DEFAULT KERNEL DISTANCES MISS COVARIANCE SHAPE AT THE SAMPLE SIZES IN USE

Anonymous authors
Paper under double-blind review

ABSTRACT

Practitioners compare generative models by distances between embedded sample sets: the Fréchet family, and kernel metrics such as the Kernel Inception Distance (KID) and the Kernel Audio Distance (KAD), which were proposed to avoid its Gaussian assumptions and each ship with a fixed default kernel inherited from the two-sample testing literature. Which differences these defaults can and cannot detect on real embedding distributions has not been measured, to our knowledge, so a practitioner reading a reported value cannot know which differences it reflects; we provide that measurement and the reporting protocol that follows from it. We treat each configuration as a statistical test and inject known distortions into real embedding distributions from symbolic music, audio, and vision. Neither default responds to a trace-preserving covariance change (a rotation) of equal or larger transport magnitude than a mean shift it always detects, and the two fail differently. KAD's median-heuristic bandwidth detects location and scale: it rejects an isotropic contraction matched in transport to the mean shift in every run in every space, and against rotations it is sample-inefficient rather than insensitive, reaching full power at sample sizes that vision datasets afford and per-genre symbolic subsets do not. KID's cubic kernel detects the mean shift and little else: it misses the matched contraction in most runs, shows no recovery at the sample sizes we reach, and does not reject on a real model against its own training data once the location component is removed. The Fréchet distance, run through the same permutation protocol, detects that difference. Aggregated multi-kernel tests restore power against shape at nominal level, except against coverage losses constructed to leave the mean unchanged, and near-duplicate samples corrupt split-based calibration unless the construction controls for them. These measurements give configuration choice a tested basis and yield a reporting protocol that separates three questions a single scalar conflates: whether distributions differ, how they differ, and whether samples are copied.

1 INTRODUCTION

Distances between embedded sample sets are the standard means of generative-model evaluation. The Fréchet family (FID, FAD, and FMD for symbolic music, Heusel et al., 2017; Kilgour et al., 2019; Retkowski et al., 2024) compares Gaussian fits; its kernel-based successors, KID in vision (Bińkowski et al., 2018), KAD in audio (Chung et al., 2025), and CMMD (Jayasumana et al., 2024), replace the Gaussian assumption with a maximum mean discrepancy (MMD) under a fixed default kernel: a cubic polynomial for KID, a Gaussian kernel with the median-heuristic bandwidth for KAD. These defaults are inherited from the two-sample testing literature, where the kernel choice is known to govern which alternatives a test can detect (Gretton et al., 2012; Sutherland et al., 2017). In the evaluation setting, however, the defaults are treated as fixed constants, and what they can and cannot detect on real embedding distributions has not, to our knowledge, been measured.

We provide that measurement. Our unit of evidence is *constructed ground truth*: we take real embedding distributions, inject a single, known distortion (a mean shift, a pure covariance rotation, an isotropic contraction, or a controlled fraction of training-set copies) and ask whether a given metric configuration detects it above a permutation threshold at level $\alpha = 0.05$. Because the ground truth is constructed, detection is a property of the configuration rather than of an interpretation;

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.

because thresholds are permutation-based, results are free of the calibration pathologies that we show affect split-based z-scores.

This paper makes four contributions. First, we measure the power of the KID and KAD defaults against constructed location, shape, and coverage distortions in five embedding spaces and across sample sizes (§4, §5), so that a reported value can be read for what it can and cannot reflect. Second, we transfer aggregated kernel tests from the testing literature to generative-model evaluation and cost them on real embedding geometries (§6), which gives practitioners a level-controlled configuration with no bandwidth guess. Third, we quantify how near-duplicates change split-based calibration and how copying can be detected instead (§7), so that split-based z-scores are not compared across corpora. Fourth, we run the Fréchet distance and a covariance-spectrum diagnostic through the same protocol (§8, §9) and derive a reporting protocol that separates whether distributions differ, how they differ, and whether samples are copied. The findings below are consistent across domains. **(1) The defaults measure the mean.** At the median-heuristic bandwidth, a weak pure covariance rotation is detected in $2/20$, $4/20$, and $1/20$ of runs on CLaMP 2 (768d, symbolic music), VGGish (128d, audio), and PANNs (2048d, audio) embeddings respectively, while mean shifts of matched strength are detected in $20/20$; KID's cubic kernel behaves identically (0–$2/20$ on rotation). The two defaults do not fail alike: a matched-transport contraction is detected by KAD's bandwidth in $20/20$ runs in every space and by KID's kernel in at most $6/20$. Mean-matched mode drop shows the same failure on a coverage loss: deleting half of the largest embedding mode is invisible to the defaults (0–8% of runs) once its incidental mean displacement is removed (§4). **(2) The failure belongs to the configuration; its fix is domain-dependent.** In the three music and audio spaces, power against rotation traces the same curve (Figure 1): a 90–100% plateau around 0.05–$0.1\times$ the median squared distance, collapsing on both sides, its location coinciding with a transition we independently located by an observational method. The plateau does not transfer to vision: on Inception embeddings of CIFAR-10 the best single bandwidth reaches only $26/50$, while the defaults reject in $3/20$ runs at n=780, so a narrower bandwidth is not available as a fallback there. The two defaults then separate as the sample grows. The median heuristic is sample-inefficient rather than insensitive: on the same difference it goes from $3/20$ at n=780 to $17/20$ at n=5,000 and $10/10$ at n=10,000. KID's cubic kernel does not follow: $2/20$, $1/20$, $2/10$ over the same range, and $2/5$ at n=20,000 (§5). **(3) The cubic kernel does not reject on low-rank alternatives.** Because real models differ from their data in a few directions rather than by a full-rank rotation, we constructed differences confined to the leading principal subspace, all of which leave the sample mean unchanged: rotations within the top two and top five components, and rank-one inflation or deflation of the first. KID's cubic kernel misses all four in all three spaces we ran them in, at rates between 0.00 and 0.16, while the median heuristic detects the rank-one rescalings in 33–50 of 50 runs and the window bandwidth and MMDAgg detect the rotations in 30–$50/50$ in music and audio and 15–$50/50$ in vision, where the rank-2 rotation is the hard case. **(4) The gap is not specific to kernel metrics.** Improved precision and recall (Kynkäänniemi et al., 2019) and density and coverage (Naeem et al., 2020), run as permutation tests on the same battery, share it: across five spaces they reject the mean-matched mode drop in 1–6 of 20 runs and the least-displacing variant in 0–2, while detecting the contraction, the intra-mode collapse and the recentered real pair (14–$20/20$; Appendix A.16).

(5) Aggregation removes the bandwidth choice; in vision it is the only fixed-kernel MMD configuration of the battery with power. Aggregated multi-kernel tests (Schrab et al., 2023) at package defaults detect the rotation in $18/20$ to $50/50$ runs in the music and audio spaces and $20/20$ in both vision spaces, at empirical level. They do not detect a mode drop constructed to leave the sample mean unchanged (§6).

(6) Near-duplicates affect two measurements. Deduplicating a symbolic corpus shifts the standard deviation of split-based null statistics at narrow bandwidths by up to $58\times$ while leaving permutation power unchanged, so z-scores against split controls are incomparable across corpora, and conclusions drawn from them (including two of our own intermediate ones, which we retract in §10) do not survive unit-free measurement. Conversely, injected training-set copies produce a monotone, dose-dependent response in a narrow-bandwidth statistic calibrated against *independent* splits ($|z| \approx 0.7 \to 8$ from 1% to 10% copies), to which the median bandwidth and the Fréchet distance respond weakly and MMDAgg does not respond, since copying leaves the marginal distribution unchanged. A nearest-neighbor hit-rate test performs better on identical injections (§7). Near-duplicate sensitivity of narrow kernels is thus a confound to control rather than an instrument to use. **(7) A covariance-spectrum diagnostic tracks the amount of collapse.** Across a sampling-temperature

sweep of a symbolic-music transformer, reference-normalized effective dimensionality rises with temperature (Spearman $+0.90$; Vendi score $+1.00$) with one inversion between $T{=}1.0$ and $T{=}1.2$, while the covariance trace is non-monotonic: scale statistics miss the collapse at low temperature that spectrum shape tracks (§8). **(8) The statistic these metrics were proposed to improve on does not share the failure.** Run through the same permutation protocol, the Fréchet distance detects the recentered real pair in $20/20$ runs under both a plug-in and a shrinkage covariance estimator, in every space we tested, including two where the sample is five times smaller than the dimension and the covariance is singular by construction. In the same comparison, at $n{=}780$ and above, the cubic kernel rejects in $0/20$.

2 Related work

Evaluation metrics. FID (Heusel et al., 2017) and FAD (Kilgour et al., 2019) compare Gaussian fits of embeddings; their small-sample bias motivates unbiased kernel estimators (Bińkowski et al., 2018; Jayasumana et al., 2024), most recently KAD (Chung et al., 2025), which adopts the median heuristic as a distribution-free default. Recent audio work examines KAD's behavior: Jeong (2026) documents sensitivity to the PANNs backbone variant and to implementation choices inside the median heuristic, proposes an optimal-transport alternative validated by perceptual correlation, and, consistent with our power analysis, reports KAD indistinguishable from noise under low-rate rank-1 contamination, a covariance-direction alternative. Biswas & Villemoes (2026) defend the Gaussian fit by showing FAD outperforms MMD in correlating with MUSHRA judgments of codec degradations. Both lines evaluate against signal-level degradations, which move primarily the embedding mean; neither measures power against shape. Our measurements agree with the first line and qualify the second. We reproduce the insensitivity that Jeong (2026) reports and locate it in the kernel and the sample size rather than the backbone. We also confirm the premise of Biswas & Villemoes (2026), that the Gaussian fit is informative, in a sense they did not test: run as a permutation test, the Fréchet statistic detects shape differences that the kernel defaults miss (§9). Our contribution differs: a power analysis against distribution *shape*, where we show the defaults fail in all five evaluation spaces and locate the effective-dimensionality boundary at which they do not (§10). O'Bray et al. (2022) report the same sensitivity in graph generation. Desiderata-based comparisons exist at one fixed kernel and sample size (Xu et al., 2018; Stein et al., 2023; Bischoff et al., 2024). Concurrently, Chen (2026) run FID and KID as permutation tests on Inception features at $n{=}200$ and propose a six-arm diagnostic; we characterize the shipped defaults across three domains and up to $n{=}20{,}000$.

Kernel two-sample testing. That kernel choice governs test power is classical (Gretton et al., 2012; Sutherland et al., 2017). MMDAgg (Schrab et al., 2023) aggregates permutation tests over a fixed collection of kernels and bandwidths with non-asymptotic level control and minimax adaptivity; MMD-FUSE (Biggs et al., 2023) is a related fusion approach. To our knowledge these tests have not been transferred to, or evaluated in, generative-model evaluation; that transfer, and its costing against realistic embedding geometries and sample sizes, is one of our contributions, mirroring how the unbiased kernel estimator itself was transferred from testing to evaluation.

Diversity diagnostics. The Vendi score (Friedman & Dieng, 2023) measures diversity as the exponentiated entropy of a similarity-kernel spectrum. Our diagnostic applies the same functional to the embedding covariance spectrum, normalized by the reference corpus; §8 compares the two directly.

3 Protocol

Embedding spaces. We use three spaces spanning two domains: CLaMP 2 (Wu et al., 2025b) embeddings (768d) of the folk-rnn v2 training corpus of $23{,}636$ transcriptions derived from thesession.org (Sturm et al., 2016); VGGish (Hershey et al., 2017) (128d) and PANNs (Kong et al., 2020) (2048d) embeddings of an eight-genre audio corpus derived from FMA (Defferrard et al., 2017). Unless noted, each comparison uses two disjoint samples of $n{=}780$ per side (the size of the smallest per-genre subset), drawn by a seeded permutation. Later sections extend this roster with two vision instantiations (InceptionV3 features of CIFAR-10 and STL-10), five evaluation spaces in total, and one out-of-domain symbolic encoder (Adversarial-MidiBERT, §10) used for boundary analysis; space counts below refer to this roster.

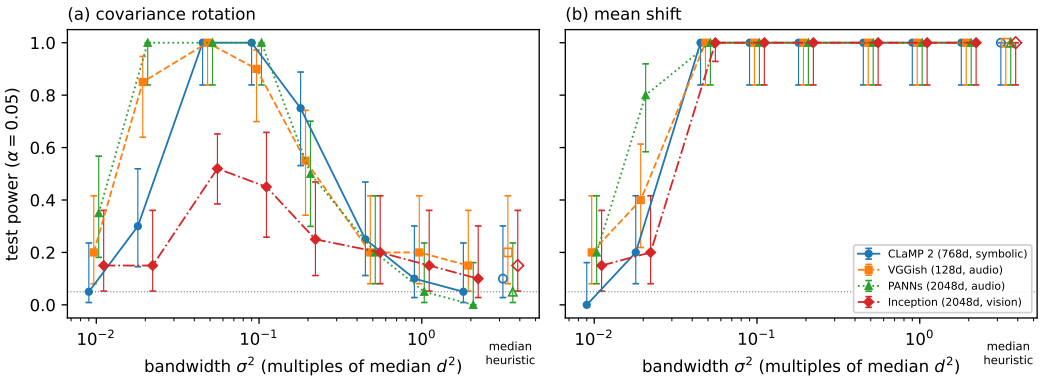

Figure 1: Panels show power against (a) a pure covariance rotation and (b) a mean shift of smaller or equal 2-Wasserstein magnitude (s=0.1, n=780 per side, 20 repetitions, 500 permutations, α=0.05). Bars give Wilson 95% intervals and hollow markers the classical median heuristic. The music and audio spaces share a narrow high-power window near 0.05–0.1× the median squared distance, with both defaults outside it; the vision space has no such window, its best bandwidth reaching $26/50$. Empirical level is $\leq 3/20$ everywhere (Appendix A.2).

Distortion battery. Given a base sample X with mean μ and cloud radius $r = \sqrt{\operatorname{tr}\operatorname{cov}(X)}$, we form a distorted counterpart by exactly one of: *mean shift* $X + s\,r\,v$ with v a random unit vector; *covariance rotation* $(X - \mu)M + \mu$ with M an exactly orthogonal Cayley transform whose skew part is scaled by $s/(2\sqrt{d})$ (location and total variance preserved); *contraction* $\mu + (1 - s)(X - \mu)$; *duplicate injection*, replacing a fraction p of one side with copies of points from the other side, either exact or with 1% per-dimension noise. We report weak (s=0.1) and strong (s=0.5) strengths.

Testing. All detection claims use permutation tests: the unbiased MMD^2 estimator with 500 permutations, α=0.05, and 20 independent draws of sample and distortion per cell, escalated to 50 where a count falls between 3 and 17 in a cell that supports a claim (the mode-drop, low-rank, real-pair, and contested aggregation cells), a rule fixed in advance; the bandwidth grids stay at 20. Larger sample sizes use 10 or 5 draws, where the per-draw cost grows quadratically. Level is reported from undistorted controls. Draws are independent subsamples rather than disjoint ones: in the two audio corpora of 1,999 embeddings two draws have a Jaccard overlap of 24%, so intervals there are descriptive rather than exact. Reference splits are drawn by seeded permutation rather than in corpus order (Appendix A.3). In the recentered arms each side is centered on its own mean once, before the permutation loop, so the test in that arm is conservative (Appendix A.3).

4 What the defaults miss at the sample sizes in use

Figure 1 summarizes the central result. At the median heuristic, power against the weak rotation is $2/20$ (CLaMP 2), $4/20$ (VGGish), and $1/20$ (PANNs), against full power on the matched mean shift. KID's cubic kernel, which has no bandwidth to tune, behaves like the median: rotation $1/20$, $2/20$, $0/20$; contraction 2–$4/20$; mean shift $20/20$; empirical level $\leq 1/20$. In the window $[0.05, 0.1]\times$ median d^2, rotation power is 18–$20/20$ in every music and audio space. The window is two-sided: at 0.01–0.02× median, mean-shift power collapses (0–$4/20$ at 0.01), so no single fixed bandwidth is uniformly safe.

The contrast is conservative in transport units. We match distortion strengths by the 2-Wasserstein distance of the induced population change, computed in the Gaussian approximation from sample means and covariances (the Bures metric). This computation is exact for the mean shift, a pure translation, and for the contraction, whose scaling map is the optimal transport map; for the rotation it is a lower bound on the true W_2 by Gelbrich's bound (Gelbrich, 1990). The rotation at s=0.1 therefore has at least 1.07–1.37× the W_2 of the mean shift it is compared against, and the margin is conservative. Recalibrating the rotation to exactly equal W_2 (bisection; s=0.073–0.093) leaves the result unchanged: the default detects the mean shift in $20/20$ runs and the equal-transport

rotation in $1/20$, $4/20$, and $0/20$ (levels 0–$1/20$). Per unit of W_2, the default's sensitivity to shape is negligible.

Mode drop is detected through its location component, and the construction matters. Removing half of the largest embedding mode (k-means, $k{=}10$) and re-adding the induced mean displacement leaves a difference the default misses: median $4/50$, $1/20$, $4/50$ and cubic $2/50$, $0/20$, $3/50$ in CLaMP 2, VGGish and PANNs. What the narrow bandwidth and the aggregated test see depends on how the deleted points are replaced. Replacing them by resampling the remainder, as an earlier version of this battery did, injects duplicates at a rate we measure at 8–10%, and duplicates alone, with no mode removed, already saturate both configurations ($50/50$ and $20/20$ across the three spaces). Replacing them instead from a disjoint reserve of the corpus removes that confound and lowers the response accordingly: the window bandwidth reaches $34/50$ (0.68, Wilson $[0.54, 0.79]$) in CLaMP 2, $15/50$ (0.30, $[0.19, 0.44]$) in PANNs and $2/20$ in VGGish, and MMDAgg $24/50$, $50/50$ and $0/20$ (controls 0–$2/20$; Table 6).

Low-rank alternatives. Real models differ from their data in a few directions: across seven real pairs the five leading eigendirections of $\Delta\Sigma$ account for 40–87% of its squared Frobenius norm (Appendix A.17), so we built differences confined to the leading principal subspace, all of which leave the sample mean unchanged: rotations within the top two and top five components, and inflation or deflation of the first component by 30%. The cubic kernel misses all four in all three spaces we ran them in, at rates between 0.00 and 0.16 (CLaMP 2, PANNs, Inception/CIFAR-10; 50 draws, Table 7). The median heuristic separates them: it misses the rotations (0.08–0.34) but detects the rank-one inflation and deflation in 33–50 of 50 runs. The window bandwidth and MMDAgg detect the rotations in 30–$50/50$ in the music and audio spaces; in vision the rank-2 rotation is the hard case at $15/50$ and $21/50$. Two controls bound the claim, both run in CLaMP 2: the same inflation applied along a random direction rather than a principal one is invisible to every configuration (0–$2/20$), as is contaminating 2% of points. The effect is confined to directions with substantial variance.

Coverage loss that is mean-quiet by construction. Mean-matching is an intervention on the statistic, so we also built coverage losses that leave the sample mean unchanged by construction. Neither reproduces the dissociation, and they fail in opposite directions: an intra-mode collapse is detected by every configuration except the cubic kernel, and a least-displacing mode drop is detected by none, MMDAgg included (Appendix A.11). The claim that a narrow bandwidth catches coverage loss the defaults miss therefore holds for the subtraction-based construction and does not generalize.

Contraction separates the two defaults. An isotropic contraction changes the trace of the covariance while leaving the mean fixed, and at $s{=}0.1$ it has the same 2-Wasserstein distance as the mean shift of the same s. The median heuristic rejects it in $20/20$ runs in all five spaces; the cubic kernel in 0–$6/20$ (level 0–$3/20$; Appendix A.15). The two defaults therefore fail differently, and the difference is not a matter of sample size: KAD's bandwidth responds to location and scale and misses trace-preserving shape, while KID's kernel responds to little beyond location, as Appendix A.1 predicts.

The same result in the units a practitioner reads. Against a noise floor from twenty independent splits of the reference, and with both metrics computed by their published implementations, the equal-transport rotation moves the reported value by less than one floor standard deviation and the mean shift by nine to twenty-nine; the contraction moves KAD by 6.1 to 20.3 and KID by at most 1.1 (Appendix A.8).

Vision. On InceptionV3 embeddings of CIFAR-10 (the native domain of KID), the defaults again show no power: median heuristic $3/20$, cubic kernel $3/20$ on rotation ($4/20$ on contraction), with empirical levels in range. The high-power window, however, does not transfer: the best single bandwidth ($0.05\times$ median) reaches only $26/50$, decaying on both sides, and this is also a space of higher effective dimensionality ($d_{\text{eff}}{=}251.8$ of 2048). The pattern replicates at higher resolution: on STL-10 ($96{\times}96$, $d_{\text{eff}}{=}342.2$) the median and cubic kernels detect the rotation in $2/20$ (the cubic reaches $5/20$ on contraction), the best single bandwidth reaches $8/20$ while already losing mean-shift power ($17/20$), Gaussian-only aggregation reaches $7/20$, and the full collection again holds $20/20$. Two findings are consistent across all five spaces at this sample size: neither default has power against the rotation, and aggregation does (§6); §5 shows that the first is a statement about n for one default only. The availability of a good *single* bandwidth is specific to the music and audio spaces, which is why we do not recommend one.

The sample-size cost. On synthetic Gaussians in 768d with equal means and covariances differing by a 5% rescaling of a quarter of the dimensions, the median bandwidth has power 0.10 at n=500, 0.46 at n=2000, and reaches 1.0 only at n≈6000; bandwidths in the window, and MMDAgg, are at 0.92–1.0 already at n=500 (level 0.06). The default thus costs more than an order of magnitude in samples on shape differences, which per-genre subsets and per-model generated sets, with 10^3 pieces, do not have.

5 How much data the defaults need

At n=780 per side, the sample size a per-genre symbolic corpus affords, the two defaults behave alike: both miss the shape difference. They separate when the sample grows. We ran the vision battery at $n \in \{780, 2500, 5000, 10\,000, 20\,000\}$ per side, drawing from a pool of 100,000 DDPM samples and the CIFAR-10 training set, with 20 draws per cell up to n=5,000 and 10 and 5 at the two largest sizes.

On a rotation matched to a mean shift in 2-Wasserstein distance, and on the recentered DDPM-versus-CIFAR-10 pair, the median heuristic rejects in $3/20$ and $0/20$ runs at n=780 and in $10/10$ and $10/10$ at n=10,000, crossing half power between 2,500 and 5,000 (Appendix A.18). It is therefore not insensitive to these differences; it needs a sample size that vision datasets afford and a per-genre symbolic subset, at 10^3 pieces, does not.

KID's cubic kernel does not follow. On the same two arms it rejects in $2/20$, $0/20$, $1/20$, $2/10$ and $2/5$, and $0/20$, $0/20$, $0/20$, $0/10$ and $2/5$: on the recentered pair, zero at every sample size up to 10,000, and a rate at n=20,000 that five draws cannot separate from either endpoint (0.4, Wilson $[0.12, 0.77]$). The window bandwidth and MMDAgg reject at or near ceiling throughout (19–$20/20$, $10/10$, $5/5$), and empirical level stays at 0–$2/20$. One further real pair, a text-to-audio model against its reference corpus in PANNs, behaves the same way after recentering (Appendix A.9, §8).

6 Aggregation restores power

We run MMDAgg, at package defaults, through the same battery, at 20 draws per cell (50 for the contested cells). On the weak rotation in raw embeddings it rejects in $18/20$ (CLaMP 2), $50/50$ (VGGish), $20/20$ (PANNs), $20/20$ (Inception/CIFAR-10), and $20/20$ (Inception/STL-10); pooled empirical level over all control runs of both kernel collections is $13/260$ (5.0%). Ablating the kernel collection attributes the power, and the attribution is space-dependent. In music and audio, Gaussian bandwidths alone reach $18/20$, $42/50$, and $20/20$: the bandwidth collection plus a proper permutation threshold provides the power, with the Laplace family adding power only in VGGish (0.84 $[0.71, 0.92]$ vs. 1.00 $[0.93, 1.00]$, Wilson 95% intervals). In vision the roles invert: Gaussian-only collapses to $5/20$ while the full collection holds $20/20$; the Laplace family provides the power in the space where no single Gaussian bandwidth works. Aggregation does not restore power against every alternative: a mode drop constructed to leave the sample mean untouched is invisible to MMDAgg as well, in every space we tested (Appendix A.11).

The recommendation follows the division of labor of the testing literature: an aggregated test with level control answers whether a model's distribution differs from the data, and separately estimated location and dispersion, which have standard estimators and significance machinery, answer how (§8).

7 Two roles of near-duplicates

Role 1: they set the units of split-based z-scores. A common calibration scores a statistic in standard deviations of a null estimated from random splits of the corpus; we used it ourselves before retracting two findings (§10), and background-versus-evaluation splits in FAD pipelines invite it. We deduplicated the symbolic corpus (removing the ∼3% of points whose nearest-neighbor squared distance lies below $0.1\times$ the median d^2) and re-ran a fixed distortion battery with both a split-based z protocol and permutation tests. The permutation power was unchanged by deduplication (rotation at s=0.05: $3/20$ on the original corpus against $6/20$ on the deduplicated one, two cells of the same battery; at s=0.1: $20/20$ both), but the split-based null standard deviations at narrow

Table 1: Cells give the mean response to duplicate injection in control-normalized units (z against independent splits; 20 draws per cell, exact copies; perturbed copies agree within ± 0.5) and MMDAgg rejections ($5/360$ pooled across all cells and both variants, $\approx 1.4\%$). Copying lowers cross-set distances, so responsive statistics go negative; the text discusses magnitudes $|z|$. Narrow is the Gaussian kernel at $0.05\times$ median d^2, Median the median heuristic, and FMD the Fréchet distance on embeddings.

Space	Copies	Narrow	Median	FMD	MMDAgg
VGGish	1%	−0.7	+0.1	−0.4	2/20
	5%	−3.8	−0.7	−0.8	0/20
	10%	−8.0	−0.9	−1.6	0/20
PANNs	1%	−0.6	−0.2	−0.5	0/20
	5%	−3.3	−0.5	−1.0	0/20
	10%	−6.8	−0.9	−1.9	0/20
CLaMP 2 (dedup.)	1%	−0.9	−0.3	−0.3	0/20
	5%	−3.4	−0.3	−1.4	1/20
	10%	−6.3	−0.7	−2.7	0/20

Table 2: Temperature dose-response of a GPT-2-class model on the folk corpus. Cells are ratios to the training reference, means ± s.d. over five disjoint 5,000-sample subsets. eff is the exponentiated spectral entropy of the covariance, n_{90} the number of components covering 90% of the variance, tr the trace, and Vendi the score of Friedman & Dieng (2023) with a cosine kernel.

T	eff ratio	n_{90} ratio	tr ratio	Vendi ratio
0.50	0.79 ± .01	0.80 ± .01	0.84 ± .00	0.75 ± .01
0.70	0.88 ± .00	0.87 ± .01	0.95 ± .01	0.83 ± .01
0.85	0.92 ± .01	0.89 ± .01	1.02 ± .01	0.90 ± .01
1.00	1.08 ± .01	1.07 ± .01	1.00 ± .01	0.99 ± .02
1.20	1.04 ± .01	0.97 ± .01	1.21 ± .00	1.15 ± .01

bandwidths shifted by up to $58\times$, inflating every z-score in the battery, for every distortion type, by the same units change. Split-based z-scores are therefore not comparable across corpora with different duplicate structure, and narrow-bandwidth conclusions should be stated in permutation power.

Role 2: copying is a dependence violation. Table 1 injects training-set copies into one side at $1/5/10\%$. The narrow-bandwidth statistic, calibrated against *independent* splits, responds monotonically and identically for exact and lightly perturbed copies. The median bandwidth changes little. MMDAgg rejects below its nominal level ($5/360$ pooled), which is correct behavior: copying leaves the marginal distribution unchanged and breaks only the independence between sides. The narrow-bandwidth response is nevertheless not the right instrument. A nearest-neighbor hit-rate test detects 2–$9/20$ runs at 1% copies where the narrow bandwidth detects 0–$3/20$, saturates equally above with 2–$3\times$ larger standardized effects, and responds the same to perturbed copies; a global data-copying test in the style of Meehan et al. (2020) responds only at 10%. We therefore treat duplicate sensitivity as a confound to control and recommend NN-based tests for copying (Appendix A.3).

Applied to six generated sets from two model families, including the continuations of a piano transformer pretrained on over a million transcribed recordings, the detector finds no copies: the hit rate against held-out training data is 0.000–0.002 against a calibration of 0.051–0.058 (per-set counts in the released `ex2_organic.csv`), at least eleven standard deviations below it in every set and uncorrelated with temperature.

8 SPECTRUM DIAGNOSTICS TRACK THE AMOUNT OF COLLAPSE

A diagnostic should track the *amount* of a failure as well as its presence. We sample a GPT-2-class symbolic-music transformer (Skierś et al., 2024) (mini-GPT-2 configuration; the reference corpus *is* its training corpus, so ratios read as model-vs-training coverage) at temperatures 0.5–1.2 ($7{,}379$–$10{,}205$ surviving pieces per setting) and embed with CLaMP 2. Table 2 shows that reference-normalized effective dimensionality and the Vendi score rise with temperature over the five settings (Spearman $+0.90$ and $+1.00$; n_{90}: $+0.90$), with a visible inversion between $T{=}1.0$ and $T{=}1.2$; at the training temperature every ratio lies within 8% of 1, and the spreads over subsets are an order of magnitude smaller than the effects. The covariance *trace* is non-monotonic ($0.84 \to 1.02 \to 1.00 \to 1.21$): scale statistics miss the collapse that spectrum shape tracks. The covariance diagnostic and Vendi are functionally equivalent here; ours costs a $d \times d$ eigendecomposition instead of an $n \times n$ kernel spectrum and is reference-normalized by construction. One caveat: measured collapse is a *lower bound*, since low temperatures lose the most degenerate samples before embedding (26% of requested samples at $T{=}0.5$ are lost to repetition loops).

Real comparisons. We ran ten un-recentered model pairs through the full battery (Table 9, Appendix A.12), one of them a model-versus-model comparison of the kind rankings are built from. Configuration choice mattered in one: on the Aria pair the fixed window bandwidth rejects in $0/20$ runs where both defaults and MMDAgg reject in $20/20$, one more reason to aggregate rather than fix a bandwidth. The other nine differ enough in location to be detected by every configuration, including a caption-conditioned text-to-MIDI model sampled on held-out captions of its own training corpus. Deleting the location component changes that. Recentering the GPT-2-class model against its training corpus leaves a difference on which the cubic kernel rejects in $0/50$ runs and the median in $25/50$, while the window bandwidth and MMDAgg detect it in $50/50$; the same holds for a diffusion model against CIFAR-10 and a text-to-audio model against its reference corpus in PANNs, where the cubic kernel is the only configuration at zero at every sample size tested (§5). Recentered splits of either side stay at level, so the operation creates nothing by itself. Recentering is a test of whether such differences exist rather than a recommended practice, and the dissociation is an observed case rather than a rule: the FolkRNN and temperature pairs remain visible to everything after recentering, their shape differences being scale-heavy.

The dissociation at practical sample sizes. The recentered GPT-2 pair also answers how much data the default needs to see the one real shape difference we found. Testing it at $n \in \{195, 390, 780, 1560\}$ per side, the median heuristic rejects in $0/20$ and $0/20$ at the two smallest sizes (Wilson 95% upper bound 0.16), reaches about one half at $n{=}780$ ($25/50$ and $28/50$ in two runs of that cell) and saturates by $n{=}1560$, while the window bandwidth and MMDAgg stay at 19–$20/20$ throughout; we did not extend the curve past 1560, where power has saturated. The sample-size cost measured on synthetic ground truth in §4 therefore reappears, with the same magnitude, on a real model-versus-corpus difference.

A reporting protocol. For practitioners we recommend, and demonstrate on the temperature sweep: (i) MMDAgg at package defaults for the difference question; (ii) location distance $\|\Delta\hat{\mu}\|$ and, where n exceeds d, the Bures term on shrinkage covariances, reported *separately*, never summed, for magnitude, with the spectrum ratios of (iii) standing in for dispersion at $n < d$; (iii) the reference-normalized spectrum ratios of Table 2 for coverage; (iv) an NN hit-rate test against independent splits for copying. Each component has its own calibration; no single scalar respecting this division exists yet.

Aggregation costs about $16\times$ a single bandwidth in wall-clock time (Appendix A.7). What the protocol does not supply is a single ranking number, and a practitioner who needs a leaderboard column must choose which question that column answers, because different configurations rank the same two models in opposite orders, and without ground truth the data do not say which order is right (Appendix A.13).

9 THE FRÉCHET DISTANCE, RUN AS A TEST

The kernel metrics we examine were introduced as improvements on the Fréchet distance, so we run the Fréchet statistic through the same permutation protocol, battery, and spaces, with two covariance

estimators and every numerically failed matrix square root logged, a comparison we have not found in the literature.

The Fréchet test detects the difference the defaults miss. On the recentered real pair, the comparison where the cubic kernel rejects in $0/20$ runs at every sample size up to $10{,}000$ (§5), the Fréchet test rejects in $20/20$ under both estimators in CLaMP 2, VGGish, PANNs and Inception alike, at empirical level 0–$2/20$. The two 2048-dimensional spaces are run at $n{=}390$, where the plug-in covariance is singular by construction, and there the Fréchet test loses power: against the weak synthetic distortions its power falls to 0.65–0.80 in Inception and 0.25–0.50 in PANNs, where the lower-dimensional spaces give 0.85–1.00. It nonetheless detects the real pair in $20/20$, that difference being larger than the constructed ones by two to three orders of magnitude in reported units (Appendix A.8). No square root failed or was clipped in $196{,}520$ evaluations counted by the harness. The Fréchet test also misses the mean-matched mode drop ($9/20$ in CLaMP 2, $1/20$ in VGGish, $4/20$ in Inception, $2/20$ in PANNs). Appendix A.10 tabulates the battery.

10 Discussion, limitations, and conclusion

Retractions. Two intermediate conclusions of this project, both derived from split-based z-scores, failed under permutation power: (i) "raw audio embeddings admit no shape-sensitive bandwidth" (permutation power at $0.05\times$ median, rotation at $s{=}0.1$: $16/20$ in VGGish and $19/20$ in PANNs on the transition-preprocessed embeddings; both reach $20/20$ at $s{=}0.5$); (ii) "near-duplicates carry the shape-sensitive band" (power unchanged by deduplication). Both artifacts trace to the mechanism quantified in §7: near-duplicates set the variance of split-based nulls at narrow bandwidths.

A fourth space, and a boundary condition. Replicating the protocol under an encoder outside both lineages (Adversarial-MidiBERT, Zhao, 2025) confirms the thesis and locates its boundary. The spectrum diagnostic replicates in ordering. The power result inverts: at the median bandwidth the test detects the weak rotation in $20/20$ runs. That encoder collapses the corpus to an effective dimensionality of 11.2 of 768, against 46–68 for the music and audio spaces and 252–342 for vision, and the σ^{-4} coefficient grows as effective dimensionality falls (Ramdas et al., 2015). Across the six spaces the ordering is consistent: at $d_{\text{eff}}{=}11$ the default itself detects shape; at 46–68 it does not but a narrow window retains full power; at 252 and above the best single bandwidth reaches half or less (0.52 and 0.40). Manipulating d_{eff} within one space dissociates the two halves of this ordering (Appendix A.6). We report it as a measured, consistently ordered boundary condition, not a fitted law.

Limitations. The symbolic-vs-audio contrast in the *observational* pair analysis (a residual-correlation transition present in symbolic pairs, absent in audio pairs) remains an unexplained observation: our causal probes eliminated its two candidate mechanisms (duplicates, PCA preprocessing) without supplying a third. The Aria pair yields nothing after recentering: both sides share a 512-token prompt inside a $2{,}048$-token sequence, so shape sensitivity is reduced by construction and we report a null result. Perceptual validation is out of scope. The copying detector is calibrated on injected copies; the measurement on generated sets in §7 is a null result on six sets. Learned and fused kernels, classifier two-sample tests, and CMMD in CLIP space were not run: the package defaults are the object of study, and data splitting at $n{=}780$ penalizes learned kernels (CLIP geometry in Appendix A.17).

Conclusion. At the sample sizes in use, both default kernels miss trace-preserving covariance change, for different reasons: the median heuristic needs more samples than most evaluation sets provide, and the cubic kernel shows no recovery at any size we reached. The Fréchet distance, run as a test, detects the shape difference a real model produces; aggregated permutation tests remove the bandwidth choice and restore power against shape; near-duplicates must be controlled before a split-based calibration is read. We recommend the protocol of §8: an aggregated test for whether distributions differ, separately reported location and dispersion for how, and a nearest-neighbor test against independent splits for copying.

AI use statement

In this work, we used generative AI tools to implement methods (analysis code and experimental tooling) and to assist in the writing of derivations, both of which are tasks with required disclosure.

We did not use generative AI tools to generate synthetic datasets, to propose or refine hypotheses, to design the methodology or experiments, or to interpret results; the remaining required-disclosure tasks (translation, dataset cleaning, qualitative data analysis) are not applicable to this work. Additionally, we used generative AI tools for tasks with recommended disclosure: drafting and editing parts of the manuscript, creating and editing software code, formatting references, and searching for related work. We have reviewed all AI-assisted work: every reference found with AI assistance was verified against its original source before citation, all experiments were executed on the authors' infrastructure from seeded scripts, AI-assisted code was tested against synthetic cases with known answers, and the derivations of Appendix A.1 were verified numerically against closed-form expectations. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Ethics statement

This work studies evaluation methodology on public datasets (CIFAR-10, STL-10, an FMA-derived audio corpus, MidiCaps, Aria-MIDI, and the folk-rnn v2 corpus) and on generative models used under their published licenses, and involves no human subjects or personally identifiable data. We foresee no harmful applications beyond those of evaluation methodology generally; the copying-detection analysis is intended to support training-data accountability rather than to circumvent it.

Reproducibility statement

Every reported number maps to a script with stated seeds and permutation counts, submitted as supplementary code together with per-cell result CSVs. Seeds derive from a stable checksum of the cell identifier; an earlier version of the harness derived them from Python's built-in string hash, which is randomized per process, so cells re-run in a separate process drew different samples. We found this by computing one cell twice, the least-displacing mode drop under the Fréchet statistic in CLaMP 2, and comparing $3/20$ against $6/20$; a different cell from the deduplication comparison of §7, which happens to give the same pair of counts on two corpora); saturated cells were unaffected and borderline cells differed by one draw. The released code is the corrected version, and the released CSVs record the seed actually used for each draw. Sample survival across the temperature sweep is reported in Appendix A.4. The symbolic corpus is the public folk-rnn v2 training set of Sturm et al. (2016), available in the folk-rnn repository (`https://github.com/IraKorshunova/folk-rnn`); we converted it to one ABC file per tune, and the conversion script is released with the code. Section 3 specifies the distortion battery and testing protocol; Appendix A.3 gives exact constructions, seed derivations, and cross-implementation caveats; Appendices A.2 and A.5 through A.17 tabulate all cells behind the figures and claims; Appendix A.1 contains the complete derivation and its numerical verification. MMDAgg is used at the package defaults of `mmdagg` 1.0.0 (PyPI; the version is pinned in the supplementary environment file) without per-comparison tuning; all analysis runs on CPU except embedding extraction.

References

Keshav Bhandari, Abhinaba Roy, Kyra Wang, Geeta Puri, Simon Colton, and Dorien Herremans. Text2midi: Generating symbolic music from captions. In *AAAI Conference on Artificial Intelligence*, 2025.

Felix Biggs, Antonin Schrab, and Arthur Gretton. MMD-FUSE: Learning and combining kernels for two-sample testing without data splitting. In *Advances in Neural Information Processing Systems*, 2023.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *Proc. Int. Conf. on Learning Representations*, 2018.

Sebastian Bischoff, Alana Darcher, Michael Deistler, Richard Gao, Franziska Gerken, Manuel Gloeckler, Lisa Haxel, Jaivardhan Kapoor, Janne K. Lappalainen, Jakob H. Macke, Guy Moss, Matthijs Pals, Felix Pei, Rachel Rapp, A. Erdem Sağtekin, Cornelius Schröder, Auguste Schulz, Zinovia Stefanidi, Shoji Toyota, Linda Ulmer, and Julius Vetter. A practical guide to sample-based statistical distances for evaluating generative models in science. *Transactions on Machine Learning Research*, 2024.

Arijit Biswas and Lars Villemoes. Towards evaluating generative audio: Insights from neural audio codec embedding distances. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2026. arXiv:2509.18823.

Louis Bradshaw, Honglu Fan, Alexander Spangher, Stella Biderman, and Simon Colton. Scaling self-supervised representation learning for symbolic piano performance. In *Proc. 26th Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2025. arXiv:2506.23869.

Hao Chen. What FID hides: Detecting, ranking, and diagnosing deviations in generative evaluation. *arXiv preprint arXiv:2608.24881*, 2026.

Yoonjin Chung, Pilsun Eu, Junwon Lee, Keunwoo Choi, Juhan Nam, and Ben Sangbae Chon. KAD: No more FAD! an effective and efficient evaluation metric for audio generation. In *ICML 2025 Workshop on Machine Learning for Audio (AI Heard That!)*, 2025. arXiv:2502.15602.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems*, 2023.

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. FMA: A dataset for music analysis. In *Proc. 18th Int. Society for Music Information Retrieval Conf. (ISMIR)*, pp. 316–323, 2017.

Zach Evans, Julian D. Parker, Matthew Rice, C. J. Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio 3. *arXiv preprint arXiv:2605.17991*, 2026.

Dan Friedman and Adji Bousso Dieng. The Vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023.

Matthias Gelbrich. On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. CNN architectures for large-scale audio classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

Peter Matthew Jacobs, Lekha Patel, Anirban Bhattacharya, and Debdeep Pati. Robust simulation based inference through robust optimal transport. *arXiv preprint arXiv:2605.18741*, 2026.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking FID: Towards a better evaluation metric for image generation. In *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 9307–9315, 2024.

Wonwoo Jeong. Optimal transport audio distance with learned Riemannian ground metrics. *arXiv preprint arXiv:2605.05554*, 2026.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Proc. Interspeech*, pp. 2350–2354, 2019.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, 2019.

Casey Meehan, Kamalika Chaudhuri, and Sanjoy Dasgupta. A non-parametric test to detect data-copying in generative models. In *International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3546–3556, 2020.

Jan Melechovsky, Abhinaba Roy, and Dorien Herremans. MidiCaps: A large-scale MIDI dataset with text captions. In *Proc. Int. Society for Music Information Retrieval Conf.*, pp. 858–865, 2024.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *Proc. Int. Conf. on Machine Learning*, 2020.

Leslie O'Bray, Max Horn, Bastian Rieck, and Karsten Borgwardt. Evaluation metrics for graph generative models: Problems, pitfalls, and practical solutions. In *International Conference on Learning Representations*, 2022.

Aaditya Ramdas, Sashank J. Reddi, Barnabás Póczos, Aarti Singh, and Larry Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proc. AAAI Conf. on Artificial Intelligence*, 2015.

Jan Retkowski, Jakub Stępniak, and Mateusz Modrzejewski. Fréchet music distance: A metric for generative symbolic music evaluation. *arXiv preprint arXiv:2412.07948*, 2024.

Antonin Schrab, Ilmun Kim, Mélisande Albert, Béatrice Laurent, Benjamin Guedj, and Arthur Gretton. MMD aggregated two-sample test. *Journal of Machine Learning Research*, 24(194):1–81, 2023.

Paweł Skierś, Maksymilian Łazarski, Michał Kopeć, and Mateusz Modrzejewski. Symbotunes: Unified hub for symbolic music generative models. *arXiv preprint arXiv:2410.20515*, 2024. Late-breaking demo, ISMIR 2024.

George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, J. Eric T. Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 3732–3784, 2023.

Bob L. Sturm, João Felipe Santos, Oded Ben-Tal, and Iryna Korshunova. Music transcription modelling and composition using deep learning. In *Proceedings of the 1st Conference on Computer Simulation of Musical Creativity (CSMC)*, 2016. arXiv:1604.08723.

Danica J. Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In *International Conference on Learning Representations*, 2017.

Shangda Wu, Zhancheng Guo, Ruibin Yuan, Junyan Jiang, Seungheon Doh, Gus Xia, Juhan Nam, Xiaobing Li, Feng Yu, and Maosong Sun. CLaMP 3: Universal music information retrieval across unaligned modalities and unseen languages. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 2605–2625, 2025a.

Shangda Wu, Yashan Wang, Ruibin Yuan, Zhancheng Guo, Xu Tan, Ge Zhang, Monan Zhou, Jing Chen, Xuefeng Mu, Yuejie Gao, Yuanliang Dong, Jiafeng Liu, Xiaobing Li, Feng Yu, and Maosong Sun. CLaMP 2: Multimodal music information retrieval across 101 languages using large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 435–451, 2025b.

Qiantong Xu, Gao Huang, Yang Yuan, Chuan Guo, Yu Sun, Felix Wu, and Kilian Weinberger. An empirical study on evaluation metrics of generative adversarial networks. *arXiv preprint arXiv:1806.07755*, 2018.

Zijian Zhao. Let network decide what to learn: Symbolic music understanding model based on large-scale adversarial pre-training. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pp. 2128–2132, 2025.

A APPENDIX

A.1 Wide-bandwidth degeneracy of the Gaussian MMD.

Proposition 1. *Let P, Q on $\mathbb{R}^d$ have means μ_P, μ_Q, covariances Σ_P, Σ_Q, and finite fourth moments, and let $k_\sigma(x, y) = \exp(-\|x - y\|^2/(2\sigma^2))$. Then as $\sigma^2 \to \infty$,*

$$\mathrm{MMD}^2_{k_\sigma}(P, Q) \;=\; \frac{\|\mu_P - \mu_Q\|^2}{\sigma^2} \;+\; O(\sigma^{-4}),$$

and covariance differences enter only at order σ^{-4}. If moreover P, Q are Gaussian with $\mu_P = \mu_Q$, the leading term is

$$\mathrm{MMD}^2_{k_\sigma}(P, Q) \;=\; \frac{(\operatorname{tr}\Sigma_P - \operatorname{tr}\Sigma_Q)^2 \;+\; 2\,\|\Sigma_P - \Sigma_Q\|_F^2}{4\sigma^4} \;+\; O(\sigma^{-6}).$$

Derivation. Write $\mathrm{MMD}^2 = \mathbb{E}k(x, x') + \mathbb{E}k(y, y') - 2\mathbb{E}k(x, y)$ and expand $k_\sigma = 1 - \frac{\|x-y\|^2}{2\sigma^2} + \frac{\|x-y\|^4}{8\sigma^4} - \cdots$. The order-$\sigma^0$ terms cancel. At order σ^{-2}, using $\mathbb{E}\|x - x'\|^2 = 2\operatorname{tr}\Sigma_P$, $\mathbb{E}\|y - y'\|^2 = 2\operatorname{tr}\Sigma_Q$, and $\mathbb{E}\|x - y\|^2 = \|\Delta\mu\|^2 + \operatorname{tr}\Sigma_P + \operatorname{tr}\Sigma_Q$, the trace terms cancel exactly and the coefficient is $\|\Delta\mu\|^2$. For the Gaussian equal-means case, $x - x' \sim \mathcal{N}(0, 2\Sigma_P)$, $x - y \sim \mathcal{N}(0, \Sigma_P + \Sigma_Q)$, and the quartic identity $\mathbb{E}(z^\top z)^2 = (\operatorname{tr} S)^2 + 2\operatorname{tr} S^2$ for $z \sim \mathcal{N}(0, S)$ gives, after collecting the three terms at order σ^{-4}, the stated coefficient $2(\Delta \operatorname{tr}\Sigma)^2 + 4\|\Delta\Sigma\|_F^2$ over $8\sigma^4$. Empirically, the ratio of the kernel statistic to the first-order prediction $\|\Delta\mu\|^2/\sigma^2$ is constant across fifteen symbolic corpus pairs at the median bandwidth (0.64–0.67, s.d. 0.010). The univariate wide-bandwidth limit is stated as a lemma by Jacobs et al. (2026), the derivation above being its multivariate counterpart; the bandwidth-dependence of kernel test power is established broadly by Ramdas et al. (2015). Both orders of the proposition were verified against closed-form Gaussian expectations (agreement to 6×10^{-5} at $\sigma^2{=}10^6$). The cubic kernel admits the analogous reading: expanding $(1 + u)^3$ with $u = x^\top y/d$, the linear term contributes $\|\Delta\mu\|^2$ up to scale while higher moments enter through u^2, u^3; at the measured $\mathbb{E}|u| = 0.23$ on the MidiCaps genre subsets, the same fifteen pairs on which the dispersion weight is measured, the predicted quadratic contribution is $\sim$18% against a measured dispersion weight of 12.1%, so the prediction is correct in form and weak in magnitude.

A.2 Full power tables. Table 3 reports per-cell rejection counts behind Figure 1 ($n{=}780$ per side, 20 draws, 500 permutations, $\alpha{=}0.05$); Table 4 the fixed-kernel arms and aggregation; Table 5 the synthetic sample-size analysis.

Table 3: Cells give rejections out of 20 draws by Gaussian bandwidth, in multiples of the median d^2; "med" is the classical median heuristic $\sigma{=}\mathrm{med}(d)$. C2 is CLaMP 2, VG is VGGish, and PN is PANNs. The main text quotes the med column of this table.

Distortion	Space	.01	.02	.05	.1	.2	.5	1	2	med
rotation $s{=}.1$	C2	1	6	20	20	15	5	2	1	2
	VG	4	17	20	18	11	4	4	3	4
	PN	7	20	20	20	10	4	1	0	1
mean shift $s{=}.1$	C2	0	4	20	20	20	20	20	20	20
	VG	4	8	20	20	20	20	20	20	20
	PN	4	16	20	20	20	20	20	20	20
control ($s{=}0$)	C2	1	2	0	0	0	0	0	0	0
	VG	1	0	1	1	0	0	0	0	0
	PN	1	1	0	1	1	1	1	1	1

Pooled empirical level across all fixed-bandwidth Gaussian control cells (the three grids of Table 3, the Inception grid, and the STL-10 arms) is $35/760$ (4.6%); the per-cell maximum is $3/20$, reached only in the Inception grid, the source of the Figure 1 caption bound.

A.3 Distortion battery conventions and toolkit replication. Perturbation seeds derive deterministically from (distortion index, strength index, draw index, global seed); the Cayley rotation uses

Table 4: Fixed-kernel arms and aggregation at weak strength s=0.1 unless noted, with 20 draws per cell and 50 where two counts appear or where marked. Median-baseline rotation cells with two counts report seeds 0;1. †: a second run of this cell; the real-comparisons battery gives $25/50$ for the same configuration (§8).

	CLaMP 2	VGGish	PANNs
cubic (KID): rotation	1/20	2/20	0/20
cubic: contraction	4/20	2/20	2/20
cubic: mean shift	20/20	20/20	20/20
cubic: control	0/20	0/20	1/20
MMDAgg (defaults): rotation	18/20	50/50	20/20
MMDAgg, Gaussian-only: rotation	18/20	42/50	20/20
MMDAgg: rotation s=0.5	20/20	20/20	20/20
MMDAgg: control (pooled, all five spaces)		$13/260$	
median baseline: rotation	2/20	3/20; 4/20	1/20; 2/20
median baseline: control	0/20	0/20	0/20
Inception (vision): MMDAgg / Gauss-only		$20/20$ / $5/20$	
Inception: median / cubic: rotation		$3/20$ / $3/20$	
Inception: best single bandwidth (0.05)		$26/50$	
Inception: MMDAgg rotation s=0.5		$20/20$	
recentered GPT-2 pair, median: n=195/390		$0/20$ / $0/20$	
n=780/1560		$28/50^{\dagger}$ / $20/20$	
window and MMDAgg, all n		19–$20/20$	
STL-10: MMDAgg / Gauss-only		$20/20$ / $7/20$	
STL-10: median / cubic: rotation		$2/20$ / $2/20$	
STL-10: best single bandwidth (0.05)		$8/20$ (mean shift $17/20$)	

Table 5: Synthetic power on 768d Gaussians with equal means and 25% of dimensions rescaled by $1+\gamma$, with 50 repetitions per cell and 20 for MMDAgg. The empirical level of the "1" arm at γ=0 is 0.06. Headers are multiples of the median *squared* distance, so column "1" (σ^2=med d^2) is the synthetic proxy for the classical heuristic, which the real-data tables report separately as σ=med(d).

γ	n	.02	.05	.1	.5	1	2	MMDAgg
0.05	500	1.00	1.00	.92	.24	.10	.06	.95
	2000	1.00	1.00	1.00	.96	.46	.24	1.00
	6000	1.00	1.00	1.00	1.00	1.00	.72	1.00
0.15	500	1.00	1.00	1.00	1.00	.96	.66	1.00
	2000	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	6000	1.00	1.00	1.00	1.00	1.00	1.00	1.00

$M = (I+S)^{-1}(I-S)$ with $S = s\,(A - A^{\top})/(2\sqrt{d})$, $A_{ij} \sim \mathcal{N}(0,1)$, so strengths are not numerically comparable across implementations that scale the skew part differently (the evaluation toolkit's convention is $\sim 2\times$ stronger at equal s; all cross-tool comparisons in this paper are of qualitative profiles, never of raw values). Replicating the copying experiment with the toolkit's own estimator on the four symbolic and audio embedding sets confirms the profile of Table 1: at 5% injected copies, the narrow-bandwidth statistic responds at $|z| = 3.5$–4.7 against 0.4–0.8 for the median bandwidth and ≤ 1.5 for the Fréchet distance, with a per-bandwidth standardized multi-bandwidth variant the most sensitive at $|z| = 8.5$–13.6. On a corpus stored in genre order, taking the first half as the reference makes every configuration reject in $20/20$ runs, where random splits stay at 2–5 of 50. Full grids over all six distortion types (including mode drop) at three strengths, for the original and deduplicated symbolic corpus and both audio spaces, are released as CSV alongside the code.

In the recentered arms each side is centered on its own mean once, before the permutation loop. The observed split therefore has zero mean difference while the permuted splits retain a small random one, which enters the kernel statistic and raises the null distribution. The observed value sits low in

Table 6: Mode drop battery with 20 draws, 50 where a count fell between 3 and 17; half of the largest of k=10 k-means clusters is removed. "clean" replaces deleted points from a disjoint reserve of the corpus; "resampled" replaces them by drawing from the remainder, which injects duplicates at the measured rate f; "duplicates only" applies that rate with no mode removed. "mean-matched" re-subtracts the induced mean displacement.

Space	Variant	median	cubic	window	MMDAgg
CLaMP 2	control	1/20	1/20	2/20	1/20
(f=0.10)	clean, mean-matched	4/50	2/50	34/50	24/50
	resampled, mean-matched	15/50	4/50	50/50	50/50
	duplicates only	6/50	4/50	50/50	50/50
VGGish	control	1/20	1/20	1/20	1/20
(f=0.084)	clean, mean-matched	1/20	0/20	2/20	0/20
	resampled, mean-matched	7/50	2/50	50/50	50/50
	duplicates only	0/20	1/20	20/20	20/20
PANNs	control	1/20	0/20	2/20	1/20
(f=0.094)					
	clean, mean-matched	4/50	3/50	15/50	50/50
	resampled, mean-matched	1/20	1/20	20/20	20/20
	duplicates only	1/20	1/20	20/20	20/20

Table 7: Low-rank alternatives at γ=0.3, with 50 draws where a count fell between 3 and 17 and 20 otherwise; all constructions leave the sample mean unchanged ($\|\Delta\hat{\mu}\|/r = 0.000$). Rank-2 and rank-5 are rotations confined to the leading principal components; inflation and deflation are rank-one scalings of the first component; "random direction" applies the same scaling along a random direction; contamination displaces 2% of points.

Space	Alternative	median	cubic	window	MMDAgg
CLaMP 2	control	1/20	0/20	2/20	1/20
	rank-2 rotation	7/50	2/50	39/50	36/50
	rank-5 rotation	16/50	3/50	50/50	50/50
	rank-1 inflation	50/50	8/50	50/50	50/50
	rank-1 deflation	50/50	5/50	50/50	50/50
	random direction	2/20	1/20	1/20	1/20
	contamination 2%	2/20	0/20	0/20	0/20
PANNs	rank-2 rotation	4/50	0/50	30/50	44/50
	rank-5 rotation	17/50	1/50	50/50	50/50
	rank-1 inflation	40/50	2/50	43/50	50/50
	rank-1 deflation	36/50	5/50	50/50	50/50
Inception	rank-2 rotation	5/50	2/50	15/50	21/50
(CIFAR-10)	rank-5 rotation	8/50	3/50	44/50	50/50
	rank-1 inflation	44/50	0/50	35/50	50/50
	rank-1 deflation	33/50	5/50	36/50	50/50

that null and the test is conservative in this arm: rejections there are lower bounds on the power a protocol that recentered inside the loop would have.

A.4 Sample-survival asymmetry in the temperature sweep. We requested 10,000 pieces per temperature; the surviving counts are 7,379 (T=0.5), 10,205 (0.7), 10,139 (0.85), 10,000 (1.0, pre-existing), 9,357 (1.2; 9,353 after feature extraction). The loss mechanisms differ by direction: at low temperature the model enters repetition loops that never emit the end token (cut by a 1,500-token guard set below the model's ∼1,955-token context) and produces sequences too degenerate for the notation converter; at high temperature losses come from format drift. Low-temperature losses remove the most collapsed samples before embedding, which is why the ratios of Table 2 are lower bounds on collapse.

A.5 Observational corpus-pair analysis. Complementing the interventional battery, we measure which term of the Fréchet decomposition $\|\Delta\mu\|^2 + \mathrm{tr}(\Sigma_r + \Sigma_t - 2(\Sigma_r\Sigma_t)^{1/2})$ the kernel statistic follows on *real* corpus pairs: fifteen pairs of genre subsets of a symbolic corpus (Melechovsky et al., 2024), five independent subsamples of 6,000 pieces, fourteen bandwidths, and two encoder versions (CLaMP 2/3, Wu et al., 2025b;a). Because the location and dispersion terms are collinear across pairs (Pearson 0.84/0.86), the regression-based dispersion weight is reported with bootstrap intervals over pairs and alongside an apportionment-free statistic: the rank correlation between a pair's location share and its kernel-statistic residual from a linear fit on the Fréchet distance. Three statements survive the uncertainty analysis: below σ^2=18 the dispersion weight is 58–66% across both encoders; above σ^2=48 it is 0.5–8.8% (at the median bandwidth $\sigma^2\approx$600: 1.9% and 1.6%; cubic kernel: 12.1%); and the residual correlation changes sign at σ^2=32 and 30 for the two encoders, the transition whose location coincides with the power window of Figure 1. No claim about a peak survives the bootstrap (interval widths up to 84 percentage points, driven by which corpora are compared rather than by sampling within them), which is why the claim rests on the sign change rather than on a maximum. The same sweep yields the ranking-reorder fact quoted in §4: the pair second most distant at σ^2=30 is fifth at the median.

A.6 Within-space effective-dimension manipulation. Table 8 reports rotation power (s=0.1, 20 draws) on PCA truncations and random orthonormal projections of Inception/CIFAR-10 embeddings.

Table 8: Rotation power along the within-space d_{eff} ladder on Inception/CIFAR-10 at s=0.1, with 20 draws unless marked. Columns .02–.1 give the Gaussian bandwidth as a multiple of the median d^2; "shift .05" is the mean-shift arm at the 0.05× bandwidth; "ctrl" is the maximum control-cell count across the row's arms.

proj.	d_{eff}	.02	.05	.1	median	MMDAgg	shift .05	ctrl
PCA-16	12.8	18	16	11	4	18	19	≤ 3
PCA-32	23.0	18	17	14	3	15	20	≤ 3
PCA-64	40.2	13	16	13	3	12	20	≤ 3
PCA-128	69.0	7	16	11	3	9	20	≤ 3
PCA-256	113.1	4	13	12	3	10	20	≤ 3
PCA-512	167.9	4	13	10	3	15	20	≤ 3
RP-64	46.2	3	9	10	3	6	20	≤ 3
RP-256	125.9	2	9	9	2	6	20	≤ 4
full	266.3	3	26/50	9	3	20	20	≤ 3

The full-space d_{eff}=266.3 here is the 50,000-sample estimate; the 251.8 quoted in the main text uses the protocol's 5,000-sample subsets for comparability across spaces. Vocabulary coverage of the fourth encoder of §10 was not a confound (rejects: 6/23,636 and 1/10,000 files). PANNs repeats the pattern (PCA-32/128: window 19–20/20, median 1/20). Aggregation is non-monotone along the ladder ($18 \to 9 \to 20$): the aggregate tracks the best member of its collection, whose Gaussian bandwidths weaken mid-ladder before the Laplace family, ineffective in the truncated subspaces but decisive in the full space (§6), takes over. A sweep over kernel families locates the mechanism, and rules out two candidates on the way. Holding the functional form fixed and changing only the norm separates it: at the median scale in the full Inception space, $\exp(-\|x-y\|_1/s)$ rejects the rotation in 20/20 runs where $\exp(-\|x-y\|_2/s)$ reaches 5/20 and the Gaussian 3/20; on PCA truncations of the same embeddings the advantage disappears (2/20 at k=64, 5/20 at k=256). Smoothness contributes little and in the direction opposite to the intuition that a peaked kernel helps: Matérn kernels with $\nu = 1/2, 3/2, 5/2$ at the narrow scale give 7, 9 and 11 of 20. Tail weight also runs the wrong way: rational quadratic kernels with heavy tails (α=0.2) reach 4–5/20 against 10–16/20 for light tails (α=5). A first-order profile of kernel values, which we report for completeness, does not track the power gap either. The Laplace family's power here comes from ℓ_1 geometry at high ambient dimension. The full-row 0.02 cell is taken from the main vision grid (20 draws). The mean-shift arm at 0.05× stays at 19–20/20 along the ladder and every control cell stays within 0–4/20, so rising rotation power at low d_{eff} reflects power rather than miscalibration; the 0.02× mean-shift arm declines from 17/20 to 5/20 as k grows (per-cell counts in the released CSVs), repeating the two-sidedness of the window along the ladder.

A.7 Cost. The table reports seconds per test on one CPU with a fixed thread count, embedding extraction excluded, on Inception embeddings (d=2048) with 500 permutations.

n per side	median	cubic	window	MMDAgg
780	0.8	0.6	0.6	10
5,000	23	20	23	359
10,000	90	76	89	1414

A.8 Effect sizes in practitioner units. The table reports the displacement of the reported metric value, in standard deviations of a noise floor estimated from 20 independent splits of the reference (n=780 per side, 5 draws per distortion). KID is computed with `torchmetrics` and KAD with the authors' `kadtk`, both at package defaults.

	CLaMP 2		VGGish		PANNs		Inception	
Distortion	KID	KAD	KID	KAD	KID	KAD	KID	KAD
rotation, equal transport	−0.2	−0.2	0.1	0.3	−0.3	0.6	0.8	0.6
mean shift, s=0.1	8.8	8.6	14.3	17.2	17.5	18.9	24.8	28.8
mean-matched mode drop	0.4	0.3	−0.6	−0.5	−0.1	0.2	−0.6	−0.4
isotropic contraction	0.5	6.1	−0.3	11.1	0.4	12.8	1.1	20.3
real model pair	48	49	128	151	588	744	71	78

The real-pair row sets the scale: an equal-transport shape difference is two to three orders of magnitude smaller in reported units than the difference between a model and its data, while a mean shift of equal or smaller transport magnitude is within one to two orders.

A.9 Real pairs through the full battery. Three pairs were selected by a geometry screen over twelve candidates as the three with the lowest location share of the Fréchet decomposition (0.198, 0.244, 0.253); the screen is reported in the released CSVs. The table gives rejections out of 20 draws at n=780 per side, with the vision pair also run at larger samples.

Pair	Arm	median	cubic	window	MMDAgg
DDPM vs. CIFAR-10	plain	20/20	20/20	20/20	20/20
n=780	recentered	0/20	0/20	20/20	20/20
	split controls	3/50	3/50	4/50	5/50
n=2500	recentered	2/20	0/20	20/20	20/20
n=5000	recentered	20/20	0/20	20/20	20/20
Stable Audio vs. FMA	plain	20/20	20/20	20/20	20/20
(PANNs)	recentered	20/20	0/20	20/20	20/20
	split controls	2/20	1/20	0/20	1/20
Aria vs. its own	plain	20/20	20/20	0/20	20/20
continuations	recentered	0/20	0/20	0/20	0/20
	split controls	1/20	1/20	1/20	1/20

The cubic kernel is the only configuration at zero at every sample size tested; for the vision pair the median also does not reject at n=780 and reaches full power at roughly n=5,000 (§5). The Aria pair yields nothing after recentering under any configuration, for the reason given in the limitations: both sides share a 512-token prompt within a 2,048-token sequence.

A.10 The Fréchet distance through the battery. We run the Fréchet distance as a permutation test. At $d \leq 768$ we use n=780, 500 permutations, and both estimators, which agree in every cell. At d=2048 we use n=390 and 100 permutations; the table reports the plug-in estimator, and where the shrinkage estimator also completed it agrees with the plug-in column to within three draws (Inception: rotation 14/20, mean shift 14/20, contraction 16/20, mean-matched drop 5/20; PANNs: radial truncation 20/20, intra-mode collapse 20/20). On the recentered real pair both estimators give 20/20 in Inception and in PANNs.

	constructed, s=0.1				coverage		real
Space	control	rotation	mean shift	contraction	mean-matched	collapse	recentered
CLaMP 2 (d=768)	0–1/20	20/20	20/20	20/20	9/20	20/20	20/20
VGGish (d=128)	0/20	20/20	18/20	17/20	1/20	20/20	20/20
Inception (d=2048)	2/20	16/20	13/20	13/20	4/20	20/20	20/20
PANNs (d=2048)	0/20	10/20	5/20	5/20	2/20	20/20	20/20

Least-displacing mode drop is 1–4/20 in every space, and radial truncation is $20/20$ except in Inception, where it is $12/20$. The pattern across rows is a loss of power against the weak constructed distortions as n/d falls, from 1.0 at d=128 to 0.25–0.50 at d=2048 with n=390, which is the expected cost of estimating a 2048×2048 covariance from 390 samples. Detection of the coverage collapse and of the real pair does not degrade; both are larger differences than the constructed ones.

A.11 Coverage loss that is mean-quiet by construction.

The mean-matched mode drop of §4 removes the location component by subtracting it, which is an intervention on the statistic rather than on the model. We therefore built coverage losses that leave the sample mean untouched *by construction* and ran the same battery on all five spaces. Two constructions matter. An intra-mode collapse shrinks every point toward its own cluster centroid, preserving the overall mean to numerical zero while destroying within-mode spread. A least-displacing mode drop deletes half of whichever mode moves the mean least, giving $\|\Delta\hat{\mu}\|/r \leq 0.016$ against 0.036–0.055 for the largest-mode drop.

Neither reproduces the dissociation, and they fail in opposite directions. The collapse is detected by the median heuristic in 19–$20/20$ runs and by the window and MMDAgg in $20/20$ in every space, but by KID's cubic kernel in 0–$2/20$ at strength 0.10 and 7–$25/50$ at 0.25; it is a mean-preserving coverage failure that the cubic kernel does not see and the median does. The least-displacing drop is detected by nothing: 0–$7/50$ for all four configurations in all five spaces, MMDAgg included. Empirical level on undistorted controls in the same runs is 0–$5/50$.

Two consequences follow. The claim that narrow bandwidths detect coverage loss that the defaults miss holds for the subtraction-based construction and does not generalize to coverage losses that are mean-quiet by construction; we scope it accordingly. And the cubic kernel's failure is the finding that survives across constructions: it is the one configuration that misses a coverage failure which every other configuration detects.

A.12 Real comparisons through the full battery. Table 9 lists every real pair through the full battery; per-pair geometry is in Appendix A.14.

A.13 Ranking reversal between two audio generators. The table gives wins out of 20 subsamples at n=780 per side against the FMA reference; the model with the smaller statistic wins.

	median	cubic	window 0.05	FMD	location	dispersion
PANNs: MusicGen	0	0	10	0*	0*	0*
PANNs: Stable Audio	20	20	10	20*	20*	20*
VGGish: MusicGen	1	2	20	7	3	8
VGGish: Stable Audio	19	18	0	13	17	12

* In PANNs (d=2048) the sample covariances at n=780 are singular, so the Fréchet-based columns are numerically unreliable there as raw scalars, without the permutation calibration of §9; the reversal rests on the kernel columns, which do not invert a covariance.

A.14 Geometry of the real pairs. The location shares of the Fréchet decomposition for eight of the pairs of Table 9 are: FolkRNN 0.375, GPT-2 0.417, Aria 0.253, DDPM vs. CIFAR-10 0.278, GPT-2 vs. FolkRNN 0.437, text2midi 0.667, T=0.85 0.728, T=0.7 0.737. The dissociating pair is not the lowest-share one: the FolkRNN and temperature comparisons remain visible after recentering because their covariance differences are trace-heavy (Table 2 and the effective-dimension ratios of §10), which every configuration detects, while the residual difference of the GPT-2 pair is rotation-like. Per-pair Fréchet terms, including the audio and model-versus-model pairs, accompany the released CSVs. The shares quoted here use the full embedding sets; the screen of Appendix A.9 uses a 3,000-sample subsample per side, which is why its DDPM value differs. Known-case validation

Table 9: Real comparisons through the full battery at n=780 per side with 20 draws; cells with mid-range counts were increased to 50 draws. "Recentered" subtracts each side's own sample mean, deleting the location component. Controls are independent splits of the corpus. text2midi (Bhandari et al., 2025) is sampled on 2,000 held-out MidiCaps captions; FolkRNN is the model of Sturm et al. (2016). Audio rows: MusicGen (Copet et al., 2023) and Stable Audio (Evans et al., 2026), 2,000 ten-second clips each against 2,000 FMA clips, both sides through one provenance pipeline. †: for the Aria row (Bradshaw et al., 2025) the plain arm is 20/20 at the median, cubic and MMDAgg but 0/20 at the window bandwidth.

	Plain	Recentered		
Pair	all four configs	median	cubic	window / MMDAgg
GPT-2 vs. training corpus	20/20	25/50	0/50	50/50
FolkRNN vs. training corpus	20/20	20/20	20/20	20/20
T=0.85 vs. T=1.0	20/20	20/20	20/20	20/20
T=0.7 vs. T=1.0	20/20	20/20	20/20	20/20
text2midi vs. MidiCaps (its training set)	20/20	20/20	20/20	20/20
GPT-2 vs. FolkRNN (model vs. model)	20/20	20/20	20/20	20/20
DDPM vs. CIFAR-10 (vision)	20/20	0/20	0/20	20/20
MusicGen vs. FMA (PANNs)	20/20	20/20	0/20	20/20
MusicGen vs. FMA (VGGish)	20/20	20/20	20/20	20/20
Stable Audio vs. FMA (PANNs)	20/20	20/20	0/20	20/20
Stable Audio vs. FMA (VGGish)	20/20	20/20	20/20	20/20
Aria vs. its own continuations	20/20†	0/20	0/20	0/20
split control	0–1/20		0/20	

of the spectrum diagnostic: FolkRNN, a model whose low diversity is documented by its authors, yields ratios 0.55 (eff), 0.63 (n_{90}), 0.59 (trace), and 0.54 (Vendi) against its training corpus in CLaMP 2, consistent in ordering with the out-of-domain-encoder ratios of §10.

A.15 Contraction against the bandwidth grid. The table gives rejections out of 20 draws at n=780 per side with 500 permutations. "matched" is the contraction whose 2-Wasserstein distance equals that of the s=0.1 mean shift, which bisection places at s=0.100 in every space, as the Gaussian expression for the two distortions predicts.

Space	Distortion	median	0.02	0.05	0.1	0.2	0.5	cubic
PANNs	control	0/20	1/20	0/20	0/20	3/20	2/20	0/20
	contraction 0.1	20/20	19/20	20/20	20/20	20/20	20/20	0/20
	contraction, matched	20/20	19/20	20/20	20/20	20/20	20/20	4/20
	mean shift 0.1	20/20	17/20	20/20	20/20	20/20	20/20	20/20
Inception (CIFAR-10)	control	0/20	1/20	2/20	2/20	0/20	0/20	1/20
	contraction 0.1	20/20	13/20	20/20	20/20	20/20	20/20	4/20
	contraction, matched	20/20	10/20	20/20	20/20	20/20	20/20	4/20
	mean shift 0.1	20/20	3/20	20/20	20/20	20/20	20/20	20/20
Inception (STL-10)	control	1/20	2/20	2/20	1/20	0/20	1/20	2/20
	contraction 0.1	20/20	7/20	20/20	20/20	20/20	20/20	5/20
	contraction, matched	20/20	7/20	20/20	20/20	20/20	20/20	6/20
	mean shift 0.1	20/20	2/20	18/20	20/20	20/20	20/20	20/20

CLaMP 2 and VGGish give the same pattern; per-cell counts are in the released CSVs. Strengths are matched by 2-Wasserstein distance throughout; no other matching criterion was tried. This table is an independent run of the contraction cells, with its own seeds; where the earlier battery of Table 4 covers the same cell the two differ by at most three draws (cubic contraction: CLaMP 2 3/20 here against 4/20 there, VGGish 3/20 against 2/20, PANNs 0/20 against 2/20). The counts quoted in the main text are those of this run, except the opening paragraph of §4, which quotes the music and audio cells of Table 4.

A.16 Precision, recall, density and coverage on the same battery. The table gives rejections out of 20 draws against split-split nulls with k=5, 200 permutations, and n=780 per side. Values in parentheses are the mean four-tuple (precision, recall, density, coverage) for CLaMP 2.

Distortion	CLaMP 2	VGGish	PANNs	Inception	STL-10
control	1/20	2/20	2/20	1/20	2/20
rotation s=0.1	17/20	14/20	16/20	7/20	1/20
mean shift s=0.1	10/20	9/20	16/20	9/20	7/20
contraction s=0.1	20/20	20/20	20/20	20/20	20/20
mode drop, mean-matched	6/20	1/20	3/20	4/20	3/20
mode drop, least-displacing	2/20	0/20	1/20	1/20	0/20
intra-mode collapse	20/20	20/20	20/20	20/20	20/20
rank-2 rotation	10/20	7/20	6/20	5/20	2/20
rank-1 inflation	20/20	11/20	13/20	11/20	6/20
real pair, recentered	20/20	20/20	20/20	14/20	n/a

The collapse leaves a signature the kernel statistics do not report: recall falls from 0.85 to 0.13 and density rises to 10.1 in Inception. The mean-matched and least-displacing mode drops leave all four quantities within 0.05 of their control values in every space.

A.17 Two properties of the embedding spaces. The left block gives the spectrum of $\Delta\Sigma$ for the real pairs, as the fraction of $\|\Delta\Sigma\|_F^2$ in the five leading eigendirections and the effective rank of its squared spectrum. The right block gives the mean absolute value of $u = x^\top y/d$, the argument of the cubic kernel. *CLIP, the embedding CMMD uses, is included for reference; no result in the paper is computed in that space.

Pair	top-5	eff. rank	Space	$\mathbb{E}\|u\|$	$\mathrm{Var}(u)$
DDPM vs. CIFAR-10	0.40	59.4	CLaMP 2	0.504	0.0037
GPT-2 vs. corpus	0.65	15.0	VGGish	1.98×10^{4}	1.16×10^{6}
FolkRNN vs. corpus	0.87	6.3	PANNs	0.040	4×10^{-5}
text2midi vs. MidiCaps	0.60	18.2	Inception (CIFAR-10)	0.144	0.0011
MusicGen vs. FMA	0.71	12.6	Inception (STL-10)	0.069	0.0005
Stable Audio vs. FMA	0.72	11.9	CLIP (CIFAR-10)*	0.404	0.0035
Aria vs. continuations	0.65	17.0			

VGGish is the outlier on the right: its features are large enough that the constant in $(u+1)^3$ is negligible, so the cubic kernel there is effectively u^3. This is the reason the KID noise floor in that space is numerically extreme (Appendix A.8) and a caution against reading its absolute values.

A.18 Power against sample size in vision. Inception embeddings of CIFAR-10 are tested against a pool of 100,000 DDPM samples. Cells give rejections out of 20 draws at $n \le 5{,}000$, 10 at n=10,000, and 5 at n=20,000. "rotation" is matched to the mean shift in 2-Wasserstein distance; "real pair" is DDPM against CIFAR-10 after recentering.

Arm	Configuration	780	2,500	5,000	10,000	20,000
rotation	median	3/20	3/20	17/20	10/10	5/5
	cubic	2/20	0/20	1/20	2/10	2/5
	window 0.05	5/20	20/20	20/20	10/10	5/5
	MMDAgg	20/20	20/20	20/20	10/10	n/a
real pair	median	0/20	2/20	20/20	10/10	5/5
	cubic	0/20	0/20	0/20	0/10	2/5
	window 0.05	20/20	20/20	20/20	10/10	5/5
	MMDAgg	20/20	20/20	20/20	10/10	5/5
control	median	0/20	2/20	0/20	1/10	n/a
	cubic	0/20	1/20	0/20	0/10	n/a
	window 0.05	1/20	1/20	0/20	1/10	n/a

Cells marked n/a were not run: the aggregated test at n=20,000 exceeded the memory budget, and the control arm was not extended past n=10,000. A mean shift of s=0.1 is detected in every configuration at every size.