TRAINING DATA REDIRECTS EVIDENCE ROUTES IN AUDIO–VISUAL DEEPFAKE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio–visual deepfake detectors can achieve near-perfect in-domain AUC yet degrade sharply under distribution shift, while aggregate accuracy reveals little about why they fail. We introduce the notion of an *evidence route*: the channel through which predictive evidence reaches a detector's decision. We show that this route is a property of the trained detector instance, rather than of its architecture alone, and introduce **AVRouteBench**, a diagnostic benchmark for measuring this. AVRouteBench uses a two-stage intervention protocol: a screening signature based on controlled perturbations such as leading silence, speaker identity, modality skew, and generator overfit, followed by targeted confirmation interventions that directly test candidate routes, for example by breaking audio–visual correspondence while preserving both streams. Two findings emerge. First, training data can redirect the evidence route even under a fixed architecture. With a frozen AV-HuBERT backbone and the same logistic-head architecture, changing only the training mixture moves fake-video/real-audio AUC from 0.495 to 0.980. The effect persists under matched controls for class, manipulation cell, and identity composition, and is reproduced across additional training sources; on two other frozen backbones its size tracks the visual evidence each representation exposes. Controlled interventions further trace these changes to the availability and accessibility of modality-specific label cues. Second, the learned route shapes out-of-distribution failure: whether a detector continues to detect a manipulation depends on whether the evidence it relies on survives the target shift, with audio–visual correspondence providing a detailed case study. Across 19 detector instances from ten families, AVRouteBench extends evaluation from *how accurate is the detector?* to *which evidence does it use, and will it survive a shift?*

1 INTRODUCTION

Audio–visual deepfake detectors can achieve near-perfect in-domain performance on standard benchmarks (Khalid et al., 2021; Cai et al., 2022; 2024), yet their accuracy often degrades sharply under changes in data source, manipulation family, or recording pipeline (Yan et al., 2023; Yermakov et al., 2026). Forward transfer to future generators remains particularly weak (Fontana et al., 2025). This generalization gap is well documented; what remains poorly understood is *why* detectors that appear similarly strong in-domain fail so differently under distribution shift. Current evaluation approaches reduce detector quality to pooled accuracy or AUC, but these numbers do not reveal what evidence the model actually uses to reach its prediction.

A natural explanation is shortcut learning (Geirhos et al., 2020). Audio–visual deepfake datasets are known to contain exploitable cues such as leading silence (Smeu et al., 2025; Müller et al., 2021), speaker identity (Dong et al., 2023), and generator-specific fingerprints (Yu et al., 2019; Corvi et al., 2023). However, we find that shortcut reliance is not a sufficient diagnostic of OOD failure. Across 19 detector instances, no individual shortcut score explains collapse consistently across targets: a detector can rely strongly on a shortcut and still transfer, while a detector that appears clean on that shortcut can nevertheless collapse. In-domain AUC is similarly uninformative. These observations suggest that generalization failure cannot be reduced to a single notion of "cheating"; what matters is *which source of evidence carries the decision*.

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.

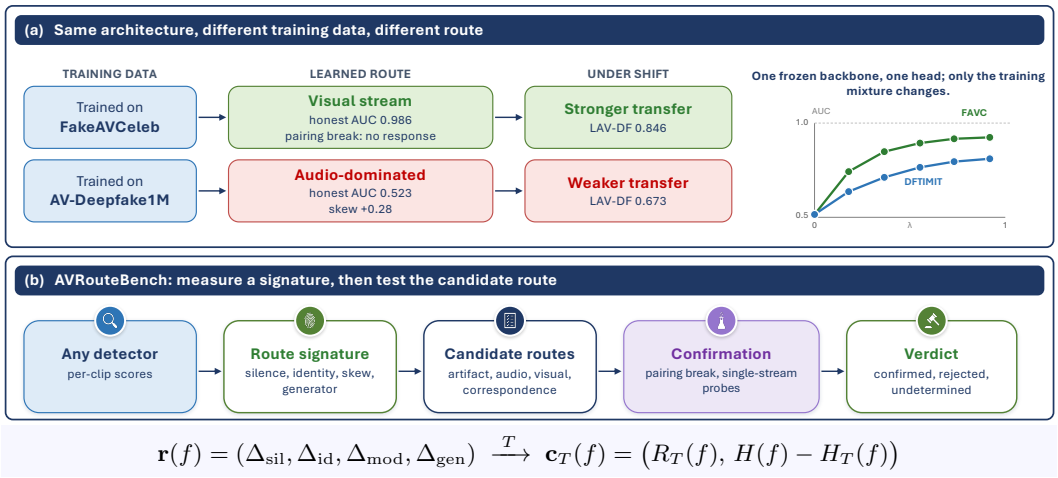
We therefore study the detector's *evidence route*: the channel through which predictive evidence reaches the decision, such as a dataset artifact, a single modality, a generator-specific cue, or audio–visual correspondence. Our central finding is that **interventions can identify the evidence a detector actually relies on, and training data can redirect that reliance within what a fixed architecture allows, changing how the detector fails under shift**. Routes differ systematically in how stable their underlying evidence remains across domains, which yields a simple mechanism: the training distribution shapes the learned evidence route, and the cross-domain stability of the evidence that route reads shapes OOD failure.

We establish this mechanism in three steps. First, we show that aggregate shortcut scores are inadequate because different evidence routes respond differently to different shifts (§3). Second, we intervene on the training distribution while fixing the model architecture. With one frozen AV-HuBERT backbone and one logistic head, changing only the mix of two training sources moves AUC on the fake-video/real-audio diagnostic subset from 0.495 to 0.980, and a published detector with identical architecture and code learns a route that reads the visual stream from one training source and an audio-dominated route from the other (§4). A matched sweep shows that the effect survives fixing class and cell balance and the within-identity real–fake pairing, a same-source intervention traces part of it to the availability of an audio-side label cue, and on two further frozen backbones the size of the effect tracks the visual evidence each representation exposes. Evidence use is therefore not fixed by the architecture alone; the training distribution can redirect which of the available evidence channels the detector exploits.

Third, we ask why routes fail differently. A route transfers when the evidence it reads survives the shift, and what survives is set by the manipulation type of the target: the visual route learned above carries over to visual edits but misses audio edits, which the correspondence route detects. We characterize the correspondence case in detail, tracing its signal to the *level* of audio–visual agreement rather than to synchronization timing, which modern generators often match, and stating where it breaks (§5–§6). We state the stability as an empirical regularity over the evaluated corpora and generators, not as a universal invariant.

To make evidence routes measurable, we introduce AVRouteBench, an intervention-based diagnostic framework. Rather than assigning a detector another scalar performance score, AVRouteBench represents its behavior as a multi-dimensional *route signature*. Each coordinate measures the detector's response to a controlled perturbation of a candidate evidence channel, including leading-silence dependence, speaker-identity dependence, modality skew, and generator-distribution overfit (§2, Figure 1). The signature narrows the route to a few candidates; a direct confirmation intervention, such as breaking the audio–visual pairing while leaving each stream intact, then tests each candidate. The resulting diagnosis characterizes *how* the detector makes its prediction, rather than merely how accurately it performs. A signature-based risk ranking, which the mechanism suggests but which rests on a small family-level sample, is reported as an exploratory analysis (Appendix H).

Contributions. Our contributions are threefold. **(i) Intervention-based route diagnosis.** AVRouteBench characterizes a detector by the evidence it relies on rather than by a single accuracy or shortcut score: a four-axis intervention signature narrows the route to candidates and a confirmation intervention tests them. Across 19 detector instances from ten families, no tested shortcut scalar or in-domain AUC captures OOD failure consistently, and a detector that is clean on every axis can still fail the confirmation step. **(ii) Training data redirects the route under a fixed architecture.** Changing only the training mixture moves FV–RA AUC from 0.495 to 0.980 on one frozen backbone, and the same published architecture learns different routes from different training sources. A matched sweep shows that the effect survives fixing class and cell balance and the within-identity real–fake pairing (the identity set still changes with the corpus), a same-source intervention traces part of it to the availability of an audio-side label cue, and on two further frozen backbones the size of the effect tracks the visual evidence the representation exposes to a linear head. **(iii) The route sets the failure mode.** Transfer depends on whether the evidence a route reads survives the target's manipulation type: visual evidence carries over to visual edits, while cross-modal agreement catches the audio edits that a visual route misses. We trace the agreement signal to its level rather than its timing, and identify where that margin narrows.



$$\mathbf{r}(f) = (\Delta_{\text{sil}}, \Delta_{\text{id}}, \Delta_{\text{mod}}, \Delta_{\text{gen}}) \xrightarrow{T} \mathbf{c}_T(f) = \big(R_T(f),\, H(f) - H_T(f)\big)$$

Figure 1: **Training data sets the evidence route, and the route sets the failure under shift. (a)** The same AVH-Align architecture trained on two sources learns a visual route from one and an audio-dominated route from the other, with stronger and weaker transfer to LAV-DF; inset: one frozen backbone and head, only the training mixture changes (§4). **(b)** AVRouteBench screens candidate routes with signature $\mathbf{r}(f)$, then tests them with intervention T. The confirmation pair $\mathbf{c}_T(f)$ records response AUC and the honest-AUC drop (§2).

Related work. Audio–visual deepfake detectors span supervised fusion (Cai et al., 2023; Oorloff et al., 2024) and correspondence-based methods that compare what is heard with what is seen (Cheng et al., 2024; Feng et al., 2023; Smeu et al., 2025). Cross-dataset benchmarks expose substantial transfer gaps (Khalid et al., 2021; Cai et al., 2022; 2024; Yan et al., 2023), but aggregate accuracy does not identify which evidence a detector uses. Shortcut studies provide more targeted tests: Smeu et al. (2025) identify leading-silence artifacts in audio–visual corpora, while Dong et al. (2023) study identity leakage in visual detection. AVRouteBench combines these tests into a route signature and checks candidate routes with direct interventions. Its fixed-architecture training sweeps further test how data changes the evidence used by a detector. A complementary literature estimates a fixed model's target accuracy from unlabeled target data (Yu et al., 2022; 2025); our primary goal is route diagnosis, with held-out-family risk ranking treated as exploratory analysis. The extended discussion appears in Appendix C.

2 AVRouteBench: Measuring the Evidence Route

AVRouteBench is the instrument that measures the middle link of the mechanism, the evidence route, which accuracy does not see. It has two steps: a signature of four intervention axes that narrows the route to a few candidates, and a confirmation intervention that tests each candidate (Figure 1b).

Definitions. Every corpus is factored into manipulation *cells* by which stream is synthetic: FV–RA (fake video, real audio), RV–FA (real video, fake audio), FV–FA, and RV–RA (genuine). The *honest subset* is the FV–RA cell of a corpus whose audio carries no class signal, so that discrimination requires the visual stream or its relation to audio; *honest AUC* is measured on that cell. *Modality skew* is the RV–FA AUC minus the FV–RA AUC: positive means audio dominated, negative visual dominated. A detector's *evidence route* is the channel its decision flows through: a dataset artifact, a single modality, a generator cue, or audio–visual correspondence. The *route signature* is the detector's value on the four axes below. A *confirmation* reports the response AUC of perturbed against intact clips (0.5 = no response) and the honest AUC when the perturbed clips stand in for the real class.

Route axes. (1) *Leading-silence trim:* the paired AUC change when leading silence is cut under the protocol of Smeu et al. (2025), which removes the matching frames from both streams; because

content is removed too, the raw change is an upper bound on silence reliance, and a duration-matched trailing trim isolates the leading-specific part. (2) *Speaker identity:* the inflation from identity exposure, measured on identity-seen versus identity-unseen controls matched on demographic composition. (3) *Modality skew*, defined above. (4) *Generator overfit:* in-domain minus cross-generator AUC. The three intervention axes are paired-bootstrap changes at the clip level; the generator axis compares independent samples with an independent bootstrap. All are measured on FakeAVCeleb (FAVC), the diagnostic corpus; skew and silence cover all 19 instances and identity 18, with FTCN's identity at its pre-registered value (protocols: Appendix E). These four contrasts form $\mathbf{r}(f)$ in Figure 1: the silence, identity, modality-skew, and generator coordinates, respectively.

Route confirmation. A detector can be clean on every axis and still not read the evidence its profile suggests (§3), so each candidate route gets an intervention with stated preserved and altered properties. The correspondence candidate is tested by a pairing break (audio rolled, shuffled, or replaced by another speaker's real audio), the artifact candidates by the trim and identity controls, and the single-modality candidates by the modality-stress cells and single-stream probes. In the pairing case studies, write $R_T(f)$ for response AUC and $\Delta H_T(f) = H(f) - H_T(f)$ for the honest-AUC drop, where $H(f)$ is intact honest AUC and $H_T(f)$ uses perturbed clips as the real class. A candidate is *confirmed* when response AUC is at least 0.70 and honest AUC drops by at least 0.05, *rejected* when all three response AUCs lie in $[0.40, 0.60]$ and honest AUC changes by at most 0.02, and *undetermined* otherwise or when the intervention is unavailable. These thresholds describe completed case studies; they locate modality and pairing dependence and do not separate generator fingerprints from other visual cues. A *rejected* verdict means that none of the three tested breaks moves the score, not that no route reads correspondence, and a *confirmed* verdict that sits at the threshold is marked as such. Verdicts are in Table 2; AVAD and VFD, re-scored through their own media pipelines with the same breaks rendered in the media, are confirmed, AVAD at the threshold (Table 11); unsup_sync, not re-scored, is undetermined (Appendix J).

Suite, corpora, and hygiene. We evaluate 19 detector or scorer instances from ten backbone or training lineages in five design groups: correspondence scorers, audio-only, visual-only, naive fusion, and supervised SOTA. Seven probes share one frozen AV-HuBERT backbone (Shi et al., 2022), which is what lets the mixture sweep hold the architecture fixed. Every signature is measured on FAVC (Khalid et al., 2021), with two axes re-measured on a TIMIT pair as a check (Appendix H); LAV-DF (Cai et al., 2022), AV-Deepfake1M (AV1M) (Cai et al., 2024), DFTIMIT (Korshunov & Marcel, 2018), VidTIMIT-Wav2Lip (Sanderson & Lovell, 2009), AVLips (Liu et al., 2024), and DeepSpeak (Barrington et al., 2024) serve as transfer targets or audit corpora, DeepSpeak held out with zero lineage overlap. A cell-purity audit flags cells contaminated in the wrong modality on every corpus; AV1M's nominally clean FV–RA cell carries audio-side class signal, so it is a transfer target but not a skew-measurement domain. Interventions are paired at the clip level with paired bootstrap; held-out tests are leave-one-family-out with permutation nulls; the DeepSpeak claims were pre-registered. A detector enters through per-clip scores and needs two to four inference passes over the FAVC test set (Appendices D, F, and G).

3 No single score captures the OOD failure pattern

Shortcut reliance and OOD failure are different properties. Different evidence routes respond to different shifts, so a single score that pools them loses the pattern that matters.

The test. The shortcuts themselves are established (Smeu et al., 2025; Dong et al., 2023); what we test is the diagnostic built on them. If a single in-domain shortcut score were a sufficient diagnostic, it should track collapse in the same direction on every target. None of the tested scores does: as family-level correlations with collapse, the leading-trim axis changes sign between LAV-DF and the TIMIT and AVLips targets, modality skew changes sign between LAV-DF/AV1M and AVLips/DFTIMIT, and so does the combined shortcut scalar (values in Appendix H). This does not refute shortcut learning, since targets preserve or break different shortcuts; the sign changes because each target's shift engages a different route, which is the multi-route structure the rest of the paper builds on.

Three counterexamples. (i) A heavy silence rider, the AV-concat probe, does *not* collapse on LAV-DF ($0.752 \rightarrow 0.716$), while the largest rider we measured, MRDF (a 2024 method outside

Table 1: **Headline AUC conflates routes (full route map).** All 19 suite instances on the FAVC diagnostic corpus: overall AUC, AUC on the honest FV–RA subset, their difference, and the modality-skew and leading-trim axis values. The drop from overall to honest lands on the shortcut routes: audio-routed detectors lose 0.25–0.27, correspondence scorers 0.014–0.114, and the FAVC-trained AVH-Align head, which reads the visual stream (Appendix J), 0.001. Lead-trim uses the reference-protocol trim (§2), with feature-level detectors measured on one feature cache built for both the trimmed and untrimmed inputs. Lead-trim is not significant for the visual-feature probes, VFD, unsup_sync, AVH-Align (FAVC), RawNet2, and BA-TFD+, and skew is not significant for AVAD. The media-level visual detectors' lead-trim values ($+0.05$–0.06) come from the protocol's video-frame removal; they vanish under the duration-matched trailing control (App. H).

Detector group	Detector	Overall	Honest	Δ	Skew	Lead-trim
Correspondence	SpeechForensics	0.988	0.974	−0.014	+0.025	−0.001
Correspondence	VFD	0.925	0.852	−0.073	+0.085	+0.004
Correspondence	AVAD	0.845	0.731	−0.114	−0.008	+0.009
Correspondence	unsup_sync	0.879	0.848	−0.031	+0.147	+0.005
Audio-only	RawNet2	0.785	0.535	−0.250	+0.465	−0.000
Audio-only	AVH probe-audio	0.759	0.485	−0.274	+0.491	+0.039
Audio-only	pooled_audio_linear	0.760	0.489	−0.271	+0.501	+0.025
Naive concat	AVH probe-concat	0.752	0.478	−0.274	+0.510	+0.069
Naive concat	AVH linear probe	0.891	0.766	−0.125	+0.234	+0.026
Naive concat	pooled_jointav_linear	0.901	0.785	−0.116	+0.215	+0.013
Visual-only	AVH probe-visual	0.914	0.907	−0.007	−0.308	−0.002
Visual-only	pooled_video_linear	0.949	0.939	−0.010	−0.335	−0.002
Visual-only	CViT2	0.785	0.773	−0.012	−0.239	+0.055
Visual-only	FTCN	0.831	0.825	−0.006	−0.308	+0.047
Visual-only	LipFD	0.870	0.806	−0.064	−0.319	+0.009
Visual-only	Xception	0.628	0.629	+0.001	−0.108	+0.060
Supervised	BA-TFD+	0.776	0.559	−0.217	+0.431	+0.002
Supervised	AVH-Align (FAVC)	0.987	0.986	−0.001	−0.097	+0.001
Supervised	AVH-Align (AV1M)	0.666	0.523	−0.143	+0.284	+0.025

the frozen suite, scored with the same instrument), does ($0.786 \rightarrow 0.512$). (ii) The shortcut-clean BA-TFD+ and RawNet2 sit at chance on the honest subset, and BA-TFD+ has the worst transfer in the suite (Table 1, Appendix B). Riding a shortcut and failing to transfer are different properties. (iii) AVH-Align (FAVC) is clean on every axis and near-perfect on the honest subset, the profile of a correspondence detector, yet a direct pairing test rejects that candidate: breaking the audio–visual pairing of every real test clip leaves its honest-subset AUC unchanged (§4, Table 2). Axis values narrow the route down; they do not certify it (family signatures: Figure 3).

Accuracy is not the missing scalar either. Recalibrating the leaderboard on the honest FV–RA subset localizes the inflation exactly on the shortcut routes: audio-routed detectors land at chance while correspondence detectors stay useful, and a weak-everywhere detector (Xception) drops nothing (Table 1). The subset penalizes audio routing, not incompetence: headline AUC pools routes that will fail differently under shift.

4 Training data can redirect the route under a fixed architecture

Training data can redirect the evidence route even when the architecture is held fixed. The architecture constrains which routes are available to a detector; within that set, the training distribution influences which route the detector learns to use, and in-domain accuracy does not identify the result. This section supplies the first link of the mechanism: Figure 2 shows the sweep and Table 2 collects the numbers.

4.1 The natural mixture sweep

One frozen AV-HuBERT backbone plus one logistic head; the only thing that changes is the share λ of FAVC in a FAVC/AV1M training mixture, FAVC being the corpus whose FV–RA cell is audio-

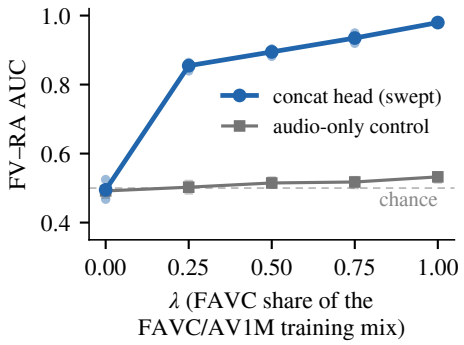


Figure 2: **The training mixture redirects the route under a fixed architecture.** One frozen AV-HuBERT backbone, one logistic head, and only the FAVC share λ of a FAVC/AV1M training mixture changes. The concat head's FV–RA (honest-subset) AUC rises monotonically (line: seed mean; open markers: three seeds per λ), while an audio-only head trained on the same mixtures stays at chance.

clean and AV1M the corpus whose FV–RA cell carries audio-side class signal. As λ runs from 0 to 1 the concat head's FV–RA AUC rises monotonically from chance (0.495) to 0.980 and its skew moves from audio-dominated to balanced, while an audio-only head trained on the same mixtures stays at chance throughout (Figure 2). A linear head on separately pooled streams has no cross-stream term, so the sweep shows increased use of visual evidence, not a landing on correspondence. The endpoints are not a tautology: at λ=0 the head trains on AV1M alone, which contains visual forgeries in every cell, and is still blind on the audio-clean FAVC FV–RA cell. Nor is the curve a train–test corpus match: the same heads scored on DFTIMIT's FV–RA cell, an independent source whose fakes reuse the real audio, rise the same way (Table 2; seeds, N and protocol in Appendix E).

4.2 What in the training source does the redirecting

The natural sweep changes identities, generators, class balance and low-level properties together with the corpus. Two controlled arms separate them. The *matched sweep* draws training units instead of clips: a unit is one source video with its genuine clip and one clip of each manipulated type, so the class ratio, the cell ratio, N and the within-identity real–fake pairing are fixed by construction at every λ; the identity set itself still changes with the corpus, and λ changes only which corpus supplies the units. The curve keeps its shape (FV–RA AUC $0.61 \rightarrow 0.99$, skew $+0.39 \rightarrow 0.00$), so the route change survives controlling the class ratio and the identity and source-video pairing of the training set. What still differs between the corpora is the identity set, the generators, and the recording and content properties of the sources; the *same-source recombination* asks whether one of these, the availability of an audio-side label cue, moves the route on its own. It keeps everything FAVC and replaces the pooled audio features of a share ρ of the visual-forgery training fakes with those of a cloned-audio clip of the same identity, holding the visual forgery, identity set, generators and class ratio fixed. A cue available for a quarter of the fakes already pulls the skew from -0.49 to -0.10 while honest AUC stays at 0.99; only when every fake carries the cue does visual discrimination fall (0.86), and the route becomes mixed rather than audio-only. The natural sweep shows the phenomenon; the controlled arms show which part of the source drives it (DFTIMIT values, the AV1M endpoint, per-seed values and unit construction: Appendix E).

4.3 Released detectors and other backbones

The supervised head of AVH-Align (Smeu et al., 2025), code and architecture unchanged, sees the honest subset when FAVC-trained and is audio-dominated and collapsing when AV1M-trained (Table 2). Nothing in that head ties its score to audio–visual agreement, so both stream-specific routes are available to fit the training labels. A direct pairing test separates the FAVC head's visual route from correspondence: rolling, shuffling, or replacing the audio of every real test clip leaves its honest AUC where it was, while the same interventions drive a zero-parameter sync score to near zero. The AV1M head takes the audio-side class signal that AV1M supplies even in its FV–RA cell: rolling its audio against the video *raises* its AUC. The pair closes the chain of §1 on one architecture: two training sources; two routes, the visual one verified by the pairing break and the audio one by the audio-only probe and the roll response (its pairing verdict is undetermined, Table 2); and two transfer outcomes, the visual route transferring better than the audio route on LAV-DF and DeepSpeak (pairing tables: Appendix J).

Table 2: **The mechanism in numbers.** (a) Sweeps under a fixed architecture: one frozen AV-HuBERT backbone with one linear head, seed means over three seeds at λ or $\rho = 0 \rightarrow 1$ (the VFD and BA-TFD+ sweeps are in Table 7). (b) Released detectors and the correspondence statistic: route verdict from the pairing break (Table 11) and transfer. Route: modality skew (RV–FA minus FV–RA AUC; positive = audio dominated). Honest: FAVC FV–RA AUC; LAV-DF: FV–RA cell; DeepSpeak: overall.

(a) Sweeps under a fixed architecture (§4.1–4.2)

Instance	Training source	Route (skew)	Honest	DFTIMIT
AVH probe	AV1M (λ=0)	audio (+0.56)	0.495	0.489
AVH probe	FAVC (λ=1)	visual (−0.06)	0.980	0.858
AVH probe	matched units, λ=0 → 1	audio → visual (+0.39 → 0.00)	0.61 → 0.99	0.60 → 0.89
AVH probe	FAVC, audio cue ρ=0 → 1	visual → mixed (−0.49 → +0.14)	0.99 → 0.86	0.90 → 0.72

(b) Released detectors and the correspondence statistic (§4.3, §5)

Instance	Training source	Route (skew; verdict)	Honest	LAV-DF	DeepSpeak
AVH-Align (publ.)	FAVC	visual (−0.10; rejected)	0.986	0.846	0.720
AVH-Align (publ.)	AV1M	audio (+0.28; undetermined)	0.523	0.673	0.568
SpeechForensics	none	corresp. (+0.03; confirmed)	0.974	0.795	0.846

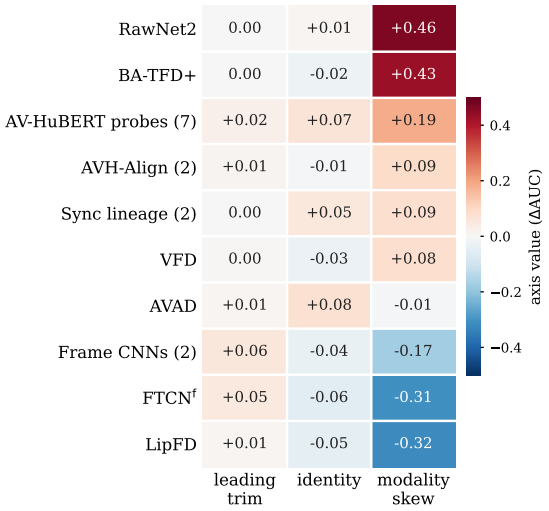


Figure 3: **The diagnostic route signature by detector family.** Family means of the three diagnostic axes, all measured on FAVC, over the 19-instance suite; these are the inputs to the family-level rows of Table 4. The leading-trim axis is the reference-protocol measurement; the same axis file feeds Tables 4 and 1. Families follow Appendix D and rows are ordered by modality skew. Reading: the sign of the skew axis separates the audio-routed families (positive, +0.43–+0.47) from the visual-routed ones (negative); a family near zero on all three axes carries no silence or identity shortcut and no dominant single-stream shortcut on the measured axes, which is where the sync and correspondence families sit. Near-zero axes do not identify the route: AVH-Align (FAVC) sits near zero and reads the visual stream (Table 2), which is why the pairing break, not the signature, is the confirmation step behind the transfer results of §5.1. FTCN's identity contrast is the value recorded at its preregistration (Table 6).

The same sweep on two other frozen representations shows how far training can move a route (Table 7). On VFD voice/face embeddings (Cheng et al., 2024) the concat head's FV–RA AUC rises from 0.53 to 0.65, carried by the visual stream: the visual-only head rises with it and the audio-only head stays at chance. On BA-TFD+ features (Cai et al., 2023) the concat head reads 0.68–0.72 at every λ and its skew stays audio-dominated. The pattern is consistent with the redirect being limited by the visual evidence a linear head can read from the representation; the limit applies to the tested linear readouts, not to the full trainable systems. A label-free router driven by training-side skew matches the best fixed route in-sweep but does not replicate on DeepSpeak. The architecture constrains which routes are available; within that set, the training distribution determines which one is used (single-stream references and per-seed values: Appendix E; router: Appendix K).

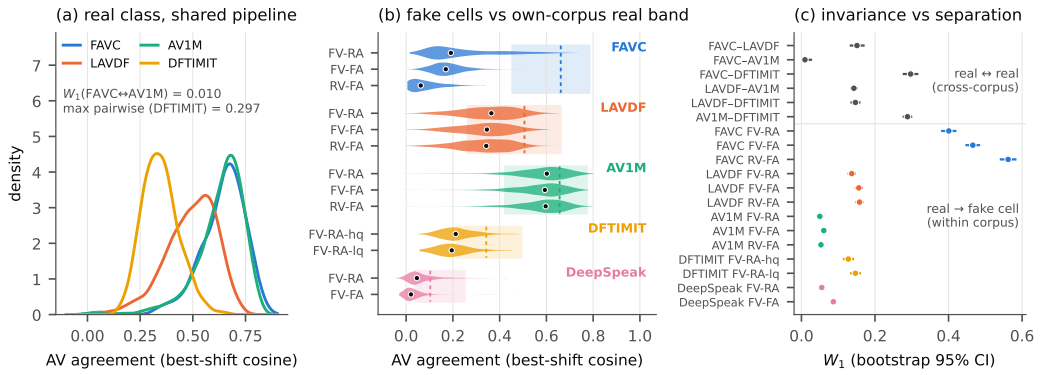


Figure 4: **The stable signal is the real-class agreement level, and it compares only under a matched pipeline. (a)** Real-class distributions of the correspondence statistic. FAVC and AV1M, which share an extraction pipeline and VoxCeleb2 source video, are near-identical (the panel annotation rounds the Wasserstein distance given in §5.3); DFTIMIT, extracted differently, is offset. **(b)** Each corpus's fake cells against its own real band: the within-corpus real-vs-fake separation replicates even where the raw levels do not line up (§5.3). **(c)** The same contrast as distances: between matched-pipeline corpora, real-to-real W_1 stays small compared with the real-to-fake separation inside any one corpus.

5 Why routes fail differently: the correspondence case

Routes differ in how stable their evidence is across domains, and transfer depends on which evidence the target shift preserves. The shortcut routes read features whose distributions shift between datasets, so each has a dataset-specific coordinate to overfit. The correspondence route reads the real-class *level* of audio–visual agreement, whose within-corpus real–fake separation replicates across the evaluated corpora. The one-architecture chain of §4 covers the visual and audio routes; no detector in the suite is moved into the correspondence route by its training data. The stability evidence below therefore comes from zero-parameter and unsupervised correspondence scorers, compared with the trained routes on the same targets, and the restore intervention of §5.2 substitutes the feature rather than the training source (its own transfer: Appendix I).

5.1 The route is clean and its transfer is real

The four correspondence scorers (SpeechForensics, VFD, AVAD, unsup_sync; the first three confirmed under pairing breaks, Table 11) are clean on every axis of the route map (Table 1): small skew against the strongly positive skew of audio-routed detectors, and no significant silence or identity inflation. Across the natural FAVC→LAV-DF shift, SpeechForensics holds 0.795 on the honest subset, where the audio-routed detectors read only what a purity-flagged, audio-contaminated cell allows; the weaker correspondence scorers fall to that level too. The FAVC-trained AVH-Align head, which reads the visual stream (§4), reaches 0.846 on the same cell, so honest-subset transfer to LAV-DF is not unique to correspondence. On the fully independent DeepSpeak the order reverses: SpeechForensics holds 0.846 against 0.720 for the visual route (Table 2). Both statistics are evaluated without target-domain fitting or readout selection (Table 3).

5.2 Controlled interventions, in both directions

Correlational cleanliness could be a coincidence, so we intervene both ways. **Break it:** rolling the audio stream by Δ frames collapses real/fake separation monotonically and then inverts it (FAVC AUC $0.87 \rightarrow 0.05$; LAV-DF $0.68 \rightarrow 0.25$): a desynchronized *real* clip scores more fake than an actual fake, directly linking the detection score to audio–visual pairing. **Restore it:** forcing a collapsed, audio-dominated detector onto an 8-d correspondence feature un-collapses it (FV–RA $0.44 \rightarrow 0.79$; worst case across the sweep $0.43 \rightarrow 0.73$).

5.3 What exactly is stable, and over what scope

Decomposing the correspondence score into a *timing* channel and an *agreement-level* channel shows that the stable signal is the level, not timing alone. Generated video attenuates cross-modal

agreement roughly $3\times$ even when perfectly synchronized, while the real-class distribution is near-identical across datasets under a shared extraction pipeline (FAVC↔AV1M Wasserstein 0.0096). FAVC, LAV-DF, and AV1M all draw their real video from VoxCeleb2, so that match is a same-source comparison; the quantity that travels across sources is the within-corpus real-vs-fake separation, which replicates on DFTIMIT, VidTIMIT, and DeepSpeak, none of them VoxCeleb2-sourced. Individual fakes can still match the real-class level, and newer generators reduce the separation (§6; distributions in Figure 4).

What the correspondence signal is not. The measured shortcuts do not explain it: leading silence is absent from LAV-DF (silence-AUC 0.51), and where a shortcut is shared (FAVC↔AV1M) the within-class correlations are small ($|r| \leq 0.04$). Encoding and visual quality remain unresolved, and the pairing breaks establish what the scorers read, not why the fakes differ, so §6 states the stability as a regularity of the evaluated corpora. Availability of a cue is not use of it: DFTIMIT's fakes reuse the real utterances, and VidTIMIT-XTTS and DeepSpeak invert the silence association (Appendix H).

Independent, pre-registered replays. Both directions replicate off-FAVC under pre-registration. On self-generated voice-clone cells (VidTIMIT video + XTTS-v2 audio (Casanova et al., 2024), 320 units) all three frozen predictions hit: corr_sync best, the visual probe at chance, and the offset collapse reproduced. On DeepSpeak, in the original pre-registered evaluation, the level-channel separation reproduces and correspondence scorers lead, alongside a pre-registered falsifier that *triggered* (§6; numbers and full record: Appendix G).

6 Where the correspondence route still breaks

Synchronized fakes can narrow the level-channel margin. Our Wav2Lip attack (Prajwal et al., 2020) (640 fully synchronized fakes) closes the timing channel, yet the level channel still separates (2.2–$2.8\times$ attenuation; corr_sync leads the tested scorers, $0.936/0.976$). A level-only score loses discrimination when real and fake agreement distributions coincide. A LatentSync-1.6 self-driven cell on VidTIMIT (n=602) approaches this condition: level attenuation is only $1.14\times$ and corr_sync AUC falls to 0.631 (cross-driven: $2.17\times$, AUC 0.956). A source-trained readout can also fail on a new target. On DeepSpeak, our pre-registered falsifier *triggered*: the absolute-logit readout of an AV1M-real-trained offset-prediction model scored AUC 0.303 against a frozen band of $[0.72, 0.86]$, an inversion no monotone recalibration fixes. The sweep form of the restored head (§5.2) also inverts on DeepSpeak (Appendix I). Fitting provenance is recorded in Table 3; none is a target refit (exploratory readout follow-up: Appendix I). Together, these cases show that stability is an empirical regularity, not a necessity. Our transfer verdict concerns lip-sync, reenactment, and content-driven manipulations; audio-untouched face swaps would exercise the visual route instead. Visual evidence transfers to visual edits, while correspondence detects audio edits that a visual route misses when they disrupt audio–visual agreement. The error audit (Appendix N) finds level-matched missed fakes and level-depressed, rarely desynced flagged reals.

7 Discussion and conclusion

AVRouteBench diagnoses evidence use beyond in-domain accuracy. Fixed-architecture sweeps and source reversals show that training data redirects routes. Matched and same-source controls retain the effect after controlling class, cell, and identity pairing and implicate an audio-side label cue; cross-backbone sweeps support the limiting role of accessible visual evidence (§4). Pairing-break tests separate visual from correspondence routes, whose stability and failure conditions explain their different transfer outcomes (§5–§6). These conclusions do not depend on the exploratory risk ranking (Appendix H); limitations appear in Appendix A.

Reproducibility statement

AVRouteBench is released with everything needed to reproduce the numbers in this paper and to diagnose new detectors: the per-clip prediction files for all 19 suite instances on every covered manifest (original, the reference-protocol leading-trim and its duration-matched trailing control, encode-only controls, identity-disjoint, and composition-matched splits), the manifests themselves,

the axis-computation and bootstrap code, the leave-one-family-out evaluation and framework-audit scripts (spec-robustness, suite-stability, noise-propagation, cell-purity, selection), and the frozen pre-registration files (the FTCN, DeepSpeak-suite, and LipFD prediction files and the voice-clone-replay and DeepSpeak runbooks) with their recorded outcomes (Appendix G). Section 2 and Appendices E and H specify the silence-trim protocol, the identity-split construction, the bootstrap and permutation protocols (10,000 draws, one-sided Monte Carlo), and the black-box adapter interface (one row per clip: identifier, label, score); Appendix D lists every detector with its training source. New detectors require two (raw signature) to four (purified variant) inference passes over the FAVC test set and no access to model internals.

Ethics statement

This work is defensive: it evaluates and diagnoses deepfake *detectors* and proposes no generation technique. Two dual-use considerations are disclosed. First, §6 states the route's failure condition (an attacker matching the real-class agreement level); we judge disclosure the right call because the condition is a property an adversary can discover by probing any correspondence detector, while defenders can only monitor for it if it is named. Second, the study constructs two controlled fake cells (640 Wav2Lip clips; 320 XTTS-v2 voice-clone clips) solely as detection stress tests: both are built from the consented research corpus VidTIMIT, use only published generation tools, and are released as scores and manifests, not as media or generation pipelines. All corpora (FakeAVCeleb, LAV-DF, AV-Deepfake1M, DFTIMIT, VidTIMIT, AVLips, DeepSpeak) are used under their research licenses; no new human-subject data was collected, and no identity is targeted by any constructed cell.

AI use statement

Generative AI assistants (Anthropic's Claude and OpenAI Codex) were used as follows. *Writing:* drafting and editing parts of the manuscript for readability and consistency, formatting references, and summarizing related literature; the authors reviewed every sentence, and every reference was verified against publisher and index records (Crossref, OpenAlex, arXiv). *Research methodology:* feedback on the design of the intervention protocols and control conditions; the hypotheses, pre-registered predictions and claims are the authors' own. *Implementation and execution:* writing and running evaluation, bootstrap, permutation and job-submission code; the authors reviewed the code and checked its outputs against the released per-clip prediction files. *Analysis:* statistical analysis and preliminary interpretation of results; every reported number and interpretation was verified by the authors against the underlying result files. No AI-generated text, code, analysis or claim entered the paper without author verification, and the authors take full responsibility for the paper's content, analyses and conclusions.

References

Christina Baek, Yiding Jiang, Aditi Raghunathan, and J Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. In *NeurIPS*, 2022.

Sarah Barrington, Maty Bohacek, and Hany Farid. The DeepSpeak dataset. *arXiv preprint arXiv:2408.05366*, 2024.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175, 2010.

Jakub Białek, Juhani Kivimäki, Wojtek Kuberski, and Nikolaos Perrakis. Estimating model performance under covariate shift without labels. In *NeurIPS*, 2025.

Dragoş-Alexandru Boldişor, Ştefan Smeu, Dan Oneaţă, and Elisabeta Oneaţă. Investigating self-supervised representations for audio-visual deepfake detection. In *CVPR*, 2026.

Zhixi Cai, Kalin Stefanov, Abhinav Dhall, and Munawar Hayat. Do you really mean that? Content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization. In *DICTA*, 2022.

Zhixi Cai, Shreya Ghosh, Abhinav Dhall, Tom Gedeon, Kalin Stefanov, and Munawar Hayat. Glitch in the matrix: A large scale benchmark for content driven audio-visual forgery detection and localization. *Computer Vision and Image Understanding*, 236:103818, 2023.

Zhixi Cai, Shreya Ghosh, Aman Pankaj Adatia, Munawar Hayat, Abhinav Dhall, Tom Gedeon, and Kalin Stefanov. AV-Deepfake1M: A large-scale LLM-driven audio-visual deepfake dataset. In *ACM Multimedia*, 2024.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. XTTS: A massively multilingual zero-shot text-to-speech model. In *Interspeech*, 2024.

Nuria Alina Chandra, Hannah Lee, Ryan Murtfeldt, Lin Qiu, Arnab Karmakar, Emmanuel Tanumihardja, Kevin Farhat, Ben Caffee, Changyeon Lee, Jongwook Choi, Sejin Paik, Aerin Kim, and Oren Etzioni. Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. *arXiv preprint arXiv:2503.02857*, 2025.

Harry Cheng, Yangyang Guo, Tianyi Wang, Qi Li, Xiaojun Chang, and Liqiang Nie. Voice-face homogeneity tells deepfake. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 20(3), 2024.

Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian. Not made for each other – audio-visual dissonance-based deepfake detection and localization. In *ACM Multimedia*, 2020.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP*, 2023.

Davide Cozzolino, Alessandro Pianese, Matthias Nießner, and Luisa Verdoliva. Audio-visual person-of-interest deepfake detection. In *CVPR Workshops*, 2023.

Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *CVPR*, 2023.

Chao Feng, Ziyang Chen, and Andrew Owens. Self-supervised video forensics by audio-visual anomaly detection. In *CVPR*, 2023.

Federico Fontana, Anxhelo Diko, Romeo Lanzino, Marco Raoul Marini, Bachir Kaddar, Gian Luca Foresti, and Luigi Cinque. Revisiting deepfake detection: Chronological continual learning and the limits of generalization. *arXiv preprint arXiv:2509.07993*, 2025.

Saurabh Garg, Sivaraman Balakrishnan, Zachary C Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *ICLR*, 2022.

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(11):665–673, 2020.

Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with confidence on unseen distributions. In *ICCV*, 2021.

Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *CVPR*, 2021.

Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Leveraging real talking faces via self-supervision for robust forgery detection. In *CVPR*, 2022.

Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. In *NeurIPS Datasets and Benchmarks Track*, 2021.

Pavel Korshunov and Sébastien Marcel. DeepFakes: a new threat to face recognition? Assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018.

Yachao Liang, Min Yu, Gang Li, Jianguo Jiang, Boquan Li, Feng Yu, Ning Zhang, Xiang Meng, and Weiqing Huang. SpeechForensics: Audio-visual speech representation learning for face forgery detection. In *NeurIPS*, 2024.

Weifeng Liu, Tianyi She, Jiawei Liu, Boheng Li, Dongyu Yao, Ziyou Liang, and Run Wang. Lips are lying: Spotting the temporal inconsistency between audio and visual in lip-syncing deepfakes. In *NeurIPS*, 2024.

John Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. In *ICML*, 2021.

Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *ACM Multimedia*, 2020.

Nicolas M Müller and Pascal Debus. The watermark shortcut: How provenance marking sabotages audio deepfake detection. *IEEE Signal Processing Letters*, 33:3476–3480, 2026. doi: 10.1109/LSP.2026.3724777.

Nicolas M Müller, Franziska Dieckmann, Pavel Czempin, Roman Canals, Konstantin Böttinger, and Jennifer Williams. Speech is silver, silence is golden: What do ASVspoof-trained models really learn? In *ASVspoof 2021 Workshop*, pp. 55–60, 2021.

Trevine Oorloff, Surya Koppisetti, Nicolò Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj. AVFF: Audio-visual feature fusion for video deepfake detection. In *CVPR*, 2024.

K R Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and C V Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM Multimedia*, 2020.

Tal Reiss, Bar Cavia, and Yedid Hoshen. Detecting deepfakes without seeing any. *arXiv preprint arXiv:2311.01458*, 2023.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *ICCV*, 2019.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020.

Conrad Sanderson and Brian C Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. In *International Conference on Biometrics (ICB), Lecture Notes in Computer Science, Vol. 5558*, pp. 199–208, 2009.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *NeurIPS*, 2020.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *ICLR*, 2022.

Ştefan Smeu, Dragoş-Alexandru Boldişor, Dan Oneaţă, and Elisabeta Oneaţă. Circumventing shortcuts in audio-visual deepfake detection datasets with unsupervised learning. In *CVPR*, pp. 18815–18825, 2025.

Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-end anti-spoofing with RawNet2. In *ICASSP*, 2021.

Yihui Wang, Yonghui Yang, Jilong Liu, Fengbin Zhu, Le Wu, and Tat-Seng Chua. Suppressing forgery-specific shortcuts for generalizable deepfake detection. *arXiv preprint arXiv:2606.01843*, 2026.

Deressa Wodajo and Solomon Atnafu. Deepfake video detection using convolutional vision transformer. *arXiv preprint arXiv:2102.11126*, 2021.

Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. DeepfakeBench: A comprehensive benchmark of deepfake detection. In *NeurIPS Datasets and Benchmarks*, 2023.

Andrii Yermakov, Jan Čech, Jiří Matas, and Mario Fritz. Deepfake detection that generalizes across benchmarks. In *WACV*, pp. 773–783, 2026.

Han Yu, Kehan Li, Dongbai Li, Yue He, Xingxuan Zhang, and Peng Cui. ODP-Bench: Benchmarking out-of-distribution performance prediction. In *ICCV*, pp. 1846–1858, 2025.

Ning Yu, Larry Davis, and Mario Fritz. Attributing fake images to GANs: Learning and analyzing GAN fingerprints. In *ICCV*, 2019.

Yaodong Yu, Zitong Yang, Alexander Wei, Yi Ma, and Jacob Steinhardt. Predicting out-of-distribution error with the projection norm. In *ICML*, 2022.

Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *ICCV*, 2021.

A LIMITATIONS

The agreement-level stability is empirical, and a 2025 diffusion lip-syncer already narrows its margin (§6). The raw-level match between FAVC and AV1M is a same-source comparison (both draw real video from VoxCeleb2); the cross-source evidence is the within-corpus separation on DFTIMIT, VidTIMIT, and DeepSpeak. Training data redirects the route only among the routes an architecture makes available (§4.3), and the data-to-route-to-transfer chain is closed on one architecture for the visual and audio routes only, not for correspondence (§5); the signature-risk sensitivity analyses share the FAVC base with their gap outcome and depend on family partition (Appendix H); the transfer verdict depends on the target pool (§6); and corr_sync is not SOTA in-domain. The supervised side of the suite is two architectures, three instances; AVFF (Oorloff et al., 2024) is not among them, and the claims about supervised heads rest on the fixed-architecture sweeps of §4, not on the breadth of that group. Face swap is represented in the target pool by one corpus, DFTIMIT (GAN face swap with the original audio), so the failure-mode claim is tested on face swap only at that corpus's scale and generation era; DFDC lacks an RV–FA cell (Appendix F) and no modern face-swap target with a full factored-cell signature is included. Every signature is measured on one corpus, FakeAVCeleb, whose RV–FA cell sits at the purity-flag boundary (Appendix F); the flag touches the skew axis of the signature, not the FAVC and DFTIMIT FV–RA cells on which the sweeps of §4 are read, and flagged target cells are marked where used. The TIMIT comparison (Appendix H) re-measures only two axes, with reduced predictive strength, a leading-amplitude artifact in place of silence and no identity axis. All corpora are laboratory-constructed; in-the-wild collections report sharper degradation (Chandra et al., 2025).

B DEFERRED TABLES AND FIGURES

Table 3: **Provenance of the transfer numbers.** What each scorer is, where any parameters were fitted, and what the target labels were used for. This table contains the original transfer evaluations, including the registered offset-model failure (§6). The exploratory readout follow-up is reported in Appendix I.

Target (cell)	Scorer	Parameters fitted on	Target labels	AUC
DeepSpeak (overall)	SpeechForensics statistic	none (zero-shot)	evaluation only	0.846
DeepSpeak (overall)	zero-parameter sync score	none	evaluation only	0.809
DeepSpeak (overall)	Offset-prediction model, original absolute-logit readout (registered)	AV1M real clips	evaluation only	0.303
DeepSpeak (overall)	AVH-Align supervised head	FAVC / AV1M	evaluation only	0.720 / 0.568
LAV-DF (FV–RA)	SpeechForensics statistic	none	evaluation only	0.795
LAV-DF (FV–RA)	AVH-Align supervised head	FAVC / AV1M	evaluation only	0.846 / 0.673
LAV-DF (FV–RA)	BA-TFD+	LAV-DF (home domain)	training and evaluation	0.997

Full-size exhibits deferred from the main text for the page budget: the terminology table.

Table 4: **Family-level held-out prediction of OOD collapse.** Leave-one-family-out Spearman ρ_{fam} (one-sided permutation p in parentheses, 10,000 refit draws) on the 19-instance, ten-family suite. Reading: $\rho > 0$ means the families the signature marks as higher-risk collapse more on that target; the test is one-sided, so a negative cell is a null (the learned ordering does not hold there), not evidence of a reversed ordering. The bold row is the registered statistic; rows marked sens. are variants of it, not independent replications (outcome, refit convention and protocols: Appendix H). Combined scalar: $|\mathrm{silence}|+|\mathrm{identity}|$. The purified rows replace the raw trim axis with the net-silence contrast, scored on the gap and on the base-free outcome. Increment rows are instance-level. Joint test across the five targets: p=0.013 (DeepSpeak sensitivity: Appendix G). Instance-level companion: Table 9.

	Target corpus				
Predictor (family-level, primary)	LAV-DF	AV1M	AVLips	DFTIMIT	VidTIMIT
Leading-trim scalar	−0.50 (.41)	−0.63 (.42)	−0.14 (.21)	+0.62 (.006)	−0.78 (.63)
Modality-skew scalar	+0.71 (.007)	+0.08 (.17)	+0.50 (.058)	+0.55 (.031)	−0.66 (.57)
Combined shortcut scalar	+0.06 (.22)	−0.87 (.81)	+0.14 (.20)	+0.18 (.15)	−0.65 (.58)
FAVC diagnostic AUC	−0.09 (.28)	+0.65 (.023)	−0.69 (.72)	−0.07 (.26)	−0.81 (.70)
Route signature (family-level, registered)	0.59 (.051)	0.35 (.14)	0.52 (.098)	0.59 (.041)	0.20 (.22)
sens.: purified net-silence substitution	+0.81 (.006)	+0.57 (.062)	+0.74 (.038)	+0.39 (.12)	+0.38 (.12)
sens.: purified, base-free outcome	+0.77 (.010)	+0.63 (.053)	+0.41 (.15)	+0.38 (.12)	+0.46 (.094)
sens.: $\Delta\rho$ over accuracy baseline (instance)	+0.48 (.005)	+0.31 (.009)	+0.98 (.005)	+0.17 (.14)	+1.02 (.005)
sens.: $\Delta\rho$ over skew alone (instance)	−0.30 (.96)	+0.26 (.023)	+0.09 (.077)	−0.23 (.87)	+0.94 (.003)

Table 5: **Terminology.** Working definitions for the load-bearing terms of this paper.

Term	Meaning
Evidence route	The channel through which a detector's decision flows: a dataset artifact, a single modality, a generator fingerprint, or cross-modal correspondence (§1).
Cell	A subset of a corpus factored by which stream is synthetic: FV–RA, RV–FA, FV–FA, RV–RA.
Honest subset	The audio-clean FV–RA cell: discrimination requires the visual stream or its relation to audio; nominal labels require a purity audit.
Axis	A candidate evidence channel probed by an intervention or comparison, with uncertainty matched to the design.
Benchmark signature	A detector's values on all four route axes.
Diagnostic signature	The three intervention-side axes (silence, identity, skew) used to diagnose OOD risk and failure mechanisms (Appendix H).
Silence vs. net silence	In signature tuples, "silence" abbreviates the raw leading-trim axis (the label frozen in the pre-registration files); *net silence* always denotes the purified contrast (Appendix H).
Family	Detector instances sharing a backbone or training lineage; the unit of generalization (ten families).
Diagnostic corpus	FakeAVCeleb (FAVC): the single corpus on which every signature is measured.
Measurement corpus	A corpus whose factored cells support route measurement (five in this study).
Transfer target	A corpus on which OOD collapse is predicted and scored (five diagnostic-loop targets; DeepSpeak as a sixth, pre-registered target).

C Extended related work

Audio–visual deepfake detection and its evaluation. Detectors in current use span supervised SOTA trained end-to-end on a labeled corpus (BA-TFD+ (Cai et al., 2023) and related fusion models), correspondence-based methods that score the agreement between what is heard and what is seen (SpeechForensics (Liang et al., 2024), AVAD (Feng et al., 2023), VFD (Cheng et al., 2024), and AVH-Align (Smeu et al., 2025), built on AV-HuBERT (Shi et al., 2022)), and single-modality baselines on the audio stream (RawNet2 (Tak et al., 2021)) or the visual stream (Xception (Rössler et al., 2019), FTCN (Zheng et al., 2021), CViT (Wodajo & Atnafu, 2021)). Correspondence reading has a lineage of its own, from audio–visual dissonance and affective-congruence scores (Chugh et al., 2020; Mittal et al., 2020) through lip-semantic and real-data-only self-supervision (Haliassos et al., 2021; 2022; Cozzolino et al., 2023; Reiss et al., 2023) to current supervised fusion (Oorloff et al., 2024); our contribution is not a new member of this class but a diagnostic framework for when and why the class transfers. Severe cross-dataset degradation is widely reported (Yan et al., 2023), across benchmark generations (Yermakov et al., 2026) and chronologically, with near-random forward transfer to future generators (Fontana et al., 2025); successive benchmarks — FakeAVCeleb (Khalid et al., 2021), LAV-DF (Cai et al., 2022), AV-Deepfake1M (Cai et al., 2024), DeepSpeak

(Barrington et al., 2024) — were built in part to expose it, and in-the-wild collections document a sharper drop still on circulated 2024 web content (Chandra et al., 2025). Prior work documents the gap but does not diagnose which detector will fall into it; we propose no new detector, and instead give a mechanism for the gap and a diagnostic framework evaluated through the reported case studies.

Shortcut learning and spurious correlations. The shortcut-learning account (Geirhos et al., 2020; Shah et al., 2020; Sagawa et al., 2020) attributes OOD failure to reliance on dataset-specific cues, and the shortcuts this paper measures are established prior work. In particular, Smeu et al. (2025) discovered the leading-silence artifact in FakeAVCeleb and AV-Deepfake1M, showed that audio and audio-visual detectors degrade when it is trimmed, and proposed AVH-Align — real-data-only audio–visual alignment — to circumvent such biases; our leading-trim axis packages their trim evaluation as one paired-bootstrap axis among four, and their supervised AV-HuBERT head appears in our suite, retrained on FAVC and on AV1M (§4). Müller et al. (2021) document silence artifacts in audio-only anti-spoofing, and Dong et al. (2023) identity leakage in visual detectors. What this paper adds is not another shortcut: we test the aggregate account *as a diagnostic framework* and find that no shortcut scalar captures consistent patterns across detectors (§3), and we replace the single-shortcut audit with multi-route diagnosis and an intervention study of how training data redirects evidence use (§4). Ordinal risk ranking remains a sensitivity analysis (Appendix H). Concurrent work reinforces both halves of this position. Single-cue mechanism studies show that shortcut *identity* matters — a provenance watermark alone induces three coupled failure modes (Müller & Debus, 2026); in our terms, a single aggregate reliance score would not separate them — and suppression methods operate on an aggregate "shortcut subspace" within one model (Wang et al., 2026), the quantity that, across detectors, does not predict transfer (§3). Separately, Boldişor et al. (2026) probe modern self-supervised representations and find them attending to semantically meaningful regions rather than superficial artifacts such as leading silence, attributing the residual in-the-wild gap to intrinsic dataset difficulty; read in our vocabulary this is a statement about feature families at the clean end of the route spectrum — consistent with the near-zero artifact axes of our correspondence block (Table 1) — whereas our suite spans heterogeneous routes, where the mixture sweep (§4.1) shows route-latching is real, responsive to training data, with different transfer outcomes.

OOD performance prediction. A separate line of work predicts a *fixed* model's accuracy on an unlabeled target sample — e.g., Projection Norm (Yu et al., 2022), with ODP-Bench (Yu et al., 2025) benchmarking ten such methods — which requires running the model on (unlabeled) target data. Our setting is complementary: a held-out detector's target scores are withheld from predictor fitting, while the risk model is calibrated on reference detectors' labeled outcomes on that same target. This is held-out-family ranking on a known target; the separate leave-one-target-out results are weaker (Appendix H). The comparison is not only architectural: score-based members of that family, computed from each detector's own unlabeled target scores (ATC- and DoC-style statistics (Garg et al., 2022; Guillory et al., 2021)), are sign-inconsistent across our targets — best cell $|\rho|{=}0.54$, and the score-distribution shift *anti*-predicts on AVLips (-0.77) — so even with target access, no such scalar predicts consistently (Appendix H). The FAVC diagnostic-AUC row of Table 4 is itself the accuracy-on-the-line baseline (Miller et al., 2021) in rank form; agreement-based and calibration-based estimators (Baek et al., 2022; Białek et al., 2025) require the same target access this setting withholds.

Domain adaptation theory. Classical bounds (Ben-David et al., 2010) tie target error to a source–target divergence term that is rarely measurable in practice. Our per-route deltas can be read as a heuristic empirical proxy for that term, exact only in a linear-Gaussian model (Appendix M).

D Detector suite and axis coverage

Table 6 lists the 19 suite instances with their detector group, training source, and per-axis coverage in the released benchmark tables. The seven starred probes share one frozen AV-HuBERT backbone (Shi et al., 2022) and constitute a single detector family in every leave-one-family-out split (Appendix H); the leading-trim axis additionally carries three RawNet2 retraining seeds. The ten families: the shared-backbone probe/pooled group (7), AVH-Align (2), SpeechForensics+unsup_sync (sync lineage), AVAD, VFD, RawNet2, CViT2+Xception (frame CNNs), BA-TFD+, FTCN, LipFD.

The two AVH-Align instances are the supervised head of Smeu et al. (2025) (`AVH_Sup`: linear projections of the frozen AV-HuBERT streams, concatenated into an MLP, per-frame real/fake cross-entropy with log-sum-exp pooling), run from their released code unchanged and trained once on FAVC and once on AV1M; the real-data-only alignment model of that paper is not in the suite. The table groups describe detector design or training objective; both AVH-Align heads remain in the supervised group. These groups are distinct from an instance's inferred evidence route. Candidate routes use the measured FAVC signature and are tested by the direct interventions in §2. The candidate labels are descriptive, so grouping by them is not independent confirmation of a route–OOD relationship. The training interventions and pairing tests supply the separate mechanism evidence. SpeechForensics and AVH-Align also consume AV-HuBERT features; a stricter partition merging every AV-HuBERT-derived instance into one family is stress-tested in Appendix H. LipFD entered last as the tenth family with its own frozen insertion prediction (Appendix H); its training corpus AVLips is a transfer target, disclosed and handled as a home pair (Appendix H). FTCN entered the study as the pre-registered ninth family (Appendix H); its silence and skew axes are measured by the benchmark pipeline, and its identity contrast enters at the value recorded at its pre-registration (Table 6, marked $^{\mathrm{f}}$). Two ensemble route substrates used by the error-case audit (level-contrast and offset ensembles) appear in the release but are not suite detectors and are excluded here.

Table 6: The detector suite: detector group, training source, and predictive-axis coverage (Sk = modality skew, Si = leading trim, Id = composition-matched identity contrast). The fourth benchmark axis, generator-distribution overfit, is excluded from the predictor (§2) and not tabulated here. FTCN's silence and skew are measured by the benchmark pipeline; $^{\mathrm{f}}$: its identity contrast is carried at the value recorded at its pre-registration (Appendix G).

Detector group	Detector	Training source	Sk	Si	Id
Correspondence	SpeechForensics (Liang et al., 2024)	self-supervised AV speech features; no deepfake supervision	✓	✓	✓
Correspondence	VFD (Cheng et al., 2024)	voice–face matching on real identities	✓	✓	✓
Correspondence	AVAD (Feng et al., 2023)	self-supervised on real talking-head video	✓	✓	✓
Correspondence	unsup_sync	unsupervised offset-prediction model trained on AV1M real clips; no deepfake supervision	✓	✓	✓
Audio-only	RawNet2 (Tak et al., 2021)	ASVspoof 2019 LA	✓	✓	✓
Audio-only	AVH probe-audio*	AV1M (frozen backbone, linear head)	✓	✓	✓
Audio-only	pooled_audio_linear*	FAVC (frozen backbone, linear head)	✓	✓	✓
Naive concat	AVH probe-concat*	AV1M (frozen backbone, linear head)	✓	✓	✓
Naive concat	AVH linear probe*	AV1M (frozen backbone, linear head)	✓	✓	✓
Naive concat	pooled_jointav_linear*	FAVC (frozen backbone, linear head)	✓	✓	✓
Visual-only	AVH probe-visual*	AV1M (frozen backbone, linear head)	✓	✓	✓
Visual-only	pooled_video_linear*	FAVC (frozen backbone, linear head)	✓	✓	✓
Visual-only	CViT2 (Wodajo & Atnafu, 2021)	DFDC	✓	✓	✓
Visual-only	FTCN (Zheng et al., 2021)	FaceForensics++	-0.308	$+0.047$	-0.056^{f}
Visual-only	LipFD (Liu et al., 2024)	AVLips (home corpus; disclosed pair)	✓	✓	✓
Visual-only	Xception (Rössler et al., 2019)	FaceForensics++	✓	✓	✓
Supervised	BA-TFD+ (Cai et al., 2023)	LAV-DF	✓	✓	✓
Supervised	AVH-Align (FAVC) (Smeu et al., 2025)	FakeAVCeleb (supervised head, retrained)	✓	✓	✓
Supervised	AVH-Align (AV1M) (Smeu et al., 2025)	AV-Deepfake1M (supervised head, retrained)	✓	✓	✓

E INTERVENTION PROTOCOLS

Silence trim. The trim is the official protocol of the work that discovered the shortcut (Smeu et al., 2025), commit-pinned: onset T is the time of the first sample with $|x| > 5 \times 10^{-4}$ on the 16 kHz decode, and $\lfloor T \cdot 25 \rfloor + 1$ whole 25 Hz frames are removed from *both* streams whenever $T > 0$ (a deliberate ≤ 40 ms over-cut), so the trimmed clip stays stream-aligned. Feature-level detectors apply the cut by in-memory slicing at feature time (bit-equivalent to the reference implementation); raw-media detectors consume a media-faithful analog (same rule; all intervention arms share one encode lineage, priced by the encode-only control). A detector-free manipulation check confirms the cut removes the cue: the trivial prefix-amplitude probe falls from AUC 0.99 to 0.62 on the cue-bearing cells, the residual being the amplitude-graded floor the protocol's own authors report. The duration-matched trailing control removes the same per-clip frame count from the clip end.

Identity splits. The identity-disjoint split removes every test clip whose target or FaceSwap-donor identity is seen in training (3,367 of 6,667 clips kept), audited for label, category, generator, and duration composition. Because identity filtering alone shifts demographic composition by up to 21.1 points, the reported axis is the composition-matched clean contrast: paired identity-seen vs. identity-unseen controls matched across 170 exact strata (1,271 clips each).

Mixture sweep. The forced-routing sweep of §4.1 holds the architecture (frozen AV-HuBERT backbone, one logistic head) and the *training-set size* fixed: at every mixture share λ the head sees N=1,800 training clips, $\lfloor \lambda N \rfloor$ drawn from the speaker-disjoint FAVC pool (all manipulation cells) and $N - \lfloor \lambda N \rfloor$ from a manifest-disjoint AV1M pool (all cells, labels from the corpus), so λ varies which corpus supplies the training fakes and never sample size or the presence of visual forgeries. Identity and generator composition within each pool are *randomised* across the three seeds rather than stratified by construction; the reported curve is the seed mean, and the seed spread is shown in Figure 2. Label balance follows the two source pools and is therefore constant in λ only to the extent that the pools are balanced — a randomisation-level control, not a construction-level one, and stated as such. *Replication on an independent FV–RA cell.* The same fifteen trained heads, scored on the DFTIMIT test cell (n=960: 320 real, 640 fake; Idiap source, no FAVC/AV1M/VoxCeleb2 lineage), give concat-head FV–RA AUC (seed means) $0.489/0.599/0.662/0.710/0.858$ at $\lambda = 0/.25/.5/.75/1$, against $0.495/0.855/0.895/0.935/0.980$ on FAVC; the audio-only head reads exactly 0.500 at every λ because DFTIMIT's fakes reuse the real utterances, and the visual-only head is not monotone (0.58–0.75). Training draws are identical to the published sweep (the FAVC values reproduce to three decimals for every seed).

Matched and same-source sweeps. Three controlled arms hold what the natural sweep randomises. *A1, matched sweep*: the training unit is one source video with its genuine clip and one clip of each manipulated type (RV–FA, FV–RA, FV–FA). For FAVC the four clips share the target identity and source video; for AV1M, whose manipulated segments are distinct segments of the same source video, one segment per type is chosen once (seed 777). At every λ the training set is 300 units, $\mathrm{round}(300\lambda)$ from FAVC and the rest from AV1M (N=1,200, class ratio 1:3, cell ratio 1:1:1, identical identities in every stratum), drawn round-robin over identities within each source (AV1M candidates: 600 videos from 600 identities; FAVC: 347 complete units from 347 identities, test identities excluded); FAVC FV–RA clips are cycled over generator tokens (Wav2Lip, FSGAN, FaceSwap). AV1M generators are not recoverable from clip identifiers and are not stratified. The unit design keeps within-identity real/fake contrast, which a one-clip-per-identity draw removes: that draw gives the same concat curve but an unstable AV1M-only visual head. *A2, same-source audio-cue recombination:* FAVC only; 300 reals and 900 FV–RA fakes drawn round-robin over identities, FV–RA balanced over generator tokens; a share ρ of the fakes has its pooled audio features replaced by those of a fake-audio clip (FV–FA or RV–FA) of the same target identity, so the visual forgery, identity set, generators, class ratio and N are fixed and only the availability of an audio-side label cue changes. The concat head is the concatenation of the pooled streams, so the recombination is exact for the concat, audio and visual heads; the fused feature cannot be recombined and is omitted from this arm. *A3, same-source natural cue sweep:* as A2, but the share ρ of fakes is drawn from the natural FV–FA cell (Wav2Lip video with cloned audio), a comparator whose visual generator is confounded. Three seeds per value; readouts on the FAVC test FV–RA and RV–FA cells and on DFTIMIT.

Outcome. Table 7 lists the seed means. The matched sweep keeps the natural curve's shape with class and cell balance and the within-identity real–fake pairing fixed by construction (identities differ between the two corpora). Its visual-only head trained on AV1M units alone reads 0.80 on FAVC FV–RA (seeds $0.86/0.72/0.80$), the same as under the natural draw ($0.80/0.81/0.78$) and below the suite's whole-pool AV1M visual probe (0.907, Table 1), while the concat head on the same units reads 0.61 ($0.64/0.55/0.64$). In the recombination arm the audio-only head stays at chance on FV–RA at every ρ (the test audio is real), the visual-only head stays at 0.99, and the concat head keeps its visual discrimination until the cue covers every fake (seeds at ρ=1: $0.85/0.82/0.90$ on FAVC, $0.71/0.68/0.75$ on DFTIMIT). The natural FV–FA comparator moves the skew the same way but keeps honest AUC at 0.95, as its visual forgeries are FAVC's own. On the two other representations (same speaker split, seeds and draws as the earlier single-run analyses, N=1,500) the visual-only reference values are the FV–RA AUCs of single-stream heads trained on the whole FAVC train split:

VFD visual 0.747, concat 0.734; BA-TFD+ visual 0.659, audio 0.661, concat 0.746. Along the VFD sweep the audio-only head reads 0.51–0.54 at every λ while the visual-only head rises $0.52 \rightarrow 0.67$, so the concat rise is carried by the visual stream; the BA-TFD+ visual-only head stays at 0.60–0.65 (Table 7).

Table 7: **Controlled and cross-backbone sweeps** (seed means, three seeds). A1–A3 on frozen AV-HuBERT features: concat-head FV–RA AUC on FAVC and DFTIMIT and FAVC skew. VFD and BA-TFD+: concat and visual-only heads, FAVC speaker-split FV–RA AUC.

	A1 matched (λ)			A2 recombination (ρ)			A3 FV–FA share (ρ)		VFD (λ)		BA-TFD+ (λ)	
value	FAVC	DFTIMIT	skew	FAVC	DFTIMIT	skew	FAVC	DFTIMIT	concat	visual	concat	visual
0	0.612	0.599	$+0.39$	0.991	0.904	-0.49	0.991	0.900	0.528	0.515	0.679	0.617
0.25	0.891	0.747	$+0.11$	0.990	0.910	-0.10	0.991	0.902	0.538	0.546	0.701	0.643
0.5	0.947	0.789	$+0.05$	0.988	0.898	-0.09	0.990	0.896	0.559	0.592	0.717	0.647
0.75	0.960	0.822	$+0.04$	0.987	0.891	-0.05	0.988	0.892	0.580	0.595	0.719	0.648
1	0.987	0.885	$+0.00$	0.856	0.715	$+0.14$	0.945	0.820	0.649	0.672	0.710	0.598

F THE EXCLUSION AUDIT

Two candidate corpora were excluded from specific roles for measured cause. **AV1M as a skew-measurement domain.** The skew axis requires a FV–RA cell whose audio carries no class signal. On AV1M's nominally-clean FV–RA cell an audio-only probe separates fake from real at $\Delta = +0.859$ where ≈ 0 is required — the cell is audio-contaminated by content-level manipulation — so AV1M is excluded as a *measurement* domain for skew while remaining a transfer target. The consequence of ignoring this audit is measurable: a route signature whose skew axis is measured on AV1M is null on every target ($\rho = -0.43$ to -0.84, at or below the refit null centre; Appendix H), which is itself evidence that the signature must be measured on a modality-clean corpus. **DFDC.** It lacks an RV–FA cell, so it cannot supply the full factored-cell diagnostic signature. This restriction does not prevent its use as an OOD target; a face-swap target evaluation remains a separate test of the present transfer scope.

Cell-purity matrix. The AV1M audit generalizes: for every corpus we run audio-only probes on the FV–RA cell and visual-only probes on the RV–FA cell (flag: any probe deviating ≥ 0.10 from chance). Clean: FAVC FV–RA (0.535 worst probe), DFTIMIT FV–RA (0.500), VidTIMIT RV–FA (0.482). Flagged: DeepSpeak FV–RA (audio probe 0.721 — the AV1M failure mode, so DeepSpeak honest-subset readings are upper bounds), LAV-DF FV–RA (0.656) and RV–FA (0.716), and FAVC RV–FA (0.604). Flagged cells are disclosed wherever their numbers are used. One flagged cell touches an axis: the skew axis's RV–FA arm on FAVC sits at the flag boundary (0.604 — the shared sample set does not imply a shared effect on every detector, so ranking invariance cannot be assumed). The cross-corpus axis comparisons in Appendix H are sensitivity checks, not a clean control isolating this contamination.

G PRE-REGISTRATION RECORD

Five predictions were frozen to files before the corresponding runs; every outcome, including the falsifications, is reported.

- **FTCN (ninth family; frozen prediction file).** Predicted collapse rank 3 of 14; observed rank 1 with collapse $+0.583$; point error 0.30, outside the pre-registered envelope. Verdict: ordinal hit, quantitative miss; the diagnostic claim is downgraded to ordinal (Appendix H).
- **DeepSpeak suite prediction (sixth target; frozen file with input hashes).** Before any eligible detector's first inference on DeepSpeak v1.1, collapse predictions for the 16 never-evaluated suite detectors were frozen (target-agnostic pooled predictor trained on the five existing targets; detectors already evaluated there — SpeechForensics, unsup_sync — excluded by rule). Revealed once: $\rho = 0.426$, exact $p = 0.0488$ on the registered pooled outcome. **Counted as a miss** (Appendix H): the corpus's FV–RA cell is audio-contaminated

(audio probe 0.72; Appendix F), and restricted to the clean FV–FA cell the frozen ranking reverses ($\rho = -0.41$; Appendix H), so the registered hit rests on a contaminated cell. The replay enters no headline count and the joint test is reported without it. The freeze predates LipFD's entry into the suite (16 = the then-18-instance cohort minus the two detectors already evaluated there); LipFD carries its own frozen prediction (next item).

- **LipFD insertion (tenth family; frozen prediction file).** Split verdict: z-scored rank prediction 1/17 vs observed 6/17 (miss; uniform-insertion $p = 0.35$), raw-gap point prediction error 0.018 in gap units (hit). Diagnosed as a feature-scaling pathology of the readout (Appendix H).
- **Voice-clone cells (untouched VidTIMIT video + XTTS-v2 audio, 320 units).** All three frozen predictions hit: (i) clone audio desyncs on both channels (level 0.060 / sync-fraction -0.088 vs. RV–RA anchor 0.346 / 0.952); (ii) corr_sync best (0.993) with the untouched-video visual probe at chance (0.511) as the built-in negative control — and the audio probe *below* chance (0.322): the legacy audio artifact does not transfer to 2024 TTS; (iii) the offset collapse reproduces off-FAVC (roll $0/5/8/15$ frames $= 0.993/0.941/0.638/0.443$, inverting outside the $\pm$3-frame window; shuffle $\approx$ chance).
- **DeepSpeak v1.1 (pre-registered runbook, evaluated once, 2,913 official test clips, zero lineage overlap).** Correspondence holds (SpeechForensics 0.846 overall; FV–FA 0.939 / FV–RA 0.817 on a cell the purity audit flags as audio-contaminated, Appendix F; zero-parameter sync score 0.809); the level-channel separation reproduces (real 0.118 vs. fake cells $0.063/0.031$). The registered falsifier *triggered*: the AV1M-real-trained raw-score readout scored 0.303 against the frozen band $[0.72, 0.86]$. The three-seed follow-up, registered after a target-informed pilot, changed only the readout on fixed checkpoints and passed its four checks (peak readout 0.68–0.70; FAVC held 0.96–0.98; no inverted dataset; seed variance 0.023). This demonstrates sensitivity to readout choice; it does not rescue the original target-blind prediction or establish an invariant underlying ordering (§6). Known structural limits of this corpus: no RV–FA cell, so only the level channel is probeable there; offset does not collapse detection ($0.60 \rightarrow 0.73$), consistent with §5.3.

One further prospective n=1, recorded for completeness: with the pipeline frozen, the never-before-scored SpeechForensics was predicted to gap $+0.195$ and observed $+0.173$ (error 0.021) — a hit, but a single instance, and not cited in the main text.

H Sensitivity analyses of signature-based risk ranking

H.1 Setup and sensitivity summary

These exploratory analyses test how signature-based risk rankings change with the detector-family partition, outcome definition and measurement corpus. They are sensitivity analyses of the diagnostic instrument, separate from the intervention evidence in §3–§6. Eight to ten families per target, mixed pre-registered replays and partition-sensitive joint tests limit their statistical reach. The post-hoc purified variant is part of this sensitivity analysis, not an independent replication. Table 8 lists every check with its outcome; the paragraphs that follow give the protocols for the checks that bear on the primary statistic, and the supplementary material gives the remaining protocols in full.

Primary statistic and status (read this first). The registered statistic is the family-level route signature on the raw axes, the bold row of Table 4: held-out ρ positive on all five targets, one of five cells below p=0.05, five-target joint p=0.013 on the ten-family partition. Every other row of that table and every paragraph below is a sensitivity check on this one statistic; none of them changes a main-text conclusion. Two checks weaken it: merging the eleven AV-HuBERT-derived instances into one family (p=0.08) and the DeepSpeak prospective replay (a miss). The purified rows, the increment rows and the remaining checks leave its sign and approximate size in place. Final status, as stated in §7: an exploratory ranking instrument, not evidence for the mechanism.

Setup. The diagnostic signature of detector d is $x_d = (\text{silence}, \text{identity}, \text{skew})$, measured only on FAVC (§2); the outcome is the collapse $y_{d,t} = \text{AUC}_d(\text{FAVC}) - \text{AUC}_d(t)$ on target t, checked also on the rank of target AUC because the gap contains the FAVC base. A ridge regression fit on the

reference families predicts the held-out family's collapse, scored by Spearman ρ with a one-sided permutation p (10,000 refit draws). Under this refit convention a held-out ρ carries the sign the model learned on the reference families, so a negative cell is a null, not a reversed prediction; the raw fixed-direction correlations of §3 can differ in sign. The protocols, variants and audits follow below; the per-target cells are in Table 4 (Appendix B).

What holds and what does not. Added to an accuracy-only model, the signature raises instance-level held-out ρ by 0.31–1.02 on four of the five targets, while accuracy alone predicts on one (AV1M). At the family level the registered signature is positive on all five targets and every single scalar and the accuracy baseline is null on at least one; modality skew is the strongest single axis where the shift is modality-driven and reverses where it is not (VidTIMIT), which is the multi-route finding of §3 in predictive form. The family-level cells are individually weak (one of five below p=0.05 for the registered signature); the five-target joint test rejects on the ten-family partition (p=0.013) but not when the eleven AV-HuBERT-derived instances are merged into one family (p=0.08; purified variant 0.04); and the pre-registered replays are mixed: FTCN, a fresh family, was predicted third-worst of fourteen and observed worst; LipFD, a fresh insertion, was priced within 0.018 AUC while its rank prediction missed; DeepSpeak is counted as a miss, since the frozen ranking held only on a cell later found audio-contaminated and reverses on the clean cell ($\rho = -0.41$; Appendix G). Table 8 collects every robustness check in one place. The partition and measurement-site sensitivities limit any general risk-ranking claim; these results are not independent support for the main mechanism.

Target-score baselines. Score-only adaptations of the OOD-accuracy-estimation family (Garg et al., 2022; Guillory et al., 2021) (mean shift, quantile W_1 shift, and threshold-crossing shift of each detector's own unlabeled target scores) fail by sign, not strength (Table 9, Table 8): a score-distribution shift records that the target moved the detector, not whether the channel it reads survived. They are adapted here from same-model accuracy prediction to cross-detector ranking, so the failure bounds the adaptation, not the original methods.

Family-level protocol (N=8–10). The ridge is fit on family means (one point per family) and validated leave-one-family-out; every permutation draw shuffles the family-level outcomes and refits the complete held-out pipeline (10,000 draws, one-sided): LAV-DF $\rho_{\mathrm{fam}} = 0.588$, $p = 0.051$ (N=10); AV1M 0.350, $p = 0.14$ (N=9); AVLips 0.524, $p = 0.098$ (N=8); DFTIMIT 0.588, $p = 0.041$; VidTIMIT 0.200, $p = 0.22$ (both N=10). At this N the one-sided critical value is 0.55–0.62 and the minimum detectable effect at 80% power is $\rho \approx 0.8$–0.9, so per-target cells are read for sign and the inferential statement is the joint test below. Dropping the three detector–target pairs that coincide with a detector's training corpus moves no conclusion: LAV-DF without BA-TFD+ $\rho = 0.200 \rightarrow 0.220$ (N=18); AV1M without AVH-Align(AV1M) $0.465 \rightarrow 0.444$ ($p = 0.041$, N=17); AVLips without LipFD $0.650 \rightarrow 0.586$ ($p = 0.013$, N=15).

Home-pair exclusion. BA-TFD+ on LAV-DF, AVH-Align (AV1M) on AV1M and LipFD on AVLips are evaluated on their own training corpus, so their gaps are not OOD collapses. Dropping each from its own target's universe leaves the registered signature positive on all five targets and the joint test rejecting (S=0.26, p=0.007), but the families common to every cell fall from eight to six and the purified variant is null there (Table 8). The raw fixed-direction signs do not change under the exclusion, so the negative raw skew→gap relation on LAV-DF and AV1M is the base term in the gap, not the home pair: audio-routed families start from a low FAVC AUC and have less to lose. Table 4 keeps the full universe.

Base-free outcome on the full universe. Rerunning the family-level protocol with the rank of target AUC itself as outcome (no FAVC term) gives, for the purified signature, ρ_{fam}=0.77/0.63/0.41/0.38/0.46 across LAV-DF/AV1M/AVLips/DFTIMIT/VidTIMIT — positive on all five, LAV-DF p=0.010, AV1M p=0.053 — and for the leading signature 0.55/−0.13/0.69/0.36/0.29 (AVLips p=0.040). The skew scalar is significant on three targets (0.75/0.38/0.64/0.52, LAV-DF p=0.004) but reverses hard on VidTIMIT (−0.71), and in-domain accuracy is null to negative on every target (−0.02 to −0.97). The family-level accuracy increment is positive on all five and significant on LAV-DF ($\Delta\rho$=+1.56, p=0.023). So removing the base

Table 8: **Robustness ledger for the risk-ranking evidence.** Family-level held-out ρ_{fam} in the order LAV-DF/AV1M/AVLips/DFTIMIT/VidTIMIT unless stated; "registered" is the raw three-axis signature frozen in the pre-registration files, "purified" its net-silence substitution. Joint p is the five-target permutation test. "Suppl." rows have their protocol in the supplementary material.

Check	Outcome	Where
Primary (ten families, gap outcome)	registered $0.59/0.35/0.52/0.59/0.20$, one of five $p<0.05$, joint $p=0.013$; purified $0.81/0.57/0.74/0.39/0.38$	Table 4; protocol, joint inference
Home-pair detectors dropped	registered $0.47/0.62/0.64/0.59/0.20$, all positive, joint $S=0.26$, $p=0.007$ on six common families; purified $0.83/0.43/0.61/0.39/0.38$, joint null ($p=0.10$)	home-pair exclusion
AV-HuBERT-derived instances removed	$0.50/0.26/0.80/0.82/-0.39$ (VidTIMIT reverses); purified $0.71/-0.43/0.90/0.43/-0.61$; the skew scalar shows the same four-positive pattern	backbone exclusion
AV-HuBERT-derived instances merged into one family	joint $p=0.077$ ($S=0.19$); purified $p=0.039$ ($S=0.28$); instance-level skill intact ($0.83/0.71/0.75$ on LAV-DF/AV1M/AVLips)	joint inference
Base-free outcome (rank of target AUC)	registered $0.55/-0.13/0.69/0.36/0.29$; purified $0.77/0.63/0.41/0.38/0.46$, all positive	base-free outcome
FAVC base residualised / placebo outcome	joint $S=0.103$, $p=0.10$; placebo (FAVC AUC as outcome) scores 0.71	joint inference
Two-axis signature re-measured on the TIMIT pair	LAV-DF $\rho=0.55$ ($p=.006$) vs. 0.76 on the matched FAVC reference; null where the reference is null. The matched reference differs from Table 4 (0.20) because the outcome is re-based to the pair ($0.39 \rightarrow 0.76$ at the instance level) and identity is dropped	corpus swaps
Skew axis measured on LAV-DF	positive on four of six targets at 27–89% of the FAVC-measured axis; Spearman 0.618 with the FAVC skew axis over 18 instances	corpus swaps
Pre-registered replays	FTCN: predicted third-worst of 14, observed worst; LipFD: gap within 0.018, rank missed; DeepSpeak: miss (clean cell $\rho=-0.41$)	Appendix G; post-reveal audit
In-domain AUC as predictor	held-out signal on AV1M only; every shortcut scalar null on at least one target	Table 4
Skew alone vs. the full signature	skew outranks the signature on LAV-DF and DFTIMIT (instance 0.50 vs. 0.20; 0.52 vs. 0.29) and reverses on VidTIMIT (-0.30 vs. $+0.65$; family -0.66 vs. $+0.20$); nested permutation: the other axes cost LAV-DF ($\Delta\rho=-0.30$, $p=.041$) and gain VidTIMIT ($+0.94$, $p=.003$). The signature is the hedge, not uniformly superior	Table 4; suppl.
Target-score baselines (ATC, DoC, W_1 shift of own target scores)	ATC uninformative (best $\lvert\rho\rvert=0.42$, $p=0.086$); DoC/W_1 anti-predict on AVLips ($-0.74/-0.77$) and predict on VidTIMIT ($+0.54/+0.51$), n.s. elsewhere	Table 9
Refit sign vs. raw sign	permutation null centred below zero (median -0.45 to -0.73 for single axes, -0.18 to -0.24 for the signature), so a negative cell is a null; raw fixed-direction single axes change sign across targets (skew $-0.86/-0.70/+0.64/+0.55/-0.14$)	suppl.
Net-silence control-arm purity	trailing silence is class-associated in FAVC's fake-audio cells (trivial AUC 0.67 vs. 0.71 leading), which biases the control against the effect; naive and controlled leading deltas agree on every instance (audio probe $+0.039$ under both)	suppl.
Axis-subset attribution	no subset dominates: LAV-DF $0.202/0.251/0.200$ for {net, id, skew} / {trail, id, skew} / registered; AVLips registered leads (0.650 vs. $0.591/0.518/0.435$)	suppl.
Effective rank of the axis basis	participation ratio 2.55 (instance) / 2.11 (family) for the raw basis; the net-silence decomposition raises the family-level rank to 2.83; an encode axis adds nothing	suppl.
Ordering vs. pricing (LOFO MAE / constant MAE)	prices four targets partially (0.67–0.80); orders without pricing on LAV-DF (1.07) and DeepSpeak (1.16)	suppl.
Predictor form (ridge $\alpha\in\{10^{-8}, 1, 10\}$, rank ridge)	every fitted form keeps a positive sign on every target ($\rho=0.14$–0.79); a fixed-direction z-sum flips sign on three targets	suppl.
Suite composition (drop any one detector or family)	instance-level bands LAV-DF $[0.10, 0.55]$, AV1M $[0.29, 0.71]$, AVLips $[0.50, 0.68]$ with no sign change; DFTIMIT $[-0.14, 0.49]$ with one (without FTCN)	suppl.
Measurement noise (parametric bootstrap of the axis CIs, 1,000 draws)	sign robust on all five diagnostic targets ($P(\rho\leq 0)=.003/.048/<.001/.033/.049$), not on DeepSpeak ($.21$)	suppl.
Partition (leave-one-detector-out)	$\rho=0.80/0.89/0.74/0.51/0.63$; leaves backbone siblings in the training fold, partition check only	joint inference
Selection regret	with a correspondence SOTA in the pool, top-1 by FAVC AUC is within 0.05 of the oracle on four of six targets; without one, signature-ranked selection beats AUC-ranked selection on four of six (LAV-DF 0.72 vs. 0.40)	suppl.

term costs the leading signature one target (AV1M) and costs skew its VidTIMIT sign; the purified signature is the only predictor that keeps a five-positive pattern on both outcomes.

AV-HuBERT-backbone exclusion. Eleven suite instances read the same frozen AV-HuBERT feature stack: the seven probe/pooled heads, both AVH-Align supervised heads, SpeechForensics, and

Table 9: Instance-level companion to Table 4. Instance-level Spearman ρ (N=16–19) for the signature and its purified variant. We report these as descriptive only, because the instances are correlated (seven probes share one backbone). The last row is the strongest target-score baseline we tried: an ATC-style threshold-crossing shift computed from each detector's own unlabeled scores on the target (Garg et al., 2022), with two-sided p. This baseline uses the held-out detector's target scores; the signature predictor uses reference detectors' target outcomes.

	LAV-DF	AV1M	AVLips	DFTIMIT	VidTIMIT
Signature, instance level	0.200	0.465	0.650	0.288	0.647
purified net-silence substitution	0.312	0.480	0.391	0.234	0.657
ATC-shift target-score baseline	+0.42 (.086)	+0.28 (.26)	−0.04 (.87)	+0.33 (.20)	+0.28 (.28)

the unsupervised sync model. Dropping all of them from every target's universe leaves 6–8 instances in 5–7 backbone-independent families. The signature stays positive on four of five targets and reverses on VidTIMIT (ρ_{fam}=0.50/0.26/0.80/0.82/−0.39; DFTIMIT p=0.024); the purified variant reads 0.71/−0.43/0.90/0.43/−0.61 (LAV-DF p=0.042, AVLips p=0.008). So the five-positive pattern depends on the AV-HuBERT families, and on this universe the skew scalar shows the same four-positive sign pattern (0.79/0.14/0.90/0.68/−0.82): the sign-consistency separation between signature and best scalar does not survive the cut. What does survive is §3's negative — every scalar still changes sign across targets (silence −0.46 on LAV-DF vs. +0.82 on DFTIMIT) — and the instance-level accuracy increment, significant on LAV-DF, AVLips, and DFTIMIT ($\Delta\rho$=+0.98/+1.09/+0.55, p=0.012/0.032/0.025), null on AV1M and VidTIMIT; on the base-free outcome the signature is positive on four of five with LAV-DF significant (0.82, p=0.019). The VidTIMIT reversal has a mechanical reading: after the cut nearly every remaining instance sits at chance there (AUC 0.49–0.55, against MRDF 0.87–0.90 and AVAD 0.925), so the outcome ranks carry almost no variance to explain. Table 4 keeps the full universe; this cut bounds its composition bias.

Net-silence purification. The reference protocol removes video frames along with the audio prefix (§2), so the raw silence axis needs a control before any silence attribution. An equal-length *trailing*-trim counterfactual (identical per-clip frame counts cropped from the tail, leading silence preserved, same encode lineage) is scored with every covered instance, and the net contrast, leading minus trailing, is the purified axis. The factorization is route-consistent: audio-route deltas concentrate at the leading edge (audio probe +0.039 leading vs. −0.000 trailing), visual-route deltas are edge-symmetric duration and re-encode responses (every visual net contrast $|\Delta| \leq 0.004$; an encode-only control on the 1,985 zero-onset clips reproduces them), and AVAD is edge-symmetric (net +0.000). Substituting the net contrast for the raw axis gives the purified rows of Table 4 (LAV-DF ρ_{fam}=0.81, p=0.006; AVLips 0.74; AV1M 0.57; DFTIMIT 0.39; VidTIMIT 0.38); at the instance level the substitution lifts LAV-DF, AV1M and VidTIMIT and gives back the duration-carried share on AVLips and DFTIMIT. The control arm's own purity and the axis-subset attribution are ledger rows.

Joint inference. The per-target nulls share detector families, so the global statement is one joint permutation test rather than per-target Holm. The statistic (fixed before computation, after the per-target results were known) is the mean family-refit ρ across the five targets on the eight families common to every cell; each of 10,000 draws shuffles the family→outcome assignment once and refits every pipeline. Leading-axis signature: S=0.295, p=0.013 (per-cell components LAV-DF/AV1M/AVLips/DFTIMIT/VidTIMIT = 0.52/0.00/0.52/0.19/0.24); the median, the mean with the strongest cell dropped, and the min/max-trimmed mean also reject (p=0.023/0.004/0.012). With DeepSpeak as a sixth cell, which the audit of Appendix G excludes, S=0.321, p=0.0075 (sensitivity only). A maxT audit over the full target×axis-variant grid confirms that the claim lives at the joint level: no single cell survives familywise correction at this N. Two controls: against a placebo outcome (the FAVC diagnostic AUC itself) the same pipeline scores 0.71, so the signature predicts the in-domain base that the gap outcome contains; residualizing the family-level outcomes on that base leaves the joint skill positive but not significant (S=0.103, p=0.10), whereas holding accuracy in the model (the increment rows of Table 4) gives significant gains on four targets. On the merged partition (the eleven AV-HuBERT-derived instances as one family; six families common to the five

targets) the statistic is S=0.19, p=0.077 for the registered signature and S=0.28, p=0.039 for the purified variant, so the joint significance depends on counting the AV-HuBERT-derived families separately. Leave-one-detector-out keeps the ordering (ρ=0.51–0.89) but leaves backbone siblings in the training fold and is a partition check only.

Silence-availability atlas. The detector-free trivial silence-AUC (fake cells vs. the corpus's reals, chance = 0.5): FAVC 0.68, AV1M 0.74, LAV-DF 0.507, DFTIMIT exactly 0.500 (its fakes reuse the real utterances), AVLips 0.52, and VidTIMIT-XTTS 0.145 and DeepSpeak FV–FA 0.018, both *inverted* because generated audio starts speech sooner than real recordings. DFTIMIT separates presence from availability: 27% of its clips carry $\geq$200 ms leading silence with a null label association, while its shipped fake tracks carry a $\sim$17 ms amplitude-prefix artifact (availability AUC 1.0, verified upstream of our processing). An availability shift is necessary for the silence channel to matter on a target, not sufficient, and availability is protocol-graded.

Measurement-corpus swaps. Every signature is measured on FAVC, selected by the cell-purity audit; a skew axis measured on a corpus that fails the audit carries no held-out signal (Appendix F). Three swaps re-measure axes off FAVC. (i) *Two-axis signature on the TIMIT pair.* The purity-clean pair (DFTIMIT FV–RA and VidTIMIT-XTTS RV–FA sharing 320 RV–RA anchors; 1,280 clips) is the only non-FAVC site where the reference trim protocol runs on every suite detector; identity is unavailable there. With outcomes re-based to $\mathrm{AUC}(\text{pair}) - \mathrm{AUC}(t)$ and the FAVC-measured two-axis signature refit on the identical 18-detector universe (10,000-draw permutation nulls), the swap predicts where the reference predicts, at reduced strength (LAV-DF ρ=0.55, p=.006, against 0.76), and is null where the reference is null (AV1M, AVLips, DeepSpeak). For media-level detectors the pair's silence axis is not a silence measurement: DFTIMIT's fakes reuse the real utterances, and the fake AVI tracks carry the amplitude-prefix artifact of the atlas, so that axis records artifact reliance in the broad sense of §2. (ii) *Skew axis on LAV-DF*, the only other corpus with both cells (a moderately impure pair; Appendix F): with outcomes re-based to LAV-DF and BA-TFD+ disclosed as a home pair, the axis is positive on four of six targets at 27–89% of the FAVC-measured axis on identical universes and null-to-negative where the reference is; the LAV-DF and FAVC skew axes agree at Spearman 0.618 over 18 instances, related but not invariant. (iii) *Skew axis on the TIMIT pair* tracks the FAVC one (Spearman 0.67) but keeps skill only on LAV-DF. These are three measurement sites, one the design anchor: they show measurement-site sensitivity and cannot isolate cell purity from other corpus differences, FAVC's own RV–FA flag included.

DeepSpeak post-reveal audit. After the frozen 16-detector readout (ρ=0.426, exact p=0.0488), a per-target LOFO on the full 19-instance universe scores -0.17 there. Two causes are identified: the added detectors include unsup_sync, whose DeepSpeak entry is the readout-inversion case of §6 (deployed-readout AUC 0.34 vs. raw-route 0.81, so the gap outcome conflates route failure with readout inversion), and the corpus's FV–RA cell carries audio-side class signal (audio probe 0.72; Appendix F). Restricted to the clean FV–FA cell the frozen ranking reads $\rho = -0.41$ — a sensitivity bound rather than a refutation, since the restriction also changes the attack population (359 fakes, one engine); residualized on the FAVC base the family-level cell is positive (0.55). The frozen readout is reported as recorded; these audits are why the replay is counted as a miss (Appendix G).

I Additional boundary conditions

Exploratory target-informed readout follow-up. The registered DeepSpeak readout failure (0.303 against $[0.72, 0.86]$) remains a miss. After inspecting DeepSpeak outcomes, a three-seed follow-up changed the absolute-logit readout to a relative offset-peak readout on the same AV1M-real-trained checkpoints, reaching 0.68–0.70 without retraining. Target labels informed readout selection as well as evaluation. This shows sensitivity to readout choice; it is exploratory and does not establish target-blind transfer, a requirement for target-domain refitting, or an invariant underlying ordering. The result remains below the untouched correspondence statistics (0.809–0.846) and is excluded from the main transfer comparison. The registration record is in Appendix G.

The restored head does not transfer either. The restore intervention of §5.2 trains a logistic readout on the 8-d correspondence feature; its sweep form is the λ=0 endpoint of the mixture sweep, trained on AV1M units. Scored on the DeepSpeak test manifest of Appendix G in the same

exploratory run as the router replication of Appendix K (cached features, no retraining), that readout inverts: FV–RA AUC $0.21/0.24/0.33$ over the three λ=0 seeds, against $0.59/0.65/0.66$ for the collapsed concat head it replaces on the same audio-contaminated cell (Appendix F) and 0.809 for the zero-parameter sync score. On LAV-DF (overall AUC over all cells) the two heads read 0.814 and 0.816. The restore therefore repairs the in-domain route but does not by itself supply a transferable readout; the correspondence transfer results of §5 rest on the zero-parameter and unsupervised scorers, and this failure is a further instance of §6, item 2.

The offset-collapse result is corpus-shaped, not universal. On corpora where fakes are synchronized (DFTIMIT, VidTIMIT-Wav2Lip, DeepSpeak), destroying timing does *not* collapse detection ($0.62/0.70$ under full shuffle; DeepSpeak $0.60 \rightarrow 0.73$) — consistent with the level channel carrying the signal (§5.3), but it bounds the causal story of §5.2 to corpora with a genuine timing cell.

Route labels do not order every detector. VFD is balanced on the route map yet collapses on LAV-DF; AVAD is weak throughout ($0.845 \rightarrow 0.731$). Relative performance also varies among detectors assigned to the same candidate route.

The route pays an in-domain price. corr_sync does not beat naive fusion in-domain — this is a diagnostic contribution, not a SOTA method.

J DOSE-RESOLVED SYNCHRONY SUPPLY AND READOUT COMPENSATION

The offset interventions of §5.2 were run to certify that the correspondence route reads synchrony. Held frozen and read across corpora rather than across detectors, the same measurements price a different quantity: how much exploitable synchrony evidence each corpus *supplies*. This appendix reports that reading, and the readout-level compensation that bounds it. Everything here is exploratory — nothing in this section enters a pre-registered test or revises a number in the main text.

Sensitivity to the tested timing breaks. Rolling each corpus's audio by $\Delta \in \{0, \ldots, 15\}$ frames under the frozen FAVC-trained corr_sync route yields, per corpus, a base AUC and a maximum AUC drop over the ladder. The two dissociate, and the dissociation is the point. AVLips reaches base AUC 0.904 with a maximum drop of 0.002 (full shuffle 0.009): a near-perfect separation that survives destroying timing altogether, so the separation survives the tested timing breaks. FAVC (0.872, drop 0.826) and VidTIMIT-XTTS (0.993, drop 0.550) are the honest case. At the other end, DeepSpeak and DeepSpeak-v2 lose no AUC over this roll ladder (maximum drops -0.011 and -0.003; their AUC rises), and AV1M loses only 0.095 from a near-chance base of 0.546 — all three supply little sensitivity to this timing perturbation, consistent with the shuffle-based boundary in Appendix I. A scalar sync-fraction does not draw this distinction, since it reads AVLips and FAVC alike; the dose maximum records the signed response of a frozen scorer to the specified temporal perturbation.

Exploratory six-target supply-feature selection. This is a post-hoc search over six supply features on six targets, not an independent validation of a transfer law. The selected base-separation feature gives Spearman $\rho = 1.00$; the theoretically motivated roll-extinguishable *max drop* gives 0.371. An exhaustive maximum-over-features permutation over all 720 target orderings gives $p = 0.0083$, compared with the unadjusted $1/720 \approx 0.0014$. The closest target pair is separated by only 0.045 on a scale spanning 8.04 (0.56%); swapping it reduces ρ to 0.943. Leaving out one point from this selected monotone sequence preserves $\rho = 1.00$ by construction and adds no validation. A leave-one-target-out linear fit has MAE 0.027 against 0.085 for the train-mean null, but feature selection used all targets, so this is also exploratory. Per-detector fits select different features: base AUC for SpeechForensics/AVAD ($\rho = 0.94/0.83$), half-drop dose for VFD (0.90, $n = 5$). unsup_sync is excluded for insufficient evaluable targets under the readout caveat of §6. The six-target fit is retained as a sensitivity result and is not used to support the main mechanism.

Readout compensation: one route, two desync demands. The supply reading above prices synchrony as *destroyed at zero lag*, which is the right quantity only for a readout that scores at zero lag. Two of our correspondence systems do not. The ladder here is a different readout from the pairing

breaks of Table 11, and the two must not be conflated. Here the audio of *every* clip of the FAVC test set is rolled against its video, reals and all fake cells alike (6,658 clips scoreable from the cache), and the quantity is the overall real-versus-fake AUC of the frozen SpeechForensics readout on that whole set. In Table 11 only the 151 reals are perturbed and the honest-subset AUC scores the intact FV–RA fakes against the perturbed reals, so a real that loses its alignment is scored as a fake and that AUC falls to near zero (0.038 under shuffle, 0.011 under swap, 0.038 at $\Delta{=}25$). Under the whole-set roll the AUC is 0.9894 at $\Delta{=}0$ and 0.9910 at $\Delta{=}15$ (every intermediate change at most 0.0006), steps down to 0.9037 at 20 frames (a drop of 0.0857) and stays there at 25 and 40 frames (0.9084 and 0.9035; drops 0.081 and 0.086); the whole-set shuffle reads 0.9584 (a drop of 0.031). Both classes are displaced together, so most of the separation survives; the size of the residue is not the point here, the shape of the response is. It is a step at the boundary of the system's own ±15-frame shift search — about ±600 ms at 25 fps — not a monotone decline: inside the search window the readout re-aligns the streams and pays nothing. unsup_sync classifies over the same $\tau{=}15$ window. Two consequences follow. First, out-of-window displacement (AUC 0.9035, a drop of 0.086) costs more than full temporal shuffle (0.9584, a drop of 0.031), so clean displacement, not destruction, is this manipulation's ceiling. Second, the supply-side quantity has to be restated for such readouts: what a corpus supplies is not synchrony destroyed at zero lag but *alignment findable within the window*, and the two coincide only for zero-lag scorers such as corr_sync — which is the route §5.2 intervenes on, and which does collapse monotonically. This is the route/readout separation of §6 reappearing on the causal side: one evidence route, two readouts, opposite desync demand.

Availability is not use, one level down. The same ladder run against AVH-Align's two supervised heads separates a route being present in the input from a route being taken. The FAVC-trained head is flat out to 40 frames (every drop ≤ 0.0015; shuffle 0.0018) despite scoring 0.971 in domain: its performance shows no material synchrony dependence under these tests. The AV1M-trained head shows *negative* demand — rolling its audio raises AUC ($0.548 \rightarrow 0.617$ at 40 frames; shuffle -0.055) — so the quantity is signed, and a detector can be hurt by the alignment its training corpus supplied. The corpus-level statement of §5.3 thus recurs at the detector level, which is the more demanding version of it.

Pairing-break verification. Timing insensitivity alone does not settle whether the FAVC-trained AVH-Align head reads audio–visual agreement at some other level, so we break the pairing directly on the 151 FAVC test reals at feature level: the audio frames are rolled by $\Delta \in \{3, 5, 8, 15, 25, 40\}$, permuted in time, or replaced by the audio of another test real from the same ethnicity–gender stratum (FAVC has one real clip per identity, so an identity-matched swap does not exist). Roll and shuffle preserve the audio-frame multiset; swapping changes speaker and content as well as pairing, so its effect is less specific. Two readouts: the AUC of perturbed against intact reals (0.5 = no response), and the honest-subset AUC (FV–RA fakes against the perturbed reals) next to its intact value. A logistic head on time-pooled audio $\oplus$ visual features, trained on FAVC with the test speakers held out, calibrates "no response": mean pooling is permutation-invariant, so it cannot react to roll or shuffle. Table 11 gives seed means (three speaker-split seeds for the shuffles and swaps). The zero-parameter sync score flags every pairing-broken real and its honest-subset AUC collapses; the pooled head sits at 0.500 as it must. Because an aggregate response AUC near 0.5 does not by itself show that individual scores are unmoved, Table 10 adds the paired per-clip shift in units of each scorer's intact-score SD, with clip-level bootstrap intervals. The FAVC-trained AVH-Align head does move, but by little: $+0.10$ and $+0.16$ SD under roll and shuffle on the released checkpoint (intervals exclude zero; 0.00–0.02 SD on the three retraining seeds) and -0.09 to -0.22 SD under swap (a real clip with another speaker's audio looks marginally more real to it), against $+5.4$ to $+5.7$ SD for the sync score; its honest-subset AUC changes by at most 0.006. "Rejected" is therefore a statement of magnitude: the tested pairing breaks move this head's score by at most a fifth of an SD. The AV1M-trained head's shifts lie within ±0.1 SD with intervals covering zero, but its swap honest-subset change is $+0.025$ ($[-0.024, +0.075]$), outside the 0.02 bound of the rule, so its verdict is undetermined. SpeechForensics, recomputed from the same cached streams exactly as its published statistic (max over ±15-frame shifts of the mean per-frame cosine), responds partially at a 15-frame roll, which sits at the edge of its search window, and fully once the roll exceeds it ($\Delta{=}25$: 1.000 / 0.038) and under shuffle and swap. AVAD and VFD score media rather than cached features, so the same three breaks were rendered in the media of the test reals (video untouched; audio rolled by 15 frames, permuted in 40 ms frames, or replaced by a same-stratum real's audio) and each released model was scored once per break through its own pipeline. The media shuffle splices audio

Table 10: **Paired per-clip shifts under the pairing breaks** (151 FAVC test reals, 150 for VFD; the readout behind Table 11). Shift: mean of (perturbed − intact) score in units of the scorer's intact-score SD; Above: fraction of clips scoring above their intact value; Response: AUC of perturbed vs. intact reals; Honest Δ: honest-subset AUC with perturbed reals minus with intact reals. Brackets: 95% clip-level percentile bootstrap (5,000 draws, replicates resampled together, fakes fixed). K replicates: speaker-split seeds for shuffle and swap, times retraining seeds or head instances where the row has several; one media rendering per break for AVAD and VFD.

Scorer	Break	K	Shift (SD) [95%]	Above	Response [95%]	Honest Δ [95%]
corr_sync (zero-parameter)	roll 15	1	$+5.49$ [5.33, 5.64]	1.00	0.998 [0.995, 1.000]	-0.946 [-0.958, -0.929]
	shuffle	3	$+5.41$ [5.25, 5.56]	1.00	0.997 [0.991, 1.000]	-0.934 [-0.946, -0.919]
	swap	3	$+5.69$ [5.53, 5.84]	1.00	1.000 [0.999, 1.000]	-0.961 [-0.972, -0.944]
AVAD (media, released)	roll 15	1	$+0.65$ [$+0.47$, $+0.82$]	0.72	0.702 [0.644, 0.759]	-0.136 [-0.176, -0.095]
	shuffle	1	$+0.55$ [$+0.38$, $+0.72$]	0.69	0.681 [0.621, 0.740]	-0.110 [-0.149, -0.072]
	swap	1	$+0.61$ [$+0.46$, $+0.78$]	0.68	0.696 [0.640, 0.752]	-0.124 [-0.160, -0.089]
VFD (media, released)	roll 15	1	$+0.16$ [$+0.08$, $+0.24$]	0.61	0.555 [0.529, 0.582]	-0.029 [-0.047, -0.012]
	shuffle	1	$+1.34$ [$+1.21$, $+1.47$]	0.94	0.839 [0.805, 0.875]	-0.327 [-0.367, -0.290]
	swap	1	$+0.95$ [$+0.76$, $+1.13$]	0.82	0.760 [0.711, 0.808]	-0.208 [-0.256, -0.157]
pooled head (3 heads)	roll 15	3	0.00 [0.00, 0.00]	0.50	0.500 [0.498, 0.502]	0.000 [0.000, 0.000]
	shuffle	9	0.00 [0.00, 0.00]	0.49	0.500 [0.499, 0.501]	0.000 [0.000, 0.000]
	swap	9	$+0.09$ [$+0.01$, $+0.18$]	0.52	0.495 [0.474, 0.515]	-0.005 [-0.010, -0.000]
AVH-Align (FAVC), released	roll 15	1	$+0.10$ [$+0.04$, $+0.16$]	0.68	0.535 [0.519, 0.557]	-0.001 [-0.002, -0.000]
	shuffle	3	$+0.16$ [$+0.11$, $+0.21$]	0.72	0.550 [0.535, 0.569]	-0.001 [-0.002, -0.001]
	swap	3	-0.09 [-0.16, -0.02]	0.47	0.481 [0.459, 0.503]	-0.000 [-0.002, $+0.002$]
AVH-Align (FAVC), 3 seeds	roll 15	3	$+0.00$ [-0.04, $+0.05$]	0.53	0.506 [0.495, 0.517]	$+0.001$ [-0.001, $+0.003$]
	shuffle	9	$+0.02$ [-0.01, $+0.04$]	0.54	0.506 [0.498, 0.515]	-0.000 [-0.002, $+0.001$]
	swap	9	-0.22 [-0.29, -0.15]	0.37	0.441 [0.422, 0.460]	$+0.006$ [$+0.002$, $+0.011$]
AVH-Align (AV1M), released	roll 15	1	-0.03 [-0.19, $+0.13$]	0.44	0.481 [0.436, 0.528]	$+0.015$ [-0.032, $+0.060$]
	shuffle	3	-0.04 [-0.17, $+0.09$]	0.43	0.478 [0.442, 0.515]	$+0.019$ [-0.017, $+0.055$]
	swap	3	-0.08 [-0.26, $+0.09$]	0.48	0.475 [0.424, 0.526]	$+0.025$ [-0.024, $+0.075$]
AVH-Align (AV1M), 3 seeds	roll 15	3	-0.00 [-0.08, $+0.07$]	0.51	0.501 [0.480, 0.522]	-0.003 [-0.023, $+0.018$]
	shuffle	9	$+0.00$ [-0.06, $+0.06$]	0.49	0.503 [0.487, 0.518]	-0.004 [-0.019, $+0.012$]
	swap	9	-0.08 [-0.21, $+0.06$]	0.48	0.486 [0.444, 0.532]	$+0.013$ [-0.031, $+0.055$]

Table 11: **Pairing-break verification on the FAVC test reals** (the confirmation step of §2). Response AUC (perturbed vs. intact reals; 0.5 = none) / honest-subset AUC (FV–RA fakes vs. perturbed reals) under three pairing breaks; seed means over three speaker-split seeds for the feature-level rows. AVAD and VFD are the released models scored once per break rendered in the media; the media shuffle splices audio frames, so roll and swap are their clean readouts. AVAD is confirmed under the pre-stated rule by its roll response (0.702, 95% interval $[0.644, 0.759]$ against the 0.70 threshold) and is therefore threshold-sensitive; its honest-subset AUC falls under all three breaks (Appendix J). The pooled head cannot read pairing by construction and calibrates "no response". The 15-frame roll sits at the edge of SpeechForensics' ±15-frame search window (at Δ=25: 1.000 / 0.038). The AV1M head's swap change (0.025) exceeds the 0.02 bound of the rule, hence undetermined. Protocol, per-clip scores, and paired per-clip shifts with bootstrap intervals: Appendix J, Table 10.

Scorer	roll Δ=15 resp / honest	shuffle resp / honest	swap (same stratum) resp / honest	intact honest	correspondence
SpeechForensics (statistic)	0.728 / 0.923	1.000 / 0.038	1.000 / 0.011	0.974	confirmed
corr_sync (zero-parameter)	0.999 / 0.023	0.997 / 0.034	1.000 / 0.008	0.968	confirmed
AVAD (media, released)	0.702 / 0.595	0.681 / 0.621	0.696 / 0.607	0.731	confirmed (at threshold)
VFD (media, released)	0.555 / 0.823	0.839 / 0.525	0.760 / 0.644	0.852	confirmed
pooled audio⊕visual head	0.500 / 0.982	0.500 / 0.982	0.495 / 0.977	0.982	calibrator
AVH-Align (FAVC), released	0.535 / 0.985	0.550 / 0.984	0.481 / 0.986	0.986	rejected
AVH-Align (FAVC), 3 seeds	0.506 / 0.981	0.506 / 0.980	0.441 / 0.986	0.981	rejected
AVH-Align (AV1M), released	0.481 / 0.538	0.478 / 0.542	0.475 / 0.548	0.523	undetermined

frames and so adds discontinuities that the feature-level shuffle does not; roll and swap are the clean readouts for these two. VFD responds by $+0.95$ SD under swap (response 0.760, honest-subset AUC $0.852 \rightarrow 0.644$) and by $+1.34$ SD under the spliced shuffle, and weakly to the 15-frame roll

(+0.16 SD). AVAD moves by $+0.55$ to $+0.65$ SD under all three breaks with honest-subset drops of 0.11–0.14 whose intervals exclude zero; its response AUC crosses the 0.70 threshold only at the roll (0.702 $[0.644, 0.759]$), so its confirmation is carried by the consistent direction and honest-subset drop across the breaks rather than by a wide response margin. Verdicts under the protocol of §2: correspondence confirmed for SpeechForensics, corr_sync, AVAD, and VFD; rejected for the tested breaks for the FAVC-trained AVH-Align head; undetermined for the AV1M-trained head (its swap honest-subset change, $+0.025$, exceeds the 0.02 bound; its audio route rests on the modality-stress cells and the audio probe of §4) and for unsup_sync, whose scores come from media through its own pipeline and were not re-scored under a pairing break.

Scope of the descriptive thresholds. The 151 real clips supply paired intact and perturbed observations, not two independent samples of that size. The response-AUC band and honest-AUC tolerance summarize the completed case studies; they do not specify a test with a controlled false-rejection rate and hence carry no formal power guarantee. Power for a future equivalence test requires a specified alternative, paired score covariance and the joint decision rule across interventions. Swaps additionally share donor identities. A prospective analysis must preserve these dependencies and control the interval coverage for the joint claim. The current result concerns the tested pairing breaks and does not distinguish generator-specific fingerprints from other visual cues.

K THE LABEL-FREE ROUTER

Inside the mixture sweep, a router that selects its route from training-side skew alone — no target labels and no target data; the skew uses the source test set's cell labels — matches or beats the best fixed route at every λ and switches at the correct regime boundary (training skew $+0.56 \to -0.06$). On the external DeepSpeak corpus the router does not replicate (concat $\geq$ router). We therefore report it as an in-sweep demonstration that the routing mechanism is operational, not as a deployable method (§4.3).

L SKEW-AXIS REPLICATION

The modality-skew decomposition replicates on an independent corpus (DFTIMIT) and across backbones: the same FV–RA/RV–FA asymmetry structure reproduces under AV-HuBERT-, VFD-, and BA-TFD+-based scorers, so the axis is not an artifact of one feature space or one corpus.

M THEORY DETAILS

Linear-Gaussian routing law. For linear scorers on Gaussian class-conditional features, the honest-subset AUC admits the closed form $\mathrm{AUC}_{\text{FV–RA}} = \Phi\big(\mu \cos\theta_V / \sqrt{2}\big)$, where θ_V is the angle between the learned weight vector and the visual-evidence subspace and μ the class separation; the learned angle follows the training-informative modality. The closed form is numerically exact on synthetic data (mean error 0.0013) and predicts the mixture sweep's shape. For nonlinear scorers the relation is empirical only: the training mix's audio–visual informativeness gap tracks the learned routing skew at Pearson $r = 0.98$ on synthetic data and $r = 0.83$ over the fifteen mixture-sweep runs of §4.1, and is stated as unproven.

Divergence bridge (heuristic). For a fixed-route scorer, the $d_{\mathcal{H}\Delta\mathcal{H}}$ term of Ben-David et al. (2010) corresponds, *within the linear-Gaussian model above*, to the routing-weighted change in manipulation composition between source and target; we have no theorem for the general case and use the correspondence as intuition. Its empirical checks: balanced routing yields an OOD gap of 0.005 and modality-dominated routing 0.33 under the same shift, and each benchmark axis can be read as a proxy for one component of the term. *Scope:* the regime-resolved decision-divergence check favours corr_sync in the collapse regime (-0.014 vs. concat $+0.535$) but the regime split there has $N{=}2$ and is illustrative; the decision-boundary divergence estimate does not favour correspondence uniformly (0.205 vs. 0.144). Hence the narrowed claim we state in the main text: routing through the stable agreement-level signal prevents the audio-dominance collapse.

N ERROR-CASE AUDIT

Deep errors are defined rank-wise over existing per-clip scores: a flagged real scores above the median fake; a slipped fake scores below the median real. Substrates: the raw correspondence route (per-clip two-channel decomposition), SpeechForensics (FAVC and DeepSpeak), and unsup_sync (FAVC/LAV-DF/AV1M/DeepSpeak).

Slipped fakes are level-matched. Fraction of slipped fakes whose agreement level lies in the real range (at or above the real-class lower quartile), route substrate: FAVC 72.2%, LAV-DF 91.5%, AV1M 88.8%, DFTIMIT 96.6% — against $1.5/21.0/51.5/14.6\%$ for caught fakes; SpeechForensics on FAVC: 80.0% vs. 1.5%. On DeepSpeak, FV–RA (level 0.063, nearest the real level 0.118) carries $153/163$ of SpeechForensics' misses (per-cell FN 13.4% vs. FV–FA 2.8%).

Flagged reals are level-depressed, rarely desynced. 73–86% of channel-covered flags sit at or below the fake-median agreement level; genuinely desynchronized reals are 1.4–5.1% of flags (enriched 4–$13\times$ over a 0.4–0.7% base rate but still marginal); silence-heavy reals show no enrichment (FAVC base 17.2%; SpeechForensics deep-flags contain zero FAVC reals; unsup_sync $2/9$, n.s.). 14–27% of covered flags carry no tag (rank noise).

Shape of the surface. Deep-FP rate tracks the corpus's level margin: FAVC $0.7\% \rightarrow$ DFTIMIT $2.2\% \rightarrow$ LAV-DF $12.6\% \rightarrow$ AV1M 28.9% (the thinnest level margin, real-vs-fake W_1 0.05–0.06).

Caveats. SpeechForensics' LAV-DF/AV1M/AVLips per-clip files were pending regeneration at audit time, so its coverage there is absent; the raw-route substrate's channel correlations ($\rho = -0.83$ to -0.99) are partially circular (the covariates are components of the score) and are reported as decomposition, not independent evidence; deep-error rates are rank-based characterizations, not deployment FPR/FNR; flagged clips have not yet been manually reviewed — the released clip list is the spot-check queue.