What Do Audio-Visual Synchronization Metrics Actually Measure?

Jai Kumar Sharma
Virginia Tech
jaisharma@vt.edu

Peeyush Tapadiya
Accenture
peeyush.tapadiya@accenture.com

Abstract

Automatic AV-sync metrics are widely used to rank and train audio-visual generators, but they are rarely audited as measurement instruments. We jointly audit AV-Align, ImageBind AV-relevance, JavisScore, and Synchformer/DeSync under a common reliability protocol: controlled-distortion monotonicity, preprocessing sensitivity, rank uncertainty, cross-metric agreement, PEAVS-proxy agreement, and learned fusion. The result is an axis split, not a single winner: Synchformer/DeSync is the strongest temporal-offset tracker ($\tau = 0.84$), ImageBind/JavisScore better match the PEAVS human-aligned proxy ($\tau = 0.20$) and content-disruption families, and AV-Align is the weakest standalone metric. The metrics mutually disagree (Krippendorff $\alpha = 0.066$), and neither linear nor simple k-NN fusion improves PEAVS agreement over the best individual metric. We recommend reporting AV-sync as a Reliability Card (metric-family breakdowns with confidence intervals) rather than a single bare synchronization score.

1. Introduction

Audio-visual (AV) generation, spanning joint text-to-AV, video-to-audio (Foley), and audio-to-video, is advancing on the strength of automatic *synchronization* metrics. Benchmarks rank models by a single sync number, and preference-optimization pipelines now *train* on a sync signal, so an unreliable metric can corrupt optimization itself. The dominant metrics are AV-Align [11] (optical-flow motion peaks vs. audio-onset peaks, IoU), Synchformer/DeSync [5] (a learned offset predictor), JavisScore [7] (windowed ImageBind AV similarity), and ImageBind AV-relevance [2] (semantic cosine). JavisScore was introduced *because* AV-Align "may produce misleading results" on complex scenes, so the community already suspects metric-specific failure modes, yet the deployed metric family has not been jointly audited under a common reliability protocol.

A synchronization score is a *measuring instrument*: it should respond monotonically to known errors, stay stable under benign preprocessing, agree with related instruments, and align with perceptual judgment. Yet the deployed AV-sync instruments have never been jointly audited on these properties. Synthetic controlled-degradation meta-evaluation, testing whether a metric tracks *known* quality changes, is an established paradigm (SLVMEval [8] for text-to-long-video quality; STREAM [6] for video corner-cases), but it has never been applied to *audio-visual synchronization* metrics, nor combined with cross-metric and stability analysis. Prior AV work either proposes a *new* human-aligned metric and validates only itself (PEAVS [3]; AVBench [10]) or *uses* the metrics to rank models (VABench [4]). The nearest precedent, JavisDiT [7], releases a 3,000-sample human set and shows JavisScore is more *accurate* than AV-Align, validating one proposed metric, not auditing the deployed set for stability, cross-metric agreement, or rank-flip. Our novelty is thus one of *domain and scope*: to our knowledge, the first joint reliability audit of the deployed AV-sync metric family under a common protocol.

We close that gap. Our contributions: **(1)** a reproducible, annotation-free **synthetic-desynchronization oracle** imposing known desync (temporal shift, audio speed, fragment shuffle, intermittent mute) on real clips; **(2)** a **preprocessing-sensitivity harness** quantifying each metric's coefficient of variation under should-not-change resampling and its rank-flip probability; **(3)** the first **cross-metric agreement** measurement across the deployed metrics, plus a PEAVS-proxy comparison; and **(4)** a **learned, conformal-calibrated meta-metric** evaluated against the metrics as baselines, testing (and finding wanting) the hypothesis that they combine into a human-aligned score. The headline: competence is *axis-specific*; Synchformer tracks temporal offset far better than the rest yet is weakly PEAVS-aligned, the metrics mutually disagree ($\alpha = 0.066$), and none wins both axes.

2. Related Work

AV-synchronization metrics. Automatic sync scoring spans two families. *Signal-level* metrics compare low-level motion and audio events: AV-Align [11] matches optical-flow motion peaks to audio-onset peaks by IoU. *Learned* metrics embed both modalities: Synchformer/DeSync [5]

is a transformer trained to predict the audio-visual temporal offset; ImageBind [2] yields a generic cross-modal cosine; and JavisScore [7] aggregates windowed ImageBind similarity, introduced explicitly because AV-Align "may produce misleading results." Both target *semantic* correspondence by design, not onset-level timing (ImageBind's training negatives swap the entire paired audio and encode only two frames), so weak temporal-oracle tracking is expected for them. PEAVS [3] instead trains a *human-aligned* predictor on 120K opinion scores. These metrics are deployed as leaderboard criteria, yet each was validated (if at all) only against its own design goal; none has been audited for stability, cross-metric agreement, or human grounding as a group. We treat all four signal/learned metrics as the black boxes the community actually uses and ask whether their *scores* can be trusted.

Meta-evaluation by controlled degradation. Testing whether a metric tracks *known*, synthetically-imposed quality changes is an established meta-evaluation paradigm: SLVMEval [8] probes text-to-long-video quality metrics, and STREAM [6] stress-tests video metrics on temporal corner cases. This annotation-free oracle has not been applied to audio-visual *synchronization*, nor combined with preprocessing sensitivity and inter-metric agreement; we port it to AV-sync and add a PEAVS-proxy axis.

AV generation and benchmarks. Video-to-audio and joint AV generation (MMAudio [1], ASVA [12], JavisDiT [7]) are ranked by the very sync metrics we audit; benchmarks (VABench [4], AVBench [10]) aggregate them but validate at most one proposed metric, not the deployed set. Our audit is orthogonal, and adds a conformal-calibrated [9] meta-metric tested as a would-be combiner.

3. Method

We score each metric as a black box, $f(\text{clip}) \in \mathbb{R}$ (higher = better synced).

Synthetic-desync oracle (agreement without humans). On a pool of real, well-synced clips we apply a controlled desync of known magnitude m from four families (temporal shift, audio speed, fragment shuffle, intermittent mute), each an increasing-desync grid, so the ground-truth ordering is known by construction. The oracle is a monotonicity test: as *added* distortion grows the score should fall, regardless of the clip's baseline sync. Per clip and family we compute Kendall τ between the metric score and *negative* perturbation severity (i.e. the ideal "more-distortion-is-worse" order), so that $\tau = 1$ denotes perfectly monotonic degradation as distortion increases and $\tau = 0$ denotes no response; a faithful metric yields $\tau \to 1$. We report per-family mean τ with clip-bootstrap 95% CIs. The four families are PEAVS [3]-inspired synchrony-related distortions; only temporal shift is a pure global AV offset, while speed, shuffle, and mute probe broader temporal-structure and content-disruption sensitivity.

Preprocessing-sensitivity harness. A metric should be invariant to changes that do not alter true sync. We resample clips (random fixed-length crops; length truncation) and report the coefficient of variation (CV). We further compute the **rank-flip probability**: under clip-bootstrap, how often a metric fails to order two conditions that differ by one known desync step, a direct test of whether single-number gaps of that size are within noise.

Cross-metric agreement. Treating each metric as a "rater" of clip sync, we compute pairwise Kendall τ (rank-based) and Krippendorff α (interval) on per-clip scores that are first *z-scored per metric*, so agreement is scale-free and unaffected by the metrics' differing ranges; we use PEAVS's human inter-annotator $\alpha \approx 0.71$ as an anchor.

Learned calibrated meta-metric (method). We test whether the metrics can be *combined* into a human-aligned score: a regularized linear model (ridge) regresses the human-aligned PEAVS score on the per-clip metric scores, evaluated strictly out-of-fold (5-fold and leave-one-out, no leakage), with split-conformal prediction intervals for calibrated per-clip uncertainty. We compare its held-out Kendall τ vs. PEAVS to every single metric.

4. Experiments

Clips. AVSync15 [12] (15 classes, 1,500 real highly-synced clips from VGGSound); the primary audit uses 75 clips (5/class), the subset on which PEAVS scoring and all auxiliary analyses (crop sensitivity, generated-audio pairing, qualitative inspection) are aligned; a 2× replication (150 clips) reproduces every conclusion (Synchformer temporal $\tau = 0.87$, $\alpha = 0.07$; §6). **Metrics.** We reimplement AV-Align from the published TempoTokens [11] algorithm (optical-flow motion peaks vs. audio-onset peaks matched by IoU) with its default peak-picking parameters (validated against the official code, App. B); ImageBind AV-relevance and JavisScore use the official ImageBind encoder (JavisScore: 2 s windows, 1.5 s overlap, mean of the 40% least-synced windows). For Synchformer/DeSync we use the released checkpoint and score each clip as $-|\mathbb{E}[\Delta]|$, the negative absolute *expected* audio-visual offset under the softmax over the model's offset-class grid (App. A: the split is unchanged under $-\mathbb{E}[|\Delta|]$ and the modal offset); inputs are reencoded to the released 25 fps / 256px / 16 kHz spec and loop-padded to the fixed 5 s window. Every other metric uses its native preprocessing; scores are oriented so higher = better synchronization. **Stats.** Kendall τ-b, ≥1k-resample clip-bootstrap 95% CIs, preprocessing-sensitivity CV, bootstrap rank-flip. All reproducible from committed code, seeds, and SLURM scripts.

Metric	Oracle tracking, Kendall $\tau \uparrow$				Preproc.	Rank-flip $\downarrow$		PEAVS
	Shift	Speed	Shuffle	Mute	CV $\downarrow$	Shift	Worst-fam.	proxy $\tau \uparrow$
AV-Align [11] (reimpl.)	0.21	0.04	0.01	0.24	0.54	0.66	0.96	0.00
ImageBind-rel. [2]	0.16	0.37	**0.38**	0.61	0.15	0.54	**0.69**	**0.20**
JavisScore [7]	0.39	0.51	0.31	**0.73**	**0.14**	0.46	0.90	**0.20**
Synchformer/DeSync [5]	**0.84**	**0.76**	0.27	0.59	n/a	**0.19**	1.00	0.07

Table 1. **Main reliability scorecard** on AVSync15 (best per column **bold**; higher τ, lower CV/flip better). Synchformer/DeSync leads temporal tracking; ImageBind/JavisScore lead PEAVS-proxy agreement (a tie); AV-Align is the weakest standalone metric. **No metric wins both the temporal-oracle and PEAVS-proxy axes.** Rank-flip *Shift*/*Worst-fam.* = pure-shift vs. maximum-family adjacent mis-ordering; Synchformer CV n/a (fixed $\geq$ 5 s input). The AV-Align row is our reimplementation; official TempoTokens scoring gives the same aggregate conclusion (App. B). CIs are in the supplement.

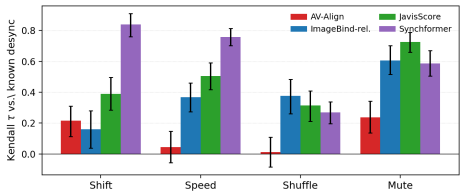

Figure 1. Agreement with controlled desynchronization (Kendall τ, 95% CI). **Synchformer leads** temporal and audio-speed tracking, while JavisScore/ImageBind respond more strongly to intermittent mute; AV-Align is weakest overall (CI crosses zero on audio-speed).

5. Results

Table 1 consolidates the full audit. The main conclusion is an *axis split*, not a single winning metric; we walk through it below.

5.1. Metrics split by reliability axis

Oracle tracking (Fig. 1). The dedicated learned offset predictor **Synchformer tracks the pure temporal-offset axes far better than the rest**: temporal-shift $\tau = 0.84$ $[0.78, 0.90]$ and audio-speed 0.76, versus 0.16–0.39 (temporal) for the others. But it does *not* generalize to content disruption: on shuffle ImageBind leads (0.38 vs. Synchformer's 0.27) and on mute JavisScore leads (0.73 vs. 0.59). So metric competence is *axis-specific*: the offset predictor owns temporal ordering, the embedding metrics own content disruption, and AV-Align is the weakest oracle tracker overall (near chance on audio-speed, 0.04, and shuffle, 0.01).

5.2. Small leaderboard gaps are not always resolvable

Preprocessing sensitivity and rank-flip. AV-Align is the *only* high-variance metric (median CV 0.54 vs. $\leq$ 0.15). The *worst-family* adjacent flip (bootstrap probability of mis-ordering two adjacent desync levels, i.e. consecutive grid magnitudes; max over non-identity pairs) is high for every metric: 0.96/0.69/0.90/1.00 (AV-Align/ImageBind/JavisScore/Synchformer), but this is each metric's weak family. On the *temporal-shift* ranking that matters most, the flip is far lower for Synchformer (0.19, and 0.00 on speed) than for the others (0.46–0.66); its 1.00 is shuffle alone. So Synchformer is reliable for temporal ranking even though no metric is reliable across all families. Adjacent-level discriminability ($d' = \Delta\mu/\sigma_{\text{clip}}$, the mean score gap between adjacent levels over clip noise) agrees: only Synchformer clears $d' \geq 1$, on temporal shift ($d' = 1.6$); elsewhere $d' < 1$ (adjacent levels overlap $> 16\%$). Leaderboard gaps below a metric's minimum detectable difference (MDD, the smallest gap whose bootstrap CI excludes zero; 13–17% of range) are noise.

Real generations: close models break the similarity metrics. We generate audio for 45 AVSync15 videos with three MMAudio [1] checkpoints and ask whether each metric reliably ranks them (paired clip-bootstrap flip probability, Table 2). A *far* pair (small-16k vs. large-44k) is easy for all metrics, but for a *close* pair (large-44k vs. large-44k-v2), the leaderboard regime of similar systems, the similarity metrics' rankings fall within noise while **only Synchformer reliably separates the generators**. Re-running the oracle on the 45 generated clips reproduces the ranking (Synchformer best, $\tau = 0.76$; AV-Align worst), so the conclusions are not an artifact of natural clips. This is not a full MMAudio benchmark but a leaderboard-resolution stress test: paired bootstrap uncertainty on matched generated outputs.

Metric	Far model gap	Close model gap
AV-Align	≤ 0.09	0.35
ImageBind-rel.	≤ 0.09	0.33
JavisScore	≤ 0.09	0.35
Synchformer/DeSync	≤ 0.09	**0.08**

Table 2. **Leaderboard risk on generated outputs**: paired clip-bootstrap flip probability $\downarrow$. Far model gaps are robust across metrics; a close pair falls within similarity-metric noise, and only Synchformer separates it.

5.3. No single score or simple fusion solves the problem

Cross-metric disagreement (Fig. 2, left). Treating each metric as a rater, the four agree almost not at all: Krippendorff $\alpha = 0.066$ (z-scored; rises to 0.21 on the controlled grid, App. C), far below the inter-annotator agreement reported in PEAVS (≈ 0.71). Every pairwise Kendall τ is near zero except ImageBind-rel. vs. JavisScore (+0.78), which is expected and *not* independent (shared ImageBind backbone). Crucially, Synchformer agrees with the embedding metrics and AV-Align at $\tau \approx 0$: an oracle-accurate metric orders real clips unlike any deployed one. A PCA confirms the four metrics span *three* orthogonal axes (only the shared-backbone ImageBind/JavisScore pair duplicates, $r = 0.93$): a single sync number projects three dimensions onto one. On already-synced clips this disagreement is over synchronization *evidence*, not *quality*; it persists under the oracle and on generated clips, so it is not a clean-clip range artifact.

Two distinct axes: oracle vs. PEAVS (Fig. 2). Scoring the same clips with PEAVS [3], a human-aligned automatic proxy reported to correlate with human ratings, Synchformer (best at the oracle) agrees with PEAVS at only $\tau = 0.07$, while ImageBind/JavisScore tie for the highest PEAVS-proxy agreement ($\tau = 0.20$) yet only moderately track the oracle; AV-Align is ≈ 0 on both. **No metric is good at both, and the two axes are nearly orthogonal**: detecting controlled desynchronization and matching the PEAVS-aligned perceptual proxy are different problems. Scene-complexity stratification (clips split at the median optical-flow magnitude, equal halves) is diagnostic: AV-Align is weaker on low-motion than high-motion clips ($\tau = 0.18$ vs. 0.25), consistent with its reliance on motion peaks, while JavisScore's PEAVS agreement is much higher in high-motion than low-motion scenes ($\tau = 0.48$ vs. 0.04); Synchformer's temporal lead is scene-invariant.

A learned meta-metric cannot fix it. A ridge meta-metric over the four scores, evaluated strictly out-of-fold with split-conformal intervals, never beats the best single metric (held-out τ vs. PEAVS ≤ 0.12 vs. 0.20). This is not a limitation of *linear* fusion: a simple nonlinear combiner (leave-one-out k-NN) does no better (≤ 0 vs. PEAVS). In our sample, then, neither linear nor simple nonlinear fusion improves PEAVS agreement over the best single metric: naive fusion is insufficient here, though it does not rule out supervised calibration with direct human labels.

6. Discussion and Guidance

We audit on AVSync15; the axis-specific split also replicates on a second, in-the-wild domain (VGGSound, App. G). The temporal oracle uses audio-lag shifts; a bidirectional variant (audio lead/lag, released harness) preserves the ranking. Our perceptual grounding uses the PEAVS *metric* (human-aligned, Pearson 0.79), a scalable proxy, not fresh human labels; PEAVS is itself a learned model and may share representational biases with the embedding metrics, so we read this axis as *PEAVS agreement*, not perceptual ground truth. Direct human preference annotation remains future work before any claim of perceptual superiority. The findings are not sample-size artifacts: the 150-clip replication (§4) reproduces every structural conclusion. The oracle imposes *controlled* distortions, not generative artifacts (complemented by the MMAudio audit above).

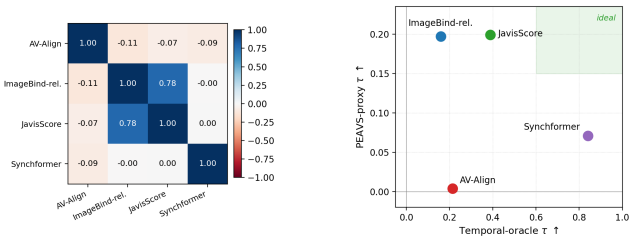

Figure 2. Left: cross-metric agreement ($\alpha = 0.066$); only the shared-backbone embedding pair agrees. Right: the two axes, oracle tracking (x) vs. PEAVS agreement (y), are nearly orthogonal; no metric is high on both.

Reliability axis	Report
Temporal oracle	τ + CI
Content/distortion oracle	τ + CI
Preprocessing sensitivity	CV + CI
Ranking resolution	flip prob.
PEAVS proxy	τ + CI
Cross-metric consistency	τ / α
Direct human calibration	pairwise acc. (if avail.)

Table 3. **Reliability Card**: report AV-sync on these axes, not one bare number (full version, App. E).

Guidance. The AV-generation community ranks (and increasingly trains) models on synchronization metrics that mutually disagree ($\alpha = 0.066$ vs. ≈ 0.71 among PEAVS's human annotators). Do not report AV-Align alone: the weakest, most preprocessing-sensitive standalone metric; for *temporal* synchronization prefer Synchformer/DeSync ($\tau = 0.84$), but its competence is *axis-specific*: it fails on content-disruption families. No metric wins both axes, and no meta-metric combines them beyond the best single one. The field should stop reporting AV-sync as a single bare number; until directly human-calibrated metrics exist, authors should report a *Reliability Card* (Table 3): metric-family scores, uncertainty, and rank resolution.

References

[1] Ho Kei Cheng et al. MMAudio: Taming multimodal joint training for high-quality video-to-audio synthesis. *CVPR*,

2025. arXiv:2412.15322. 2, 3

[2] Rohit Girdhar et al. ImageBind: One embedding space to bind them all. *CVPR*, 2023. 1, 2, 3

[3] Lucas Goncalves et al. PEAVS: Perceptual evaluation of audio-visual synchrony grounded in viewers' opinion scores. *ECCV*, 2024. arXiv:2404.07336. 1, 2, 4

[4] Daili Hua et al. VABench: A comprehensive benchmark for audio-video generation. *CVPR*, 2026. arXiv:2512.09299. 1, 2

[5] Vladimir Iashin et al. Synchformer: Efficient synchronization from sparse cues. *ICASSP*, 2024. arXiv:2401.16423. 1, 3

[6] Pum Jun Kim et al. STREAM: Spatio-temporal evaluation and analysis metric for video generative models. *arXiv:2403.09669*, 2024. 1, 2

[7] Kai Liu et al. JavisDiT: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. *arXiv:2503.23377*, 2025. 1, 2, 3

[8] Ryosuke Matsuda et al. SLVMEval: Synthetic meta-evaluation benchmark for text-to-long video generation. In *CVPR*, 2026. arXiv:2603.29186. 1, 2

[9] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *JMLR*, 2008. 2

[10] Jialiang Yang et al. AVBench: Human-aligned and automated evaluation benchmark for audio-video generative models. *arXiv:2605.24652*, 2026. 1, 2

[11] Guy Yariv et al. Diverse and aligned audio-to-video generation via text-to-video model adaptation. *AAAI*, 2024. arXiv:2309.16429. 1, 2, 3

[12] Lin Zhang et al. Audio-synchronized visual animation. *ECCV*, 2024. arXiv:2403.05659. 2

Appendix (supplementary material)

A. Synchformer scoring-reduction ablation. The main text reduces the predicted offset posterior to $-|\mathbb{E}[\Delta]|$, which can cancel for a bimodal posterior. Two non-cancelling reductions, $-\mathbb{E}[|\Delta|]$ (expected absolute offset) and $-|\Delta_{\arg\max}|$ (modal offset), leave the temporal/perceptual split intact (Table 4): temporal-oracle τ stays high and PEAVS τ stays low under all three, so Synchformer's conclusions do not depend on the reduction.

Synchformer reduction	Temporal-shift τ	PEAVS τ
$-\lvert\mathbb{E}[\Delta]\rvert$ (main)	0.84	0.07
$-\mathbb{E}[\lvert\Delta\rvert]$	0.81	0.10
$-\lvert\Delta_{\arg\max}\rvert$ (modal)	0.78	−0.05

Table 4. The temporal/perceptual split survives the score reduction (75 clips).

B. AV-Align: official vs. our implementation. We ran the official TempoTokens AV-Align through the same oracle (Table 5). Every *aggregate* conclusion holds under the official code: both versions are weak oracle trackers (temporal $\tau \approx 0.22$; near-zero on speed/fragment), highly variable (CV 0.34–0.54), unreliable for temporal ranking (temporal flip ≈ 0.67), and near-zero PEAVS. Per-clip scores correlate only weakly across the two (Kendall $\tau = 0.11$, $n = 45$), reflecting AV-Align's sensitivity; because the aggregate audit agrees, the main table reports our reimplementation and the AV-Align conclusions are robust to implementation, not a reimplementation artifact.

AV-Align	Temp.	Speed	Frag.	Mute	CV	PEAVS
Official (TempoTokens)	0.23	0.10	−0.10	0.35	0.34	−0.04
Ours (main table)	0.21	0.04	0.01	0.24	0.54	0.00

Table 5. Official vs. reimplemented AV-Align through the oracle (75 clips): the aggregate pattern (weak tracking, high variance, near-zero PEAVS) is the same. Temporal rank-flip is 0.68 (official) vs. 0.66 (ours).

C. Cross-metric agreement across settings. All scores are z-scored per metric (scale-free) before Krippendorff α. Agreement rises with the strength of the sync signal (Table 6): near zero on clean clips, 3× higher on the controlled grid where sync varies by design. The low clean-clip α thus reflects subtle clean-clip variation, not uninterpretable disagreement.

Setting	Krippendorff α	n
Clean clips	0.07	75
Controlled desync grid	0.21	1650
Generated (MMAudio)	0.11	45

Table 6. Cross-metric α by setting (z-scored).

D. Preprocessing-sensitivity measure. We report CV over should-not-change resampling on each metric's positive score range, complemented by the scale-free rank-flip probability; the two agree (AV-Align is the outlier on both), so the CV comparison is not an artifact of differing score ranges.

E. Metric Reliability Card. We distill the audit into a reusable reporting standard (Table 7): an AV-sync metric should be characterized on all seven axes, not by a single number. We recommend authors report this card for any metric they use to rank or train models.

Reliability axis	What it tests	Report
Temporal oracle	global-offset sensitivity	τ + CI
Distortion oracle	speed/shuffle/mute sensitivity	τ + CI
Preproc. sensitivity	crop/length invariance	CV + CI
Ranking resolution	leaderboard uncertainty	flip prob.
PEAVS proxy	perceptual alignment	τ + CI
Cross-metric consist.	agreement with peers	rank τ/α
Direct human calibration	direct preference check	pairwise acc.

Table 7. The Metric Reliability Card: the axes any deployed AV-sync metric should be reported on.

F. Qualitative disagreement examples. Concrete AVSync15 clips where the metrics diverge (per-metric z-scores over the audit set): *(i)* a clip the embedding metrics rate well-synced (ImageBind/JavisScore $z \approx +1.0$) but PEAVS rates poorly ($z = -2.2$), semantic AV-relevance mistaken for synchrony; *(ii)* a clip Synchformer flags as badly offset ($z = -6.0$) while AV-Align, ImageBind, JavisScore *and* PEAVS all rate it synced ($z \geq +0.8$), the offset predictor firing where perception sees none; *(iii)* a clip AV-Align rates well-synced ($z = +1.4$) while the other three metrics and PEAVS disagree ($z \leq -0.6$), a spurious optical-flow/onset coincidence. These make concrete the axis-specific, mutually-inconsistent behavior quantified in the main text.

G. Second-domain replication (VGGSound). We rerun the oracle on 75 in-the-wild VGGSound clips (unconstrained YouTube AV, distinct from curated AVSync15). The axis-specific split replicates (Table 8): Synchformer is by far the strongest temporal tracker ($\tau = 0.63$ vs. ≤ 0.27 for the others), the embedding metrics lead on content disruption (ImageBind/JavisScore up to $0.36/0.59$ on fragment and mute), and AV-Align is weakest overall (near-zero on speed and fragment). Absolute τ is lower than on curated AVSync15 (expected for noisier in-the-wild clips) but the *ordering*, the paper's headline, holds.

Metric (VGGSound)	Temporal	Aud.-speed	Fragment	Mute
AV-Align	0.16	0.04	−0.03	0.24
ImageBind-rel.	0.04	0.26	**0.36**	0.46
JavisScore	0.27	0.36	0.25	**0.59**
Synchformer	**0.63**	**0.55**	0.17	0.33

Table 8. Oracle Kendall τ on VGGSound: the temporal (Synchformer) / content (embedding) split and AV-Align's weakness replicate on a second domain.