-encoded subsets, not a score under the benchmark's own leaderboard metric, and it leaves the GenVideo family out of scope. Ben Hayun et al. (2026) publish the GenVideo container asymmetries our rules execute against (including the 3 fps real pool) and standardize the protocol rather than score a rule; Michels et al. (2026) independently confirm the same asymmetries and decompose four motion detectors into preprocessing and motion-prior bias. What none of the three does is decompose *how* a learned detector comes to ride the container (Section 2.3), expose the metric floors that convert the shortcut into leaderboard points (Section 2.2), execute zero-decode container rules on GenVideo-Val and GenVideo-100K, or verify, on disk, that a benchmark standardizes all four container axes and then use that clean control as a measurement substrate (Section 2.4; the construction is BusterX's (Wen et al., 2025)). SAGA (Kundu et al., 2025) and Real-Data-Lies (Fang et al., 2026) are complementary: generator attribution and a content-level quality shortcut, not container contrast at the source.

Family	Venue	Rule result	Axes matched	Verdict
GenVidBench (Ni et al., 2025)	AAAI'25/26	99.83[a] Top-1	0 of 4	Fallen (executed)
GenVideo-Val (Chen et al., 2024)	SCIS'24	98.96 ACC	0 of 4	Fallen (executed)
GenVideo-100K (Chen et al., 2024)	SCIS'24	98.43[a] ACC	0 of 4	Fallen (executed, sample)
GenWorld (Chen et al., 2025)	arXiv'25	not executed	0 of 4	Fallen (own Table 1)
RA-Bench (Liang et al., 2026)	arXiv'26	fps AUC 0.9237 (full release)	1 of 4 (codec)	Fallen (executed, full release)
GVD / AIGVDet (Bai et al., 2024)	PRCV'24	not executed	frame-format patched	Confounded (admitted)
DVF / MM-Det (Song et al., 2024)	NeurIPS'24	not executed	codec mismatch	Confounded (ext. audit (Corvi et al., 2025))
GVF / DeCoF (Ma et al., 2024)	ICME'25	not executed	res+length admitted	Confounded (admitted)
AIGVDBench (Ma et al., 2026)	CVPR'26	not executed	codec only	Confounded (fps 5–24)
MVAD (Hu et al., 2025)	arXiv'25	not executed	durations mismatched (own Table 2)	Confounded (admitted, partial)
GenBuster-200K (Wen et al., 2025)	arXiv'25	cannot fire	4 of 4	Clean (verified locally)

Figure 2: Field-wide container audit: ten benchmark families, shown as eleven rows because GenVideo-Val and GenVideo-100K are two splits of one family (DeMamba's GenVideo). "Fallen" means a zero-feature container rule was executed locally, or the family's own paper admits the split: 4 of 10 families (GenVideo counted once, both splits executed locally; RA-Bench executed on every file of its full release), with 5 more confounded by their own documentation (MVAD included: its Table 2 documents mismatched durations, the only axis it reports) and 1 clean. "Rule result" is the accuracy of a zero-feature threshold on container metadata (duration unless noted), scored on the as-released local copy; no frames are decoded. [a]Threshold selected on the test files themselves (oracle); the train-fit duration rule scores 79.87% official on GenVidBench and 98.04% at $t{=}8$ s on GenVideo-100K, whose local copy excludes 1,079 zero-byte files and 71 NA-probed rows (committed stratified sample). Per-column GenVidBench duration AUCs on the full cache: CogVideo 0.9917, SVD 0.9993, MuseV 0.9996, Mora 0.9963. "Container axes" are duration, fps, resolution, codec. Local rule results are computed from per-video ffprobe rows committed with the paper; the remaining verdicts quote each paper's own documentation, with sources in the supplement.

Metric-exhibit extensions. The balanced-accuracy inversion behind the main text's exhibit, in full: the same scores place the system 10th of 12 at 70.49 versus 88.64 for the published best. The genre of the constant-floor observation is established, and we claim only its instantiation on this family of leaderboards: constant-predictor and prevalence floors are classical measurement points in evaluation (Provost et al., 1998; Schaeffer et al., 2023); per-dataset averaging has been shown to overstate deepfake-detector generalization and reverse method rankings (Nguyen et al., 2026); VidAudit recommends floor-aware metric tuples over any single number (Cakiroglu et al., 2026); and BusterX already reports real/fake accuracy separately (Wen et al., 2025). The all-fake constant's 80.00 under GenVidBench's specific 4:1 column structure, above 10 of the 11 systems (median published mean 62.83) in the benchmark's own 143k Table 7, and the flagship table's non-reconstruction from its per-source rows are, to our knowledge, the first such instantiations on an AIGV leaderboard. The prevalence-instrument exhibit's remaining controls: the same system calls 51.33% of its own training reals fake in-sample (where prevalence is identical by construction, so the calibration error survives even with no train/test-prevalence gap to explain it), and 52.62% of container-clean reals longer than 8 s fake.

Reconstruction conventions and threshold sensitivity. The defender's reply and ours: a defender can reply that the container columns were never hidden; we accept the premise and reject the conclusion, because a benchmark may print its own confound and still average five columns into a metric with an 80.00 constant floor, a non-reconstructing flagship table, and a one-column threshold that solves the task. The metric-instrument exhibit's margin over the floor: 4.29 points above the all-fake constant (paired cluster bootstrap 95% CI [4.05, 4.52] on 67,853 test clips). The reconstruction gap in full: the raw sum of DeMamba's seven published per-source accuracies falls twenty points below the published headline, and whatever convention reconciles them is not stated in the paper. The raw sum of DeMamba's seven published per-source accuracies is 65.47; the gap above is measured from it. The n-weighted reconstruction lands within 1.20 points, suggesting an unstated weighting convention rather than an arithmetic error, and five-largest-column aggregation gives 84.52. Threshold sensitivity: holding the score vector fixed and moving only the decision threshold, the official mean of the metric-instrument system spans 85.40 at threshold 0.40 to 69.82 at 0.70, and on the matched 143k protocol every published system above 80 except DeMamba's 6M-protocol 85.47 sits inside that band; the leader is excluded only because it is a different protocol. Under the motion-magnitude-matched control the in-band model falls to 0.4999 on MuseV (chance) and 0.6465

on SVD; and the E2 retraining moves real accuracy at the fixed 0.5 threshold from the instrument's 47.48% to its collapsed value among the supplementary values below.

E3 per-column detail and E2 mirror. The band's edges are the exact edges of the executed selection: the lower edge is the shortest fake column (MuseV, every clip 1.084 s), the upper the 95th percentile of fake durations (band edges 1.084 and 4.125 s), leaving 54438 fakes in band: every SVD clip is 1.4 s and every MuseV clip 1.084 s. Every CogVideo clip is inside the duration band, all 13850 of them at exactly 4.125 s, and the reals that survive the band are the short tail of HD-VG, not a representative sample ($n_{\text{real}} = 106$). Per-column in-band AUCs against the band reals are 0.6776 (CogVideo), 0.6813 (SVD), 0.67 (MuseV), and 0.7488 (Mora); the same caveat on the 106 reals applies to each. The E2 retraining's threshold relocation is the mirror image of the published I3D and SlowFast rows (real accuracy ≥ 93.6, MuseV accuracy ≈ 8).

Layered decompositions have prior form. VidAudit decomposes every audited score into the dataset-identity floor it shares with a real-vs-real probe and an above-floor margin (Cakiroglu et al., 2026); Michels et al. (2026) decompose four motion detectors into preprocessing bias and motion-prior bias; and in the image domain, container-adjacent bias layers (JPEG quality, resolution misalignment, semantic shortcut) are established (Grommelt et al., 2024; Choi et al., 2025; Guillaro et al., 2025). Our decomposition differs in target and grain: it splits a leaderboard number under the official metric into a container layer that enters both the labels and, through the sampling window, the features; a content layer isolated inside the duration overlap band; and a threshold layer that the 4:1 column metric converts into points, and it identifies the specific sampler mechanism (spacing = the duration capped at the 3 s window, over seven) by which a learned detector inherits the container field without ever reading it, which is the part no prior audit traces.

D.4 Instrument: gate narratives and survey detail

g1 narrative. The latent-branch detector's fragility profile is perturbation-specific, not a general high-frequency dependence: a detector that gains from a resample destroying more high-frequency content than JPEG-40 was not reading high frequencies; it was reading the recompression artifacts themselves: blur moved it by only -0.042 and a downsample to 112 px *gained* it 0.057. The full before/after row, including the clean score and the fps-4 control, is Table 12.

g2 narrative. The two g2 exhibits quote two checkpoints of the same bank on the same matched protocol, and we name both: the full 17-feature leaderboard model (8k-subset checkpoint in Table 5 row 4's 0.776 $\rightarrow$ 0.400; deployed full-data checkpoint in the main text, same verdict). The falsification is the same verdict either way. The jerk-family exhibit it falsified outright: jerk statistics fell from 0.728 to 0.529 under matching.

g3–g8 narratives. The flow-cycle demonstration's three signed AUCs are in Table 12; the token-grid family's grid-specificity numbers (stride-3 versus stride-4 contrast, and the phase-null destruction) and the resolution-equalization flips of the image-forensics classics are likewise in the table, as is the final symmetric 30 fps adjudication of the token-contrast family, sign-inverted at 0.429. g4's named falsification is the below-chance regularity of Section 3.2 itself, and the band-limited claim of Table 5 row 1, which g4 falsifies jointly with g6. The g8 inversion in points: our flow bank, officially first of twelve, is inverted by 18.15 points under balancing. g7 and g8 were added after the rest of the battery had already run; they belong to the instrument because they generalize beyond one benchmark.

Multiplicity accounting detail. The family is the survey's committed admission surface: every feature-by-domain AUC cell of the sixteen tested families that produce AUC cells (the co-training ablation produces none), defined before the null was computed, with the admission rule the bank bar of Section 3.2 fixed at 0.10; gate re-tests (compression, resolution, matched-control retests of already-admitted cells) are controls rather than additional admission tests, so no cell enters or leaves the family after its AUC is seen. The survey's feature-by-domain AUC comparisons (total in the main text) split into 168 recomputable from committed per-video feature files and 14 existing only as summary statistics and bounded analytically ($168 + 14 = 182$); of the 182, the 138 cells that met the bank's admission bar are the ones the null and BH control adjudicate (the remaining 44 fall below the bar and enter no verdict), and the best-cell null's 95th-percentile point estimate is 0.0852 with

Gate	Specimen before	After the control
g1	latent detector clean 0.897; JPEG-40 Δ-0.223; blur Δ-0.042; fps4 Δ0.001	resize112 Δ0.057: the failure is recompression-specific, not high-frequency loss
g2	jerk 0.728 (n=260, ratio 1.028); MuseV bank 0.766	matched 0.529; 0.4999 (chance)
g3	flow-cycle fakes *lower* than reals on CogVideoX (0.272), Wan (0.187)	Seedance *higher* (0.729): direction inverts across eras
g5	token-contrast same-sign 3 domains (0.671/ 0.695/ 0.585), CRF-robust	stride-3 vs. stride-4 contrast 0.054 vs. 0.040 (monotone, not peaked); phase null 0.043 → -0.011
g6	noise-residual 0.169; HF energy ratio 0.142 (native, seedance)	res480-equalized 0.733 / 0.934: sign flips

Table 12: One demonstration specimen per gate, before and after its control. Full provenance for each number is inline in the source and re-verified at build time; the g7/g8 demonstrations are metric exhibits reported in Section 2.2.

interval [0.0845, 0.0859], so all 138 admitted cells clear the bar, and Benjamini–Hochberg control at q=0.05 rejects 158 of the 182 comparisons with no admitted cell losing significance.

Decay-test detail. The permutation count behind the multiplicity null is 10000 label permutations per domain. The paired decay bootstrap: the flow-statistics expert scores 0.471 after 0.769 on the 2025 track, a difference of 0.299 with 95% CI [0.247, 0.352], excluding zero; the wild value's own interval is [0.417, 0.525] at $n = 150$.

Attribution leak-fix numbers. The NaN-mask channel alone now scores 0.096 (chance), against 0.893 before the fix.

Fence elaboration. The zero-survivors fence, in full: Michels et al. (2026) simultaneously close the motion-detector family and certify a learned, augmentation-trained detector, WaveRep (Corvi et al., 2025), across all 5 of their datasets, which is why our zero-survivors verdict and their certification do not conflict; and the zero is checkable: a missed family can be added and the battery applied unchanged; a single survivor would overturn it.

Battery and survey co-evolved; stated, not hidden. The battery was not fixed before its targets were known: the demonstration exhibits of g1, g2, and g5 are claims of ours born in this project's own detector line (Table 12), and g7/g8 were fixed after the GenVidBench metric audit, not before the survey began. The survey's family list is likewise not independent of its authors: part of it descends from our own probe line (the latent branch, coherence, token-grid, jerk, lag-k among them), the rest from literature classics (the image-forensics axes, flow statistics, VFI), the per-family rows of Table 4 naming each. The fence above is the guard this needs: the zero is a verdict about the tested list under the stated gates, and any reader can extend either.

Survey bookkeeping. Of the families entering the survey (the total count is the main text's), 1 remains queued (feature-series structure and ballistic-trajectory fits) and 1 is blocked (audio-visual synchrony: no paired-audio real pool locally); the main text's tested and survivor counts are the survey itself.

Bank-level per-domain detail. The flow-magnitude prior's CogVideoX interval, in full: 0.955 (95% CI [0.950, 0.959]). On Seedance-2.0 the flow-divergence variance family reaches 0.926 (95% CI [0.912, 0.939]): an in-domain leader whose g2 motion-matched control has not yet been run, so it is bank-level on magnitude alone, and the static-noise floor reaches 0.917 (95% CI [0.895, 0.937]) with its g2 control passed (matched 0.913) but computed on the msrvtt real pool, whose 3 fps source duplicates frames (zero inter-frame flow stairs).

Iron-law protocol detail. The single-domain MLP experiment: an MLP over the full 16-feature survey flow bank (the cross-domain feature set; the GenVidBench leaderboard instrument of Section 2.3 is the same bank plus a 17th saturation feature), trained on CogVideoX alone ($n_{\text{train}} = 8466$, 7687 after leakage control, $n_{\text{test}} = 1600$). Below-chance per-generator cells have been reported for individual detectors on these benchmarks (Ben Hayun et al., 2026; Cakiroglu et al., 2026); what we add is the elevation of that observation to a property of single-domain joint fits under leakage control. On the msrvtt real pool (ledger A2's condemned 3 fps staircase domain) the same protocol yields 0.019–0.244 depending on which feature subset is used; we quote it only as subset-conditional context, not as the law's value, which is the clean-domain estimate in the main text. The single features inside the same model transfer at up to 0.574.

The distinct-pair correction. The raw Seedance-2.0 ratio 3.011 is computed on a real pool whose 3 fps frames duplicate (6.5 duplicate zero-flow pairs per video); counting only distinct frame pairs gives the corrected value quoted in the main text. The physics expert's cross-generator 2024-era transfer, for completeness: 0.830 on Wan-2.2, between its in-domain and Seedance values in the main text.

Flow-magnitude intervals and the Vript replication. The four ratio intervals, in full (2000-draw bootstrap on the committed per-video flow magnitudes): 0.362 [95% CI 0.259, 0.486], 0.139 [0.111, 0.172], 1.217 [1.038, 1.441], and 3.342 [2.643, 4.161], each excluding 1. The independent-family replication, against Vript reals (a different collection lineage; 300 clips, 299 on the distinct-pair ratio's frame budget): 0.215 on CogVideoX, 0.090 on Wan-2.2, 2.561 on Seedance, and 1.681 on Seedance-2.5, every interval again excluding 1 and every sign matching the main-pool value.

The inversion's internal structure (survey-side preview). The inversion has an internal structure that the wall measurement of Section 4.3 localizes feature by feature on container-normalized Seedance at $n{\approx}1000$: the bank's central-tendency features (the ones that lead the 2024 domains, te_mean foremost at 0.955 on CogVideoX) carry almost nothing there in either direction (direction-free AUC 0.5106), while its dispersion features hold moderate signal in either direction (0.6481–0.7352; direction-free AUCs use full-set labels and are mechanism characterization, not system rows): the representation that made the bank strong on 2024 generators is the part that inverts, and what survives the flip is its weaker half.

Epistemic status of the direction-shift reading. If signal direction tracks generator generation, then by 2026 the motion prior itself should fail as a *detector* on generators whose motion budget has converged toward real footage. The survey evidence for the direction instability was committed on 15 August 2026; the GenBuster-wild track (eleven 2026-era commercial generators, container verified uniform across all files, so no container axis is available to blame) was evaluated three days later. We state the epistemic status plainly: no quantitative prediction for this benchmark was committed before the evaluation; the direction-instability evidence precedes it by three days, which is temporal precedence of the observation, not a registered prediction. The one pre-registered set that does exist for this line (Seedance-2.5) hit two of four on the main protocol and 1 of four after its pre-fixed controls; its misses are Table 5 row 1 and the token-grid falsification of g5. Their weakest shared cell (for the two strongest systems; the weaker systems bottom out elsewhere) is Seedance-2.0, where the co-trained backbone scores 0.401 on fifteen videos (CI [0.227, 0.572], which includes 0.5); we report this as an observation on the smallest cell of the track, not as a localized failure point, and excluding that one generator does not close the inter-system gap (Section 4.3).

D.5 Systems: protocol and per-feature detail

Scored-video counts. The mini split's zero-feature controls score exactly 0.5 (the control stated in Section 4.2). Scored-video counts on the same GenVidBench Pair2 protocol differ across rows because decode failures and scoring dropouts differ per system: the manifest carries 67854 files (the full official test pool, used by the duration rules of Section 2.1); the feature bank (PM-C) evaluations score 67853; PM-A and PM-B score 67824 and 67828; and every paired analysis in this paper uses the intersection, $n = 67822$; the two backbone rows scored 67824 and 67828 videos respectively. Each count is the honest denominator of its own row; none is a subsample chosen for effect. The `trained_on` field of the earlier GenBuster evaluation JSONs predates a provenance correction: it records GenVideo-100K

for PM-A and PM-B, whose training manifests (committed, 70,985 Pair1 rows) and training scripts are GenVidBench Pair1. The manifests and `scripts/evaluation/eval_genbuster.py`, corrected at source, are authoritative; all numbers in this paper are verified from the committed artifacts, and the stale field is metadata only.

TSM normalization: divergences, construction, and controls. Our retraining (Kinetics-400 initialization, fine-tuned on the official Pair1 train split, 8 frames at 224^2, final-epoch checkpoint, single seed) reaches its native mean Top-1 on the full 67854-video official test pool. The divergences are declared rather than hidden: the benchmark paper publishes no fine-tuning schedule, ours is a standard recipe at five epochs, and the run is single-seed; nothing was tuned toward the published number or toward any normalized outcome. The retrained run is not a bit-exact reproduction but a different training draw, and its per-column profile disagrees with the published row even where the means agree: ours scores 95.88% on Mora and 83.50% on SVD against the published row's 70.37 and 54.70, and 78.29% on the real column against its 96.76, an anti-correlation that means the container claim is made for our retrained detector, not attributed to the published row. The four normalization directions in full, between the CogVideo endpoint 58.32% and the MuseV endpoint quoted in the main text: Mora 45.81% and SVD 34.81%. One provenance note on the anchor itself: the v1 arXiv version of this table printed a higher TSM headline (76.40) that its own per-source row does not reconstruct to, a non-reconstructing headline of exactly the kind Section 2 documents, corrected in v3 to the value that row means to. Construction: duration cannot be equalized by subsetting this benchmark; only 142 of the 13416 reals fall inside the 1–5 s span the fake columns occupy, so in four separate passes every real is re-encoded to one fake column's measured median container profile: center-trimmed to its duration, center-cropped to its aspect and scaled to its exact resolution, resampled to its frame rate, h264-encoded at its median bitrate (13415 of 13,416 reals per column; one encoder failure). The fakes-to-real direction is not performed: fakes are 1.1–4.1 s and cannot be extended to 11 s without synthesizing content, so normalization runs in the one direction the released data permits. Fake scores are bit-identical across columns for 99.8–100% of fakes (the remainder is floating-point summation noise from batch composition). Column AUCs fall from 0.7937–0.9447 natively to 0.6116–0.8368 normalized ($n = 13369$ fakes against 13415 reals on the MuseV column, whose AUC falls to 0.5517, a fall that leaves weak but statistically nonzero residual discrimination at this sample size). The bitrate targets are each fake column's own measured median (0.52–14.2 Mbps across columns), and the MuseV profile's 1.18 Mbit per frame is 20.9× the reals' native 56.7 kbit per frame, so gross quality loss is an unlikely carrier of the collapse. The same-side control behind Section 4.1's measured split, in full: re-encoding the reals to their own recorded profile (1280x720, their own median bitrate, no fps move) costs -2.41 points of real-class Top-1 (78.29 to 75.88, paired interval [-2.87, -1.95], permutation $p = 0.0002$), and adding the real side's own duration axis (the median-duration trim) moves it a further -0.19 points [-0.66, 0.27], indistinguishable from zero; the remainder of the fake-profile collapse is carried by the moved fake-side container axes. Single seed, this exhibit's retrained detector only; the re-encoded MP4 trees stay server-side, regenerable from the committed script and logs. And the no-re-encode duration-band control is real-side underpowered by construction (the band keeps every fake but only 106 reals), so its in-band AUCs (0.6165–0.8564) are completeness, not estimates.

Released-checkpoint check: BusterX++ under the same intervention. The container claim is not an artifact of our retraining: the released BusterX++ checkpoint (Wen et al., 2025) (an MLLM, Qwen3.5-4B with an RL-trained reasoning chain, lineage and training corpus both disjoint from the Table-7 classifiers) loses real-class accuracy under the identical intervention. It emits verdicts rather than scores, so the readout is verdict accuracy under a parse rule committed before any output was seen (unparseable verdicts count as wrong for their true class, worst arm 129 of 2000; no AUC anywhere), on a declared paired subsample: $n = 2000$ reals drawn seed-42 alphabetical and committed before inference, scored native and in all four equalizations, plus 1500 fakes per column. Native real-class accuracy 79.50% falls to 61.45% under the CogVideo profile, 72.60% (Mora), 73.75% (MuseV), and 73.80% (SVD); the pooled paired drop is 9.10 points, CI [7.42, 10.75] (2,000 bootstrap resamples, seed 42; sign-flip permutation 5,000 draws, zero exceedances), every per-column interval excluding zero. The point drops (5.7–18.1 points) sit at the low end of the retrained TSM range (20–48 points), the direction identical. The model runs under its released settings (its own prompt verbatim, greedy decoding, 2.0 FPS frame sampling, 64 vision tokens per frame, 8,192-token context, revision-pinned checkpoint with per-file sha256); two released-protocol notes matter. First, its 2 FPS sampling leaves the ultra-short MuseV and SVD fake columns near-blind as-released

System (all zero-shot, fixed 0.5)	mini AUC	id AUC	ood det.@0.5	wild det.@0.5
PM-A (Pair1-trained, probes off)	0.865	0.771	94.8%	75.3%
v7e1 (GenVideo-trained)	0.779	0.824	68.2%	56.0%
PM-C (motion-stat expert)	0.724	0.668	94.6%	62.0%
PM-B (plain baseline)	0.709	0.721	48.4%	16.0%

Table 13: Zero-shot transfer onto GenBuster under the container-clean protocol, single seed, train-label access only, containers verified uniform across all four columns ("mini" is the 9,800-video train split, 4,800 real OpenVid-1M (Nan et al., 2024) / 5,000 fake, widest reported CI half-width ± 0.0103). Two disclosures govern the reading: (i) the "ood"/"wild" detection rates are the fixed-0.5 operating point against the id reals, so that column ranks calibration, not detection; (ii) the checkpoint rule is asymmetric by system (PM-A/PM-B final-epoch, v7e1 stage-1 epoch-1), as stated in the text.

(fake-class verdict accuracy 37.87% and 23.20%; fakes are untouched in both conditions, so this is the released protocol on as-released columns, not the intervention). Second, if any Pair2 real source overlaps BusterX++'s GenBuster-200K real pool, real-class accuracy is inflated in both conditions equally, which is conservative for the collapse direction. Raw per-video verdicts, committed draw lists, and the checkpoint manifest: `results/paper/busterxpp_container_collapse/`.

Clean-anchor detail. The anchor's training run: full train split, 202,570 rows, four epochs, final-epoch checkpoint fixed in advance, and 0 filename overlap with any test track; the published VideoMAE row it improves on is 79.1 ID and 76.4 OOD, and its reference-pool AUCs on our protocol are 0.9897 id and 0.9691 ood before the wild value in the main text. The margins, in full: the anchor sits $+14.1$ and $+9.9$ points over the published VideoMAE row (cross-recipe, not controlled), with BusterX still ahead on OOD at 87.8. The register the main text compresses to a pointer: in-domain training on the benchmark's own 200K clips does not survive the benchmark's own 2026 track; this is the wall Section 4.3 measures at per-generation scale.

Clean-anchor divergences. The anchor's cross-recipe margins carry declared divergences: we sample eight frames at 224^2 where the benchmark's evaluation samples 2.0 FPS ($\sim$10 frames) near 256^2, their training schedule is unpublished, and the run is single-seed with no variance estimate. The anchor's id-track balanced-accuracy interval ([92.1, 94.3]) excludes the published VideoMAE row outright, BusterX stays ahead on wild (77.5), and the anchor's wild pooled AUC carries interval [0.777, 0.853]. The anchor sits above the published DeMamba row (82.0 ID, 80.4 OOD) and above BusterX's published ID row (86.7).

GenBuster track detail. The intro's system-level headline numbers, in full: PM-A id-track AUC 0.771 (also Table 13) and the attribution triage layer at macro-F1 0.889 after its own leak was closed at source (both stated in Section 4.4). The mini split pairs $n = 5000$ 2024-era open-source fakes (CogVideoX, EasyAnimate, HunyuanVideo, LTX-Video) against 4800 OpenVid-1M reals, and PM-A's mini AUC interval is [0.858, 0.871] (video-level bootstrap). The id-track comparison: PM-A 0.771 against v7e1's 0.824, paired delta -5.32 points, CI $[-7.52, -3.05]$, with PM-A calling 40.9% of id reals fake; on the wild track PM-A's interval is [0.654, 0.744] and v7e1's [0.826, 0.890]. The id-track paired PM-A-versus-PM-B delta carries interval $[2.88, 7.30]$. The mini-split paired PM-A-versus-PM-B delta is computed over $n = 9800$ paired videos (2000 resamples), the id-track delta over $n = 2000$ paired videos, and the PM-A-versus-v7e1 delta over $n = 2000$; the mini-split replication is not independent evidence (the same two single-seed training runs, loader, and decision rule are applied to new videos). For any reader who insists on bracketing the zero-shot gap against the in-domain baselines anyway: our zero-shot PM-A id-track AUC is 0.771 against their in-domain accuracies, a comparison that mixes label access and metric at once, which is why the rows are kept apart. PM-A's id real accuracy is 59.1% against v7e1's 92.0%. On the ood track, one construction disclosure: Pika is a product line whose earlier open model appears in both training sets (13,484 Pair1 clips for PM-A; 9,515 GenVideo-100K clips for v7e1), so the ood column is zero-shot with respect to clips but not to product family; a partial-MD5 hash check found zero byte-identical videos, and Pika is PM-A's *lowest* ood cell (AUC 0.745), so the overlap does not inflate the pooled number. With that stated: PM-A detects 94.8% at the fixed threshold while v7e1 detects 68.2%, even though their reference AUCs against the id reals are statistically indistinguishable on overlapping marginal

intervals (0.897 CI [0.883, 0.910] versus 0.895 CI [0.881, 0.909]), a statement about intervals, not a paired test. PM-A's ood advantage is an operating-point fact read against its 40.9% real false-positive rate, not a discrimination fact; the same caveat governs the two extreme cells of the det.@0.5 column: PM-C's 94.6% and PM-B's 16.0% are the two degenerate corners of the same fixed threshold, since PM-C calls 69.6% of id reals fake and PM-B misses 63.6% of id fakes; a reader comparing that column across rows is ranking calibration, not detection. On the wild track PM-B's reference AUC falls to 0.625 (CI [0.573, 0.671]), and PM-C's wild interval [0.417, 0.525] includes 0.5 at $n = 150$. The wild-track point estimates the main text demotes here: PM-C at 0.471 and v7e1 holding the suite's highest wild ranking at 0.860, on a track shipping 150 fakes. Between the 2025 and 2026 tracks v7e1's detection rate falls from 68.2% to 56.0%, while the ranking order between the systems flips twice (v7e1 leads id, ties ood, leads wild).

The three official tracks, in full. The three official tracks add a distribution axis: on the id track PM-A's AUC is *below* that of our GenVideo-100K-trained backbone (v7e1) while PM-A's detection rates ride a threshold shifted toward fake: ranking and operating point are different questions with different leaders (checkpoint asymmetry, deltas, intervals, disclosures: this appendix). The wild track decays or inverts every system but one, ranking before detection: PM-A's reference AUC (fakes scored against the id track's 1,000 reals, the protocol's only cross-track reals) falls, id 0.771 to 0.702, the flow-statistics bank expert's (PM-C; the Section 2.3 feature bank deployed as a system) from 0.668 to below chance, while v7e1 alone rises, 0.824 to 0.860, the suite's highest wild ranking.

Attribution detail. The five-class classifier of Section 4.4 (5-fold grouped CV over 1388 videos) rises to 0.933 on the four-generator fake-only problem and 0.98 for CogVideoX-versus-Wan-2.2 at a uniform container; under grouping that separates near-duplicate videos across folds it reaches 0.8855; its finest separations are between adjacent generations of one product line (Seedance-2.0 versus 2.5 is the largest residual confusion), which bounds how fine the triage granularity is. The operational reading in full: attribution tells an operator *which generator generation* a suspect video belongs to, which selects the direction-calibrated bank of Section 3.3 and, at the 2026 wall, which family's labels to buy for the fine-tune step of Section 4.3; it is not a certified detector under the gates: its margins ride the priors Section 3.3 falsified and its feature set partially aliases container protocol (macro-F1 drop under protocol-matching among the supplementary values below), which bounds the triage granularity.

Fusion rows and the rule artifact. Relocated from Section 4.4: on the container-clean benchmark, solo PM-A's interval is [0.858, 0.871] against three-system fusion's pooled AUC 0.8375 with disjoint-lower interval [0.830, 0.845] (paired delta versus solo $[-3.27, -2.12]$): adding the weaker systems subtracts, with no decision rule available to blame. The official-protocol fusion rows we once reported fail for an additional reason, which is a finding about the decision rule rather than about fusion: normalizing each system's scores over the entire test pool calibrates silently to the pool's 4:1 prevalence, and replacing only that rule with an inductive per-video one moves the same scores, in official Mean Top-1 points on GenVidBench Pair2, from 64.17 to 45.44 (two-system) and from 65.09 to 53.28 (three-system), an 18.73- and 11.81-point drop attributable to the rule alone. We do not re-report those rows as exhibits because the prevalence-calibration mechanism they demonstrate is a test-pool-level rule artifact, the same family as the constant-predictor floor and official-versus-balanced inversion of Section 2.2 rather than a gate-falsified detection claim.

Wall: k-shot breach criterion and bitrate native references. Whether 0.72 also bounds fine-tuning was the next pre-registered question, with the breach criterion fixed before any run: past the oracle's 0.7203, interval clear of the calibrated 0.712. The bitrate control re-encoded the full collected pool (the fresh encode also repairing the five native decode failures), where the native pass had scored 995 of them; its re-scored reference AUCs are stated in Section 4.3, and they leave PM-C and PM-A within two thousandths of their native 0.348 and 0.153 (Table 6), with intervals [0.3251, 0.3726] and [0.136, 0.1681].

Wall: corpus-not-era detail. The WaveRep wall cells, in full: WaveRep G1 and G4 (Corvi et al., 2025), zero-shot, read 0.369 (inverted) and 0.547 on Seedance-2.0 while holding Seedance-2.5 (0.722, 0.801). The clean-corpus DeMamba row's provenance, in full: the architecture and clean-room recipe are those of the Pair1 row, seed 42, retrained on the container-standardized GenBuster-200K

train split (`data/manifests/genbuster200k_train.csv`). The run's summary carried a stale `trained_on` field copied from the Pair1 template; it is corrected at source with the correction documented in the released directory, and every numeric cell was recomputed exactly from the committed per-video prediction files. Its threshold column: detection at 0.5 is 0.042 on Seedance-2.0, 0.503 on 2.5, and 0.16 on the wild mix: the ranking transfers across the corpus change, the operating point does not. The anchor row's seed-42 wall cells survive only as quoted references inside the seed-43 artifact's provenance block (0.889 and 0.921), so the table prints the seed-43 replicate, whose per-video reference and fake scores are both committed and recompute its cells exactly. The controlled-retraining rows (the PM-B container-normalized retrain at the 540p CRF-23 encode and its 1024-resolution replicate; preregistration documents committed under `docs/prereg/`): the retrain's below-chance permutation p is 0.0 as committed (no permutation draw above chance); the two retrains' reference scores recompute exactly from the committed per-video wall prediction files (both the fake pools and the id-reals reference side), and the 1024-resolution replicate is the in-repo defense against the compression-level reading, at 0.065 [0.0545, 0.0765] and 0.0195. The container-normalized retrain's intervals, in full: 0.056 [0.047, 0.066] on Seedance-2.0 and 0.016 [0.0095, 0.0239] on 2.5. The seed-43 replicate values behind the main text's summary: the retrained TSM reads 0.129 [0.112, 0.146] on Seedance-2.0 and 0.086 [0.0614, 0.1138] on 2.5, the clean-room DeMamba 0.256 [0.232, 0.280] and 0.276 [0.250, 0.303], every interval entirely below chance, so all three inverted lineages hold below chance at both seeds tested, a robustness check rather than a variance estimate (Table 6). The content channel, in full: the corpus carries content as well as containers, and the wall readout certifies nothing: swapping only the fake pool's content moves PM-A on CogVideoX from 0.749 to 0.293 (paired -0.456), and a PM-B trained on a 2,000-clip LTX corpus reads 0.975 on the wall yet collapses off-corpus, a letterbox control cutting up to 0.298 of the wall reading (the GenBuster-native corpus holds everywhere). The mixed-era retrain, in full: corpus composition carries the same signature: retraining with every batch balanced 1:1 between the 2024 and 2026 fake pools recovers the held-out 2026 pools to 0.829 and 0.803, coverage of the new class at a paired -0.145 AUC (p=0.0005) against the 0.974 sequential fine-tune, not a representation carrying both classes. The mixed-era retrain's declared asymmetries: the 100 fixed-split 2026 clips repeat about 1,080 times each across four epochs (forced by the 1:1 batch balance), the real-side exposure differs from the sequential reference run, and 69.2% of its own training clips sit below the 0.5 threshold (66.7% on 2.5); its GenBuster-mini 2024 AUC is 0.7045 [0.6938, 0.7145], within the committed PM-B interval; its paired deficit against the sequential fine-tune carries the interval [-0.165, -0.125]. The content-channel control's other systems: PM-B moves by $+0.0011$ [-0.0324, 0.0346] (interval spanning zero), PM-C by $+0.0487$, v7e1 by -0.1964 (PM-A's paired interval [-0.488, -0.425]; PM-A's own cell intervals [0.718, 0.779] collected and [0.270, 0.315] caption-matched); the caption-matched pool is fully padded (73/121 static-clone frames), and the raw-container PM-A diagnostic before normalization is 0.525 (the padding plus re-encode footprint and the content-container split of the remainder are not separately identifiable in-protocol; the run's REPORT declares both). The 2k-corpus rows in full (Seedance-2.0 / Seedance-2.5 / wild / mini / letterbox-controlled Seedance-2.0): GenBuster native 0.918/0.932/0.742/0.9935/0.834; LTX image-to-video 0.975/0.980/0.482/0.548/0.687; LTX text-to-video 0.967/0.971/0.476/0.553/0.669; every cell carries a committed interval, the letterbox control (center-crop both sides to 1024×576) is computed on the Seedance-2.0 face only, and the wild cells' intervals span chance (the wall fakes are roughly two-thirds letterboxed, which is why the control exists). WaveRep protocol notes: official demo verbatim, all 121 frames, the same 1000 bench-id reals, with the declared 10-bit-HEVC demux deviation applied identically to both classes; its detection at 0.5 is 0.339 (G1, Seedance-2.0) and 0.847 (G1, 2.5).

Wall k-shot defaults. Default-direction fusion scores 0.2974, an anti-detector, and ten calibration clips correct 83.5% of feature directions (the k=100 value is in the main text).

Topology probe: per-feature numbers. The probe scores the same 995 Seedance-2.0 fakes against the same 1000 reals, dividing each of the bank's flow statistics by the motion magnitude the statistic is computed over, which removes the motion-amount axis. First stratum's interval, in full: curl mean per unit motion stays inverted (point estimate 0.449; permutation p among the supplementary values below; CI [0.424, 0.474]). Second stratum, in full: divergence and curl dispersion separate at 0.716 and 0.714 unnormalized, but divided by motion-magnitude dispersion the divergence statistic inverts to 0.336 (CI [0.312, 0.360]) and the curl statistic to 0.368 (CI [0.344, 0.392]): what survives the flip on the dispersion side is motion-amount information, not structure in how motion varies. Third

stratum, in full: the bank's te_mean is dead at 0.511 raw but separates per unit motion at 0.643 (CI [0.619, 0.669]). The wall also deepens within the family: on Seedance-2.5 even the motion-amount prior no longer separates (motion-magnitude mean 0.481, CI [0.423, 0.539] spanning chance).

Wild k-shot diagnostics. Calibrating the bank with wild labels at k=10 and k=25 moves fused AUC from its zero-calibration value to 0.4119 at k=10 and 0.411 at k=25, from a zero-calibration value of 0.4134: flat. Worse, the calibrated directions agree with the wild truth on only 19.1% of features (against 17.65% before calibration): old-family calibration does not recover the directions, it reinforces the wrong ones. The wall-side carry values: on Seedance-2.5, directions calibrated on Seedance-2.0 fuse at AUC 0.5555 (69.7% direction agreement), and on the wild mix even the full-label direction oracle fuses to only 0.5975.

Fine-tune label counts: intermediate values. Fifty and one hundred fake labels take PM-B to 0.990 and 0.974, and v7e1 at one hundred to 0.965; at one hundred PM-A reaches 0.996 while PM-C stays at 0.339, and PM-A's Seedance-2.5 transfer holds 0.995. The k=10 seed dependence behind the main text's summary: seed-42 reads 0.831 and seed-43 0.752, both past the 0.7203 ceiling, and at k=100 the seed replicate spans 0.956 to 0.968 on the 2.5 pool; the Seedance-2.5 transfer rows are to be read against the calibrated bank's cross-generation 0.5555. The k=10 point is seed-dependent (both seeds' values past the oracle ceiling), while from k=100 upward the seed replicate moves values by at most 0.012 on the 2.5 pool.

Additional intervals and values. Wall pool sizes: the Seedance-2.5 pool is 185 collected clips. The k-shot ladder's fused AUC: 0.7125 at k=10 and 0.712 at k=100 (the main text's value). The curl-topology permutation p: $p = 0.0003$. PM-B's ten-label interval: [0.808, 0.851]. The wrong-family point estimate: 0.245; the paired interval: [-0.369, -0.314]. The spillover interval: [0.698, 0.786].

Spillover intervals and the operating-point fall. PM-A's k=895 wild row: 0.771 (CI [0.720, 0.817]); the paired PM-B gain interval is [0.071, 0.173] (two-sided bootstrap-tail $p = 0.0005$). The operating-point fall behind the main text's overestimate: PM-A's detection rate drops from 94.8% on the 2025 ood track to 75.3% on wild. The priced ladder's across-generation entry, duplicated for the ladder's own table: PM-A at 0.995.

Fine-tune spillover: per-generator heterogeneity. The fine-tuned PM-B's k=895 wild row in full: pooled AUC 0.744 at 8% detection of the wild mix against 96% on its held-out family. The wild mix carries 150 clips, 15 of them the fine-tune family itself and 20 the same brand's previous generation. The spillover is family-heterogeneous: PM-A ranks Sora-2 at 0.900 and Kling-2.6 at 0.848, yet Seedance-1.5-Pro, the same brand's previous generation, at only 0.568 (PM-B) and 0.519 (PM-A). The family boundary that labels must match is per generator generation, not per brand.

Supplementary values (forensics). Oracle pool size: 67854 test files. Flagship-table floor: the 6M-protocol all-fake constant is 85.71, within 0.24 of DeMamba's published 85.47. The bank's official-pool AUC: 0.7656. In-band interval: [0.6414, 0.7487]. E2 real accuracy at the fixed threshold: 4.30%; E2 pooled AUC under the reweighting: 0.7345. Clean-control verification: 12950 files probed with 0 probe errors.

Supplementary values (wall). GenBuster paper baselines (in-domain): DeMamba 80.4, VideoMAE 76.4, BusterX 87.8. k-shot direction diagnostics: default directions match truth on 17.65% of features, k=100 calibration on 91.5%. The ten-label fine-tune: 0.831. The spillover paired gain: $+0.120$ over zero-shot PM-B on identical videos. The reference-swap cells behind the main text's summary, verbatim: PM-A at 0.2445 [0.2244, 0.2657], the clean-room DeMamba at 0.2785, the retrained TSM between 0.0395 and 0.2189, PM-B at 0.5725, v7e1 at 0.3529. The operating-point overestimate: 19.5 points between the 2025 and 2026 tracks. The 2024-era calibration curve: from an anti-detecting default of 0.181 to 0.812 (95% CI [0.784, 0.831]).

Supplementary values. The honest-leaderboard exhibits in full: PM-A and PM-B sit below the 80.00 all-fake constant and the 99.83 duration if-statement. TSM retraining margin: 6.29 points over the published row; MuseV real-accuracy point delta -47.51, interval $[-48.39, -46.66]$. Anchor wild

track: 65.0 accuracy at fixed-0.5 detection 31.3%. Attribution container-alias: macro-F1 0.952 to 0.835 under protocol-matching. Two-system fusion: 0.8201. Bitrate control: fakes 2.232 Mbps vs reals' 1.463 Mbps natively; after re-encoding all 1000 files, fake mean 1285.7 kb/s against the reals' 1460 kb/s.

The priced ladder, re-tabulated. The closing ladder's steps with their values: zero-shot inverted at 0.151; wrong-family adaptation at 0.245; frozen-bank calibration capped at 0.712; in-family fine-tuning at 0.996 (0.998 at k=895), with off-family spillover 0.744 at 8% detection. The mini split's probed count: 9800 files, all passing the container probe.

The motion-matched cross-source re-read, in full. Preregistered before any motion feature touched the cross-source pools (`docs/prereg/motion_matched_cross_source.md`, frozen decision rule: a flip is motion-robust if its matched estimate keeps its side of 0.5 with the interval excluding 0.5 in at least 150 of 200 subsample draws). Real pools were binned into ten quantile bins of mean flow magnitude (the g2 statistic, frozen RAFT estimator, 24 frames per clip) and subsampled to the fake pool's bin proportions (seed 42, 200 draws, bootstrap interval per draw); committed per-video scores, no model inference. Coverage was 100% on every pool (kill criterion: abort below 90%). The pools do differ in motion (median flow magnitude: OpenVid 1.8429, HD-VG 1.3694, Vript 1.6823, against the Seedance-2.0 fakes at 4.4443), which is exactly why the control matters. Matched medians for the primary panel (all against Seedance-2.0 fakes; unmatched values are the H/V columns of Table 2): DeMamba-Pair1 0.7272 (HD-VG), 0.7479 (Vript), 0.1992 (OpenVid); TSM 0.6796, 0.6608, 0.1787; PM-A seed 42 0.5533, 0.5743, 0.1536; PM-A seed 43 0.5636, 0.5867, 0.1037; PM-B seed 42 0.5684, 0.5667, 0.5218; PM-B seed 43 0.5408, 0.5715, 0.2405; VideoMAE-GB 0.5846, 0.6039, 0.8746; DeMamba-GB200K 0.4452, 0.4949, 0.9255. Under the frozen rule three cells lose robustness, all borderline holdings near chance: PM-B seed 42 on OpenVid (10 of 200 draws), PM-B seed 43 on HD-VG (95 of 200), and the corpus-retrained DeMamba's HD-VG collapse (145 of 200, point estimate unchanged); every inversion and every clear holding survives, most strengthened. The motion-magnitude prior itself, matched, collapses to chance on every reference family (0.5038, 0.5099, 0.5428): the pairings are not readable off motion amount. Matching cost: the achievability-capped matched subsets use 181–222 reals per draw.

The 2024-era calibration curve, in full. On 2024-era survey domains the label budget works: default directions anti-detect (the k=0 value is among the supplementary values above; it is deterministic, so no interval attaches to it), ten calibration clips restore the mean listed above (bootstrap over the twenty calibration-draw repeats per domain, then the domain mean; worst domain 0.749), and twenty-five clips 0.832 (CI [0.831, 0.833]). The flip risk that budget covers is concrete: the same bank statistic, divergence variance, separates at AUC 0.822 on Wan-2.2 in one direction and 0.926 on Seedance in the opposite one (signed AUC 0.178 versus 0.926); a detector deployed with the wrong direction is not weakened but inverted.

D.6 Related work: extended discussion

The compressed positioning of Section 5 omits the following per-lineage detail. Corvi et al. (2025) acted on the codec confound before anyone else we are aware of, and their codec-matched retraining result is the earliest published instance we know of a container axis deciding a comparative claim in this subfield. Within the detector lineages: L3DE (Chang et al., 2025) in particular evaluates 3D consistency through frozen motion, depth, and appearance estimators, close to the frozen RAFT-plus-depth probes our gates later caught riding the motion-amount prior; the survey's flow-magnitude prior, token-grid family, and spatial-spectral classics are instances of these lineages, and the verdict of zero universal survivors is a statement about their frozen-estimator members, not about any one implementation: learned, augmentation-trained members of the same lineages, including the certified WaveRep (Corvi et al., 2025), sit outside the survey population (Section 3.2). On the decay genre: Wang et al. (2020) named the image-domain version (detectors winning "for now"), Cozzolino et al. (2023a) documented the GAN-to-diffusion transfer failure, and Thakur et al. (2026) name "generator shift" outright; our addition is the sign inversion, which contradicts VidAudit's unconditional under-motion prediction (Cakiroglu et al., 2026) while realizing, as a measurement, the obsolescence Michels et al. themselves forecast for the prior (Michels et al., 2026). In the image domain, the same container-bias layers are established prior art: JPEG and resolution biases in generated-image

detection datasets (Grommelt et al., 2024), dataset misalignment and its repair (Choi et al., 2025), and content-matched bias-free synthesis (Guillaro et al., 2025). In measurement science: Torralba & Efros (2011) showed that vision datasets are recognizable from their samples alone; Beery et al. (2018) traced a deployed recognizer's errors to a spurious correlation between species and camera-trap location; Lapuschkin et al. (2019) unmasked Clever Hans predictors by attributing decisions to image regions the task should not depend on; and Geirhos et al. (2020) framed the general mechanism: networks solve the objective given, and any feature correlated with the labels that is cheaper than the intended one will be used.

Displaced from the compressed Section 5 (page budget): VidAudit's clip-length existence proof (LOGO AUC 0.998) isolates the phenomenon our per-video ffprobe rows translate into official metric points; Michels et al. (2026) certify WaveRep (Corvi et al., 2025) across the 5 datasets they evaluate, a learned, augmentation-trained detector outside our survey's frozen-estimator scope (Section 3.2), so our zero-survivors verdict and their certification do not conflict; the appearance-axis, model-internal, and physics- or coherence-oriented lineages (Ojha et al., 2023; Cozzolino et al., 2023b; Corvi et al., 2025; Wang et al., 2023b; Liu et al., 2024; Internò et al., 2025; Zheng et al., 2025; Chang et al., 2025; Ma et al., 2024), the last the lineage our own prior submissions belonged to, define what the field counts as a detection signal, and our survey treats their frozen-estimator members as the population under audit; the field's own signal-space map (Hou et al., 2026) runs no falsification battery, precisely the gap our survey occupies; and image-domain container-bias layers, content-level quality shortcuts (Fang et al., 2026), and generator attribution (Kundu et al., 2025; Grommelt et al., 2024; Choi et al., 2025; Guillaro et al., 2025) are complementary prior art we use rather than compete with (Section 4.4).

D.7 Extended limitations and conclusion

An earlier revision compacted the main text's Limitations to one clause per caveat category and its Conclusion to a single sentence; the full forms follow, with the wall's six protocol caveats and the fine-tune caveats consolidated from the two earlier extended statements.

The full statement of Section 6. The audit instrument is itself a measurement, and it carries the threats we list here rather than the ones already controlled. Every training run behind every system row is single-seed except where a second seed is stated, so every between-system margin in Section 4 is a difference between two draws (the wall ladder's seed replicate and the corpus-standardized pair's multiseed multiref retest are reported in Section 4.3 and Appendix C.6); and the evidence chain behind the direction-shift reading spans three days between the direction-instability commit and the wild-track evaluation: temporal precedence, not a registered prediction.

The 17-family survey's domains share one real pool: the local CogVideoX, Wan-2.2, and Seedance domains draw their real videos from a single pool (97% measured overlap), so a feature's learned direction is defined relative to one real pool's capture signature as much as to its generator; and one survey family remains queued rather than tested, so the survey count of 17 tested families is not the count of families that exist. GenBuster is the only evaluation whose real pool is independent of ours, one reason its tracks carry the deployment-relevant claims. Those tracks are asymmetrically constructed, and the paper's most deployment-relevant result rests on their smallest subset: the wild track ships 150 fakes at 10–20 per generator, and neither it nor the ood track ships real videos of its own: every reference AUC on both borrows the id track's 1000 reals, a mini-instance of the real/fake asymmetry Section 2 audits in the field's benchmarks. Results conditioned on the msrvtt real pool, including the static-noise bank member of Section 3.2, inherit that domain's flawed 3 fps staircase protocol and are labeled as such wherever they appear.

The wall measurement's six declared caveats in full: (i) the fakes are re-encoded at CRF-18 (2.232 Mbps mean on Seedance-2.0, 1.801 on 2.5) against reals at 1.463, so bitrate sits deliberately outside the container gate and the matching control of Section 4.3 closes the rate axis only: an encode always perturbs codec internals as well, and the source-quality difference between generator output and camera footage is not equalizable by any re-encode, so pixel-level compression cues remain an honest residual confound of any re-encode protocol, GenBuster's own included; (ii) clips are truncated to their first 121 frames, so late-video evidence is discarded; (iii) the 1024^2 target is reached by letterbox padding, never cropping; (iv) the reference real pool is the id track's 1,000 reals, reused by every track in the suite, so the wall inherits the same cross-track real reuse, and

the reference-swap check stays within that pool's OpenVid upstream, so robustness to a reference from a different source remains untested; (v) the label-peeking rows of Section 4.3 (the direction oracle, the direction-free per-feature AUCs, and the magnitude-normalized probe) use full-set labels and are mechanism characterization only, never system rows; and (vi) the Seedance-2.5 pool is an order of magnitude smaller than the 2.0 pool ($n = 183$ scored of 185 collected), so its intervals are correspondingly wider.

The fine-tune caveats in full: within-2.0 numbers may ride same-prompt near-duplicates (the 2.5 transfer rows are the memorization-free estimates and carry the generalization claim); the PM-C $\times$100-upweight variant was declared after the base result was seen, is post-hoc exploratory, and moves the bank only to 0.500; the control run's id-era fakes were drawn from the full GenBuster train split rather than the mini split, whose fake pool had been partially deleted on disk, a declared substitution, with treatment and control sharing identical real clips and label count, so the substitution moves which clips were drawn, not the comparison's structure; the fine-tune and reference reals are both OpenVid-1M derivatives (filename-disjoint; source-level overlap cannot be excluded); and the k=895 extension rotates the committed split: the 100 former fine-tune clips become the held-out test (direction-diagnosed historically but never trained on by the k=895 model), so its within-2.0 held-out numbers rest on n=100 with correspondingly wide intervals, while its Seedance-2.5 and wild rows rest on the full 183- and 150-clip pools.

The k-shot flip mechanics: under the Mann-Whitney model of the direction estimator (the calibration count U follows $\mathrm{Binomial}(k^2, \mathrm{AUC})$ over the $k \times k$ real-fake pairs), the flip probabilities stated in the boundary list below follow, and the committed router experiment shows the corresponding empirical behavior, in which a flipped member inverts its rank contribution under median fusion; a ten-clip budget is therefore adequate for the strong bank members and unsafe for weak ones.

Four further boundaries complete the list. The container-normalized re-scoring of Section 4.1 now carries its complementary same-side control (reals re-encoded to their *own* recorded profile at matched quality: the measured quality share and the zero duration share stated there), but the split is measured for one exhibit's retrained detector, not for the other Table-7 systems. The field survey of Section 2 is a sample of convenience (selection rule: the benchmarks we hold locally plus every prominent benchmark whose paper documents its own container columns), and the ten-family counts inherit that denominator. The 17-feature bank plays two roles not fully separable: audit instrument behind the metric exhibit of Section 2.2 (84.29 is a metric instrument, not a detector we offer for use) and audited system behind the PM-C rows of Section 4, a circularity we disclose rather than resolve. And the k-shot budget Section 4.3 proposes is statistically fragile at the clip counts we recommend: a feature whose true AUC is 0.55 has its direction estimated backwards with probability 0.1346 at k=10 (0.0168 at true AUC 0.60), the committed router experiment shows fused AUC falling to between 0.199 and 0.2508 in the corresponding empirical draws, and a deployment must choose k against a stated flip-risk budget rather than adopt our working value. The attribution triage of Section 4.4 is container-inclusive by construction (protocol-matching the container axis alone moves five-class macro-F1 from 0.952 to 0.835), so it routes between generator generations but does not certify detection; and the release of per-video audit paths depends on benchmark terms we cannot verify from the local copies alone. GenVidBench ships no license file on disk, and we will release code, extracted features, and prediction CSVs, with per-video paths only where benchmark terms permit.

The full closing statement of Section 6. This paper began with two of our own method claims and the controls that falsified them; it ends with the instrument those controls became. The finding behind it is not that detectors are weak but that the field's measurements do not yet isolate detection: ten families surveyed, most separable by an if-statement on the file container, scored by metrics whose constant floors sit above the published mid-table, with the gap measured inside the benchmark's own retrained TSM baseline and populated by signal families that fail under compression, motion-matching, direction, and resolution controls. The defensible contribution is the instrument: eight gates and two audit modes, every gate with a demonstrated falsification, five of the falsified claims ours; a 17-family survey closing the universal level while populating a per-domain bank; clean-protocol measurements on the one container-standardized benchmark; and a label-priced adaptation ladder, with the code, feature banks, and per-video prediction CSVs behind every number released where benchmark terms permit. The field's task is small and concrete: report container controls beside every accuracy: until a benchmark prints its fps, duration, resolution, and codec per side, states its metric's

constant floor, and survives a zero-feature rule, a high number on it is not evidence of detection, and the field should stop reading it as one.

E GLOSSARY

The recurring terms of the paper, defined once and used consistently. *Container*: the file wrapper (duration, frame rate, resolution, codec) whose header precedes every pixel; a *zero-decode rule* reads container fields only and decodes no frames. *In-band*: inside the duration window in which the real and fake sides of a benchmark overlap. *Real side / fake side*: the two halves of a benchmark's real/fake split. The paper's systems carry self-explanatory names in the main text (the physics-co-trained backbone, the matched plain baseline, the motion-statistics expert, the GenVideo backbone, the clean anchor, the standardized-corpus DeMamba, the retrained TSM); the appendix tables keep short codes as row labels for width, and Table 14 is the legend.

Code (appendix row label)	System
PM-A	the physics-co-trained backbone (ViT-B/16)
PM-B	the matched plain baseline (VideoMAE, identical split)
PM-C	the motion-statistics expert (the frozen feature bank as a system)
v7e1	the GenVideo backbone (trained on GenVideo-100K)
DeMamba (clean-room, Pair1)	the published DeMamba architecture retrained on Pair1
DeMamba (GenBuster-200K corpus)	the standardized-corpus DeMamba
VideoMAE-GB anchor	the clean anchor (in-domain GenBuster training)
TSM (retrained)	the retrained TSM (GenVidBench baseline recipe)

Table 14: Legend for the short system codes used as row labels in the appendix tables; the main text uses the self-explanatory names in the right column throughout.

Bank: the frozen 17-feature handcrafted flow/physics statistic set (the survey families are its relatives). *Gate / battery*: one of eight controls (named in Section 3.1) plus two audit modes; *survey*: the application of the battery to the computable signal families. *Floor*: a constant predictor's score under a given metric. *Official metric*: GenVidBench mean Top-1 under its 4:1 column structure; we always report *balanced accuracy* and *pooled AUC* alongside it. *Wall*: the 2026 Seedance-generation setting in which the tested zero-shot systems trained on the field's confounded corpora fail or invert against the OpenVid-family reference on Seedance-2.0, none with a point estimate reaching 0.6 AUC, and most of our suite inverts or fails on Seedance-2.5 while one published augmentation-trained family retains ranking; measured against that reference, the pairing the Vript re-scoring of Appendix C.5 and the multiref multiseed retest of Appendix C.6 bound. *Ladder*: the adaptation routes ordered by label budget (calibration, in-family fine-tuning, wrong-family control). *Direction calibration*: signing/flipping feature directions with k labeled clips, model frozen. *Label-peeking oracle*: an upper reference that uses test labels, reported only as a bound, never as a system. *Sign inversion / below-chance transfer*: the discriminative direction of a signal flips across generator generations, driving transfer below chance. *Leakage control*: dropping every training real that also appears in the test pool.