ESCAPING THE BARYCENTRIC CEILING: A TRAINING-FREE DIAGNOSTIC FOR OUT-OF-DISTRIBUTION CONDITIONAL GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A conditional generative model learns to produce data at the settings it was trained on, and is then asked for a setting it has never seen. A recent line of analysis gives a discouraging answer: if the model can only blend the conditions it was trained on, there is a hard floor on how well it can ever do, set by how far the unseen target lies outside the region those conditions span. We revisit that floor and find that it rests on a hidden assumption - that the blending weights must be positive. Allow them to go negative, so that the model extrapolates along the geometry of covariance matrices rather than averaging within it, and the floor stops applying. We prove this, and we correct an error we found in our own earlier argument for it along the way; the corrected proof is short enough to be checked by machine, and we check it. But escaping the floor is not always possible, and whether it is depends on *how* the data's shape drifts as the condition changes. If the drift is a stretching of the same axes, extrapolation works; if it is a rotation of those axes, nothing we tried helps. We turn this into a diagnostic that costs two covariance estimates and no training at all, and it predicts the answer before a single gradient step. It holds up on natural image and biological data, and we report plainly where it does not: it reads rotation-like drift correctly everywhere we looked, it is confounded on cluttered colour scenes, and it says nothing useful about unordered categories, which we document rather than hide. Finally we ask whether any of this survives an actual trained network, and here the most useful result is a negative one. Handing the network the extrapolated statistics directly - the design the theory naively suggests - destroys the diagnostic and makes the model worse than a plain baseline. Building the extrapolation into the architecture as a closed-form starting point, and asking the network only for a correction, restores it. The lesson is that the escape route has to be part of the model's structure; a network will not find it on its own.

1 INTRODUCTION

A conditional flow-matching model is trained on conditions $c \in [0, 1]$ and then asked to generate at an out-of-distribution (OOD) condition $\bar{c} > 1$. How well can it possibly do - *before* we train anything?

The ceiling. The published analysis we build on (Pham, 2026) idealises the OOD output as a Wasserstein barycenter of the training conditions. The reachable set of such a predictor is the set of barycenters over the simplex - the *barycentric hull* $\mathcal{B}$ - and the best achievable OOD cost is the distance from the true target to that hull, which we write $\mathrm{Term\,I}^*$. The framework's assumption is explicit: the unseen condition is "approximated by a convex combination of seen conditions." Under that assumption $\mathrm{Term\,I}^*$ is a genuine floor.

The escape. The assumption fails exactly when the OOD condition is reachable by an *affine* combination - coefficients summing to one but allowed to go negative. For Gaussians, log-Euclidean extrapolation of $(\mu(c), \Sigma(c))$ along an SPD geodesic is precisely such a map in $\log \Sigma$-space, and it is not restricted to $\mathcal{B}$ (Figure 1). Section 3.1 makes this precise and self-contained.

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.

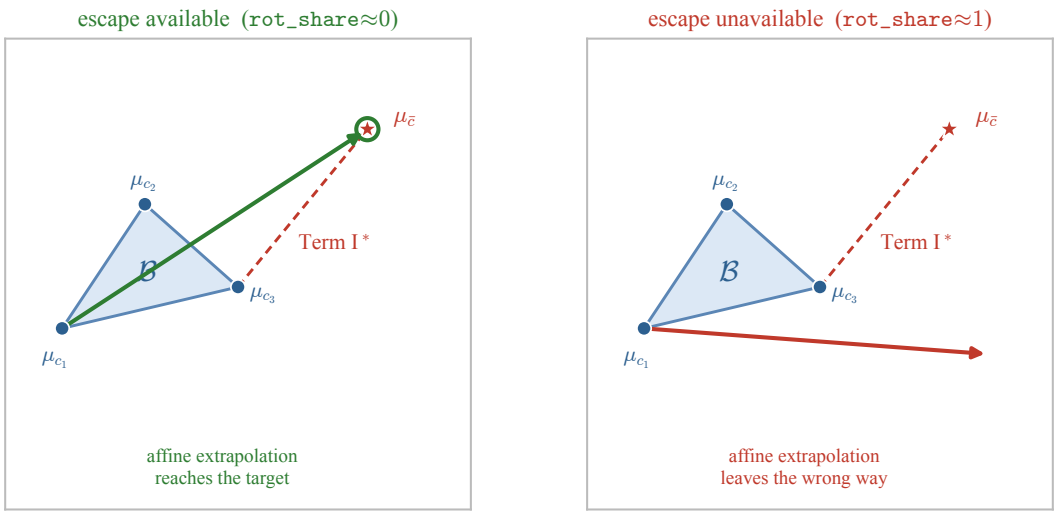

Figure 1: The ceiling and the escape. Convex (barycentric) predictors are confined to the hull $\mathcal{B}$ of the training conditions, so the target $\mu_{\bar{c}}$ is at best $\mathrm{Term\,I}^*$ away. An *affine* predictor - weights summing to one but allowed to be negative - is not confined to $\mathcal{B}$. **Left:** when the conditional covariance drifts by spectral scaling, the affine ray runs through the target and the ceiling is irrelevant. **Right:** when the drift is eigenframe rotation, the same ray leaves in the wrong direction and the escape is unavailable. Telling these two regimes apart *before training* is what `rot_share` (§3.2) does.

A correction we make to our own argument. While preparing this paper we machine-checked the proof and found that one step of our earlier argument is *false*. That step claimed the barycentric hull of a commuting family coincides with the curve segment $\{\Sigma(c) : c \in [0, 1]\}$ through a "componentwise" reparameterisation. The reparameterisation in fact yields a *vector* of per-eigenvalue parameters rather than a scalar, and the barycenter lies strictly *off* the curve whenever the log-spectral slope has unequal components (§3.1, Appendix A). The proposition itself survives: we give a shorter, correct proof that minimises the relevant quadratic directly and never needs the hull to be a curve. The corrected argument is formalised in Lean 4 and machine-checked against Mathlib (Moura & Ullrich, 2021; The mathlib Community, 2020); the refuted step is exhibited numerically. We report this because the corrected proof is both stronger and easier to verify, and because a reader deserves to know which parts of a geometric argument were checked by a machine and which were not.

The diagnostic. Whether the escape is available in practice is a property of *how* the conditional covariance drifts. Section 3.2 introduces `rot_share`, the fraction of the drift between two covariances that lives in the eigenframe rather than the spectrum, computable from two covariance estimates with no training, no target samples, and no generative model.

The reality check. Section 4 asks whether any of this survives when generation is done by a real trained network rather than by closed-form moments. The answer is nuanced and, we believe, the most useful result in the paper: the architecture that the affine-escape theory naively recommends is the one that fails.

Contributions.

1. A self-contained convex-vs-affine ceiling proposition with an explicit Gaussian counterexample, **machine-checked in Lean 4**, together with an explicit correction and numerical refutation of an erroneous step in our earlier proof (§3.1, Appendices A-B).
2. `rot_share`: a training-free, two-covariance diagnostic (§3.2), with synthetic validation and an external-validation map over seven published datasets recording what transfers, what is confounded, and what is out of scope (§4, Appendix D).
3. A replicated trained-network finding with a known architectural fix, plus the first trained-network test on real images, documenting where the fix holds and where it fails (§4, Appendices F-G).

2 Background and Related Work

Wasserstein geometry. $\mathcal{P}_2$ denotes the Wasserstein-2 space over $\mathbb{R}^d$ with distance W_2 (Santambrogio, 2015; Peyré & Cuturi, 2019). For Gaussians $N(m_1, \Sigma_1)$, $N(m_2, \Sigma_2)$,

$$W_2^2 = \|m_1 - m_2\|^2 + B^2(\Sigma_1, \Sigma_2), \qquad B^2(\Sigma_1, \Sigma_2) = \mathrm{tr}\big(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\big), \quad (1)$$

the Bures metric (Bhatia et al., 2019). The W_2 barycenter of Gaussians is Gaussian, with mean $\sum_i w_i m_i$ and covariance the Bures barycenter $\mathrm{Bar}_w(\{\Sigma_i\})$ (Agueh & Carlier, 2011; Álvarez-Esteban et al., 2016). For two *commuting* SPD matrices the Bures barycenter is the square of the linearly interpolated square roots, $\mathrm{Bar}_w(\Sigma_0, \Sigma_1) = ((1-w)\Sigma_0^{1/2} + w\Sigma_1^{1/2})^2$.

Barycentric conditional generation. Barycentric constructions for conditional generation go back to Wasserstein barycenter regression (Agueh & Carlier, 2011; Bigot & Klein, 2018; Chen et al., 2023). The OOD generalisation bound we probe is that of Pham (2026), whose standing assumption is that the unseen condition is a convex combination of seen ones. Our Proposition 1 isolates that assumption and shows it is what does the work.

Flow matching and OT-informed training. Flow matching (Lipman et al., 2023) and rectified flow (Liu et al., 2023) train continuous normalising flows without simulation; OT-informed variants (Pooladian et al., 2023; Tong et al., 2024) modify the coupling or the vector-field parameterisation to straighten the learned paths. These change *training*; our diagnostic is post-hoc and training-free, and is complementary - it says whether OOD extrapolation is geometrically available at all, before any of these methods is applied.

SPD geometry. Log-Euclidean (LE) (Arsigny et al., 2006; 2007), affine-invariant (AI) (Bhatia, 2007) and Bures-Wasserstein (BW) (Bhatia et al., 2019) are the standard Riemannian structures on SPD matrices. All three give extrapolation rules for $(\mu(c), \Sigma(c))$. Our contribution is not a new geometry but the empirical question of *which* one a trained network can follow, and the measurement of when it cannot (§4).

3 Method

3.1 The ceiling does not bind affine extrapolation

Let the training Gaussians be $N(m_i, \Sigma_i)$, $i = 1, \dots, N$. The *barycentric hull* and the ceiling are

$$\mathcal{B} \;=\; \Big\{ N\Big(\textstyle\sum_i w_i m_i,\; \mathrm{Bar}_w(\{\Sigma_i\})\Big) \;:\; w \in \Delta_N \Big\}, \qquad \mathrm{Term\,I}^*(\bar{c}) \;=\; \min_{\nu \in \mathcal{B}} W_2(\nu, \mu_{\bar{c}}). \quad (2)$$

The *affine* (log-Euclidean) moment predictor uses coefficients a_i with $\sum_i a_i = 1$ and no sign constraint: $\hat{m} = \sum_i a_i m_i$, $\hat{\Sigma} = \exp(\sum_i a_i \log \Sigma_i)$.

Proposition 1 (the ceiling does not bind affine extrapolation)**.** *For Gaussian training measures there exist Σ_0, Σ_1 and a target $\Sigma_{\bar{c}}$ such that the affine predictor attains $W_2 = 0$ while $\mathrm{Term\,I}^* > 0$. Concretely, for the commuting family $\Sigma(c) = Q\,\mathrm{diag}(e^{\ell_0 + c\,\ell_1})Q^\top$ trained at $c \in \{0, 1\}$ with $\ell_1 \neq 0$ and tested at $\bar{c} > 1$, the affine predictor with $a = (1 - \bar{c},\ \bar{c})$ reproduces $\Sigma(\bar{c})$ exactly, while every convex weight $w \in [0, 1]$ satisfies $W_2(\mathrm{Bar}_w, \Sigma(\bar{c})) \geq W_2(\Sigma(1), \Sigma(\bar{c})) > 0$.*

Proof sketch. In the common eigenframe write u, v, t for the square-rooted spectra of $\Sigma(0), \Sigma(1), \Sigma(\bar{c})$. The affine claim is the identity $(1 - \bar{c})\ell_0 + \bar{c}(\ell_0 + \ell_1) = \ell_0 + \bar{c}\ell_1$. For the convex claim, $w \mapsto \|(1-w)u + wv - t\|^2$ is a convex quadratic whose unconstrained minimiser is $w^\star = \langle v - u,\, t - u\rangle / \|v - u\|^2$. A coordinatewise sign lemma - for $\bar{c} \geq 1$, $(v_j - u_j)$ and $(t_j - v_j)$ always share a sign - gives $w^\star \geq 1$, so the constrained minimum over $[0, 1]$ sits at the endpoint $w = 1$ and equals $\|v - t\| > 0$. The full proof is in Appendix A. □

Remark 1 (a step we retract)**.** *Our earlier argument instead claimed $\mathcal{B} = \{\Sigma(c) : c \in [0, 1]\}$ via a componentwise reparameterisation $c_w = (2/\ell_1)\log((1-w) + we^{\ell_1/2})$. This is false for $d \geq 2$ with unequal ℓ_1 components: c_w is a vector, not a scalar, and Bar_w lies strictly off the curve. We verify the refutation numerically (Appendix A, Figure 7) and note that the corrected proof above never needs the claim.*

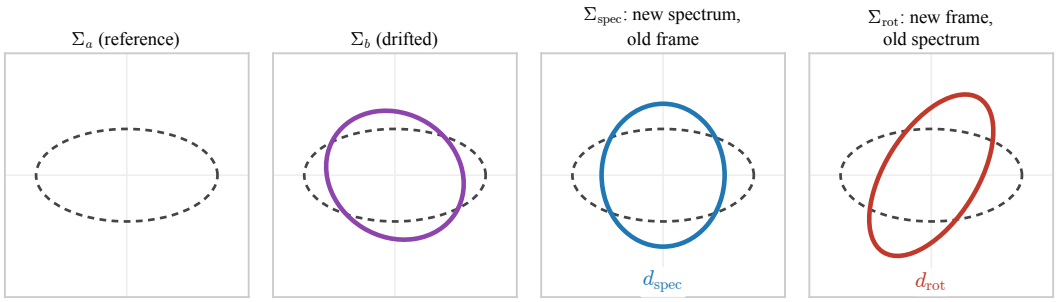

Figure 2: What `rot_share` measures. Given two conditional covariances, we build two surrogates that each change only one thing: Σ_{spec} takes the new spectrum in the old eigenframe, Σ_{rot} takes the new eigenframe with the old spectrum. `rot_share` is the share of the squared Bures drift carried by the frame. Spectral drift (`rot_share` ≈ 0) is extrapolable; frame rotation (`rot_share` ≈ 1) is not.

Remark 2 (what this does and does not say)**.** *Proposition 1 shows the ceiling is not* in general *a floor. It leaves open whether the affine escape is available* for a given family*: on eigenframe-rotation families the affine predictor is far* worse *than the barycenter. Telling the two regimes apart cheaply is the job of* `rot_share`*.*

Machine-checked. Proposition 1 - the affine identity, the sign lemma, the endpoint minimisation and the strict positivity - is formalised in Lean 4 against Mathlib and compiles with no `sorry`, depending only on the three standard axioms. Well-posedness of `rot_share` (below) is checked in the same file. See Appendix B.

3.2 ROT_SHARE: A TRAINING-FREE DIAGNOSTIC

Given two covariances Σ_a, Σ_b of the same dimension, align their eigenframes (signs fixed, $\det = +1$) and form $\Sigma_{\text{spec}} = U_a \Lambda_b U_a^\top$ (new spectrum, old frame) and $\Sigma_{\text{rot}} = U_b \Lambda_a U_b^\top$ (new frame, old spectrum). Then

$$\texttt{rot_share}(\Sigma_a, \Sigma_b) = \frac{d_{\text{rot}}^2}{d_{\text{rot}}^2 + d_{\text{spec}}^2}, \qquad d_{\text{rot}} = B(\Sigma_a, \Sigma_{\text{rot}}), \quad d_{\text{spec}} = B(\Sigma_a, \Sigma_{\text{spec}}). \tag{3}$$

It answers one question (Figure 2): *of the covariance drift between two conditions, how much is eigenframe rotation and how much is spectral scaling?* It needs two covariance estimates - no training, no target samples, no generative model - and lands in $[0, 1]$ by construction (machine-checked, Appendix B). The full instrument adds finite-sample calibration controls that convert the raw score into a GO / NO-GO / AMBIGUOUS verdict; these are described in Appendix C.

3.3 FOUR WAYS TO CONDITION A FLOW-MATCHING NETWORK

To test whether the geometry survives a trained model we compare four conditional rectified-flow networks (Figure 3), all small MLPs trained on $c \in [0, 1]$ only: **naive**, conditioned on the raw scalar c; **moment**, conditioned on the flattened $(\mu(c), \Sigma(c))$, LE-extrapolated at $\bar{c}$ - the architecture the affine-escape theory naively recommends; **moment_film**, the same condition delivered by FiLM modulation (Perez et al., 2018); and **moment_res**, the same condition but with the network predicting only a *residual* on top of the analytic rectified-flow velocity of the extrapolated Gaussian endpoints. Full architectural and training detail is in Appendix F.

4 EXPERIMENTS

We score OOD generation by the headroom fraction $\text{hfrac} = 1 - W_2(\text{pred})/\text{Term I}^*$: $\text{hfrac} = 1$ means the barycentric ceiling was reached, $\text{hfrac} < 0$ means the predictor did *worse* than the best convex blend. The diagnostic's claim is that `rot_share` and hfrac are *negatively* correlated.

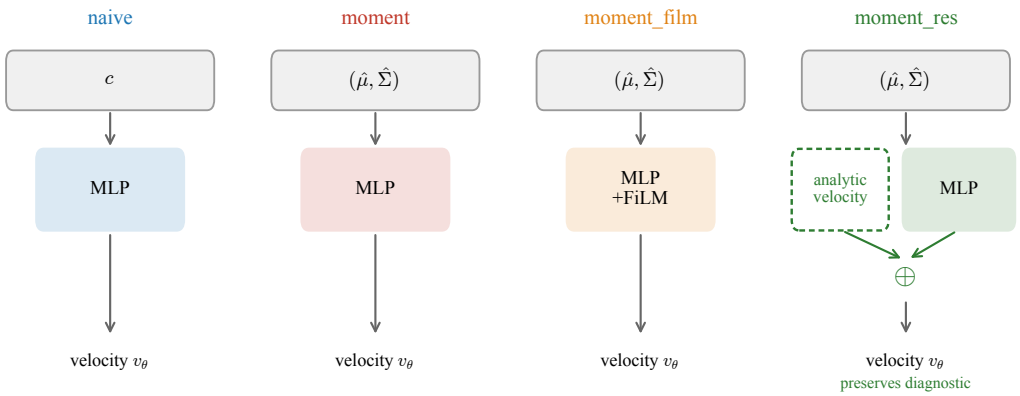

Figure 3: The four conditioning schemes of §4. Only `moment_res` - which supplies the closed-form rectified-flow velocity of the extrapolated Gaussian and learns a bounded residual on top - preserves the diagnostic at both dimensions tested.

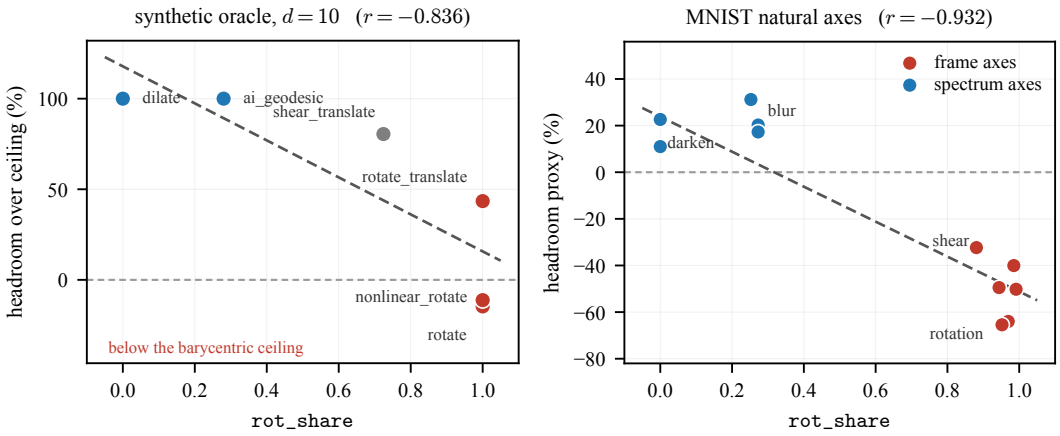

Figure 4: Left: synthetic battery at d=10 with exact moments - spectral-drift families leave the hull essentially for free, rotation families are worse than the barycenter under every SPD geometry tried. **Right:** the same prediction on MNIST with natural continuous axes, where the pre-registered GO/NO-GO verdicts are correct on both headline axes. Per-axis numbers: Appendix D, Table 3.

4.1 Oracle level: the diagnostic works

On the synthetic battery (d=10, exact moments, $\bar{c} = 1.5$) the headroom of the best SPD extrapolation over the barycentric ceiling correlates with `rot_share` at $r = -0.836$ over six finite families; the calibrated covariance-only figure gives $r = -0.972$ (Figure 4, left).

The instrument then transfers to external data. On MNIST (LeCun et al., 1998) with natural continuous condition axes (digit '3', PCA to d=32, 7 training conditions $\times$ 300 images, $\bar{c} \in \{1.25, 1.5, 2.0\}$) it predicts headroom at $r = -0.932$ over the 15 axis $\times$ $\bar{c}$ points, with correct calibrated verdicts: NO-GO on rotation and shear, GO on darkening, decontrast and blur (Figure 4, right).

Published datasets: what transfers and what does not. Run on seven published datasets with *natural*, not self-designed, axes, the frame-detection half transfers universally: COIL-20/COIL-100 (Nene et al., 1996b;a), dSprites (Matthey et al., 2017), 3DShapes (Burgess & Kim, 2018) and UTKFace (Zhang et al., 2017) all read their natural rotation/pose axes as frame-dominated, and COIL-100 at n=100 per condition earns a *calibrated* NO-GO. The spectrum-detection half transfers to grayscale single-object scenes (MNIST, dSprites scale) but is confounded on multi-object colour scenes (3DShapes), where object-background interaction makes spectrum axes read as frame-dominated. On single-cell pseudotime (Paul et al., 2015; Wolf et al., 2018) the instrument honestly reports AMBIGUOUS. Full tables, per-round detail and the pooled correlations are in Appendix D (Tables 4-6, Figure 8).

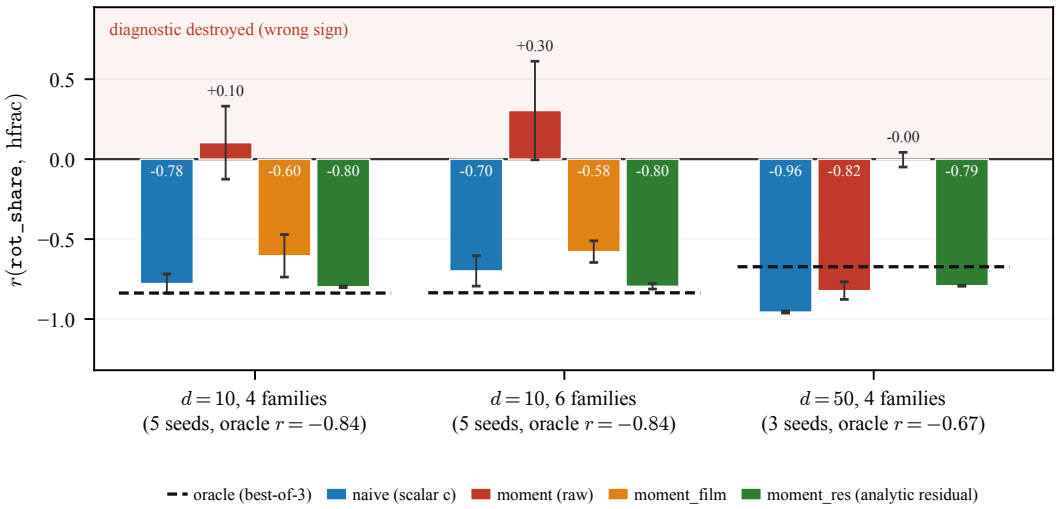

Figure 5: The headline result. Correlation between `rot_share` and the achieved headroom fraction, for each conditioning scheme, at two dimensions and two family counts. Error bars span the seed range. Conditioning on the raw extrapolated moments - the design the theory naively suggests - lands in the shaded region at d=10, i.e. it *inverts* the diagnostic; FiLM is unstable across dimension; supplying the analytic velocity and learning only a residual tracks the oracle at both dimensions. Per-seed values: Figure 10; per-family: Figure 11.

Table 1: Mean $r(\texttt{rot_share}, \text{hfrac})$ over seeds. Bold marks the two results that matter: raw moment conditioning inverts the diagnostic at d=10, and the analytic-residual head restores it at both dimensions.

variant	d=10, 4 fam. (5 seeds)	d=10, 6 fam. (5 seeds)	d=50, 4 fam. (3 seeds)
oracle (best-of-3)	−0.838	−0.836	−0.673
naive	−0.778	−0.699	−0.956
moment (raw)	**+0.103**	**+0.304**	−0.822
moment_film	−0.604	−0.578	−0.003
moment_res	**−0.798**	**−0.795**	**−0.791**

Scope: continuous conditions only. A 26-letter point-cloud cross-check gives $r(\texttt{rot_share}, \text{slack}) = +0.25$ and $r(\texttt{rot_share}, \text{gap}) = -0.00$ (n=26), and EMNIST letters (Cohen et al., 2017) give $r(\texttt{rot_share}, \text{Term I}^*) = -0.01$ over 26 held-out classes. The continuous-axis prediction does **not** transfer to unordered class identity. We report this as a documented boundary, not a refutation: the instrument is for *ordered* conditions - dose, time, angle, morph parameter (Appendix E, Figure 9).

4.2 Contact with a trained network

Sections 3.1-3.2 are oracle-level. Does the diagnostic survive when a *trained* network does the generation? We train the four schemes of §3.3 on families spanning the `rot_share` spectrum (`dilate` 0.00, `ai_geodesic` 0.28, `shear` 0.72, `rotate` 1.00), 6000 steps, RK4 sampling, and score by the same hfrac convention. The single-seed table reproduced **row-for-row** across two different GPUs, and the oracle pipeline reproduced exactly, so the numbers below are not single-machine artefacts (Appendix F).

Table 1 and Figure 5 give four findings. **(i) The sign flip replicates.** Raw moment conditioning destroys the diagnostic at d=10 - every seed near zero or positive, mean +0.10 on four families and +0.30 on six. At d=50 it no longer flips but still *degrades* it relative to naive (−0.82 vs −0.96). **(ii) The plain scalar-conditioned network preserves it** (−0.70 to −0.78 at d=10; −0.956 at d=50): with no explicit manifold machinery at all, it is a better moment extrapolator than the raw moment head. **(iii) FiLM is not a reliable fix** (−0.58 at d=10 but −0.003 at d=50). **(iv) The analytic-residual head is the fix**: it matches or exceeds the oracle correlation at both dimensions and is the most seed-stable variant.

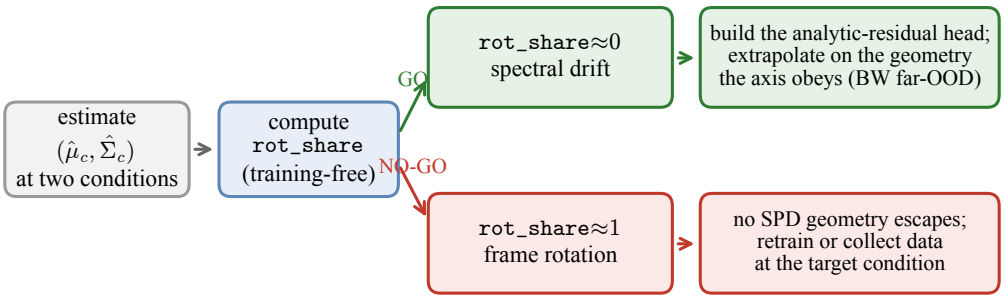

Figure 6: The practitioner's recipe. Two covariance estimates and no training decide whether OOD extrapolation is geometrically available; the verdict then selects the architecture and the test-time SPD geometry.

Interpretation. Conditioning on a raw (μ, Σ) vector asks the network to rediscover SPD geometry in a space it has only ever seen along a 1-D parameter curve; the extrapolated conditioning point lies off that curve and the network has no inductive bias to handle it. Supplying the analytic velocity of the extrapolated Gaussian, and training only a bounded residual, is the missing bias. **The escape mechanism must be built into the architecture, not learned from a raw condition vector.**

4.3 The same protocol on real images

Rerunning the protocol on real pixels (dSprites and 3DShapes, three natural axes each, PCA to d=32, three seeds, three held-out $\bar{c}$ per axis = 18 rows) gives a mixed and instructive picture. `moment_res` has the lowest generated-vs-truth W_2 on **11 of 18 rows**, and on dSprites the synthetic finding replicates cleanly: it is $\sim$3.3$\times$ closer to the true OOD distribution than the naive scalar head on all three axes. But the *correlation* does not transfer verbatim: on dSprites the barycentric ceiling is nearly reachable ($\mathrm{Term\,I}^* \approx 0.002$, two orders below the W_2 scale) so hfrac degenerates, and on 3DShapes the confound of §4 propagates into the trained models. A failure mode the synthetic families never exhibited also appears: at far $\bar{c}$ the LE-extrapolated covariance on the 3DShapes scale axis becomes near-singular and both moment heads explode (up to $W_2 \approx 2 \times 10^8$) while the scalar head stays finite. Switching the *test-time* extrapolation from log-Euclidean to Bures-Wasserstein removes the instability completely - $2.1 \times 10^8 \to 3.78$ at $\bar{c} = 3.0$ - and makes the analytic-residual head stable everywhere while winning all three orientation rows. Full tables: Appendix G (Tables 7-8, Figure 12).

The design rule, with its scope: **build the escape into the architecture, and extrapolate the test-time Gaussian on the SPD geometry the axis actually obeys.** Figure 6 states the resulting recipe.

5 Limitations

A retracted proof step. One step of our earlier argument was false (Remark 1). The proposition survives with a corrected, machine-checked proof, but the episode bounds how much unverified geometric reasoning should be trusted - including ours.

E3-real is mixed, and recorded as such. The analytic-residual advantage replicates on dSprites, but the hfrac correlation degenerates there because $\mathrm{Term\,I}^* \approx 0$, and LE test-time conditioning explodes on far-OOD 3DShapes scale. Bures-Wasserstein fixes the latter. The honest claim is therefore: `moment_res` is the best OOD generator on clean axes and stable everywhere under BW, with naive still winning the colour-scene spectrum axes where the diagnostic itself is confounded.

Spectrum-side confound on complex scenes. On 3DShapes the spectrum axes read as frame-dominated. The frame-detection side transfers universally; the spectrum side is validated only on grayscale single-object data.

MNIST validation is single-digit and transform-based. The axes are natural but applied by us; the published datasets are fully external and their mixed outcome is reported above.

Scope. The discrete-condition null result is a limitation of the diagnostic's scope and should be cited whenever `rot_share` is recommended.

Budget and coverage. d=50 uses 3 seeds (vs 5 elsewhere); the synthetic E3 conditioning uses LE extrapolation only; networks are small MLPs.

Few-shot correction (demoted). An earlier k-shot correction of the barycentric reference is *not* a contribution here: it was noisy, sign-unstable across repetition counts, and its only surviving predictor was weak. It is left to future work (Appendix H).

6 CONCLUSION

The barycentric ceiling on OOD conditional generation is real but conditional: it binds convex combinations of the training conditions, and an affine extrapolation on the SPD manifold is not a convex combination. Whether that escape is actually available is decided by how the conditional covariance drifts, and that is measurable from two covariance estimates before any training happens. The measurement survives contact with trained networks - but only if the escape is built into the architecture. Handing a network the extrapolated moments and hoping it learns the geometry inverts the diagnostic and makes the model worse than a scalar-conditioned baseline; giving it the closed-form velocity and asking for a residual restores oracle-level behaviour. We also correct, and machine-check, our own earlier proof. We hope the negative results - the confounded spectrum side on colour scenes, the null result on unordered classes, and the retracted proof step - are as useful as the positive ones.

IMPACT STATEMENT

This work is a diagnostic for when conditional generative models can be trusted to extrapolate beyond their training conditions. Its intended effect is conservative: it tells a practitioner *not* to trust an extrapolation when the geometry does not support it, and to collect data or retrain instead. We see that as a mild positive for safety, in that it makes a specific failure mode cheap to detect before deployment rather than expensive to discover after.

AI USE STATEMENT

In this work, generative AI tools were used to (i) create or modify visual figures based on our experimental results, (ii) edit English grammar, phrasing, and vocabulary, and (iii) assist in producing Lean formalizations for checking selected mathematical equations. The tools were not used to generate data, formulate the research questions or theoretical claims, design or run experiments, analyze or interpret experimental results, or make scientific decisions. They were also not used to replace the authors' judgment about the scientific content.

We reviewed all AI-assisted work. The experimental data and results were checked against the original runs and stored result files; AI-assisted figures were checked against those results; and the Lean formalizations were compiled and checked by the authors to ensure that they corresponded to the equations stated in the paper. AI-assisted language edits were manually reviewed for meaning and accuracy. The authors take responsibility for the final content of this work, including all text, claims, figures, and formalizations produced with AI assistance.

REFERENCES

Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

Pedro C. Álvarez-Esteban, Eustasio del Barrio, Juan A. Cuesta-Albertos, and Carlos Matrán. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Log-Euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic Resonance in Medicine*, 56(2):411–421, 2006.

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 29(1):328–347, 2007.

Rajendra Bhatia. *Positive Definite Matrices*. Princeton Series in Applied Mathematics. Princeton University Press, 2007.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.

Jérémie Bigot and Thierry Klein. Characterization of barycenters in the Wasserstein space by averaging optimal transport maps. *ESAIM: Probability and Statistics*, 22:35–57, 2018.

Chris Burgess and Hyunjik Kim. 3D shapes dataset. `https://github.com/deepmind/3d-shapes`, 2018.

Yaqing Chen, Zhenhua Lin, and Hans-Georg Müller. Wasserstein regression. *Journal of the American Statistical Association*, 118(542):869–882, 2023.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. EMNIST: Extending MNIST to handwritten letters. In *International Joint Conference on Neural Networks (IJCNN)*, pp. 2921–2926, 2017.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dSprites: Disentanglement testing Sprites dataset. `https://github.com/deepmind/dsprites-dataset/`, 2017.

Leonardo de Moura and Sebastian Ullrich. The Lean 4 theorem prover and programming language. In *International Conference on Automated Deduction (CADE)*, pp. 625–635, 2021.

Sameer A. Nene, Shree K. Nayar, and Hiroshi Murase. Columbia object image library (COIL-100). Technical Report CUCS-006-96, Department of Computer Science, Columbia University, 1996a.

Sameer A. Nene, Shree K. Nayar, and Hiroshi Murase. Columbia object image library (COIL-20). Technical Report CUCS-005-96, Department of Computer Science, Columbia University, 1996b.

Franziska Paul, Ya'ara Arkin, Amir Giladi, Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Deborah Winter, David Lara-Astiaso, Meital Gury, Assaf Weiner, Eyal David, Nadav Cohen, Felicia Kathrine Bratt Lauridsen, Simon Haas, Andreas Schlitzer, Alexander Mildner, Florent Ginhoux, Steffen Jung, Andreas Trumpp, Bo Torben Porse, Amos Tanay, and Ido Amit. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, 163(7): 1663–1677, 2015.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, 2018.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

Long HC Pham. Decomposing out-of-distribution error in conditional flow matching via wasserstein geometry. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=orzcDIBxa0.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T. Q. Chen. Multisample flow matching: Straightening flows with minibatch couplings. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28100–28127, 2023.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. Birkhäuser, 2015.

The mathlib Community. The Lean mathematical library. In *ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP)*, pp. 367–381, 2020.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research (TMLR)*, 2024.

F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19:15, 2018.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4352–4360, 2017.

A Full proof of Proposition 1, and the step we retract

A.1 Setting

Fix $d \geq 1$, an orthonormal frame Q, and spectral vectors $\ell_0, \ell_1 \in \mathbb{R}^d$ with $\ell_1 \neq 0$. Define the one-parameter *commuting* family

$$\Sigma(c) \;=\; Q\,\mathrm{diag}\big(e^{\ell_0 + c\,\ell_1}\big)\,Q^\top, \qquad c \in \mathbb{R}, \tag{4}$$

with all means zero, so that W_2 reduces to the Bures metric B. Train on $c_0 = 0$, $c_1 = 1$; test at $\bar{c} > 1$.

Because every member shares the eigenframe Q, all quantities below are functions of the spectra alone. Writing $\lambda_j(c) = e^{\ell_{0j} + c\ell_{1j}}$ and using the commuting-case Bures formula, for any two members

$$B^2(\Sigma(a), \Sigma(b)) \;=\; \sum_{j=1}^{d} \Big(\sqrt{\lambda_j(a)} - \sqrt{\lambda_j(b)}\Big)^2, \tag{5}$$

and the Bures barycenter of $\Sigma(0), \Sigma(1)$ at weight w is

$$\mathrm{Bar}_w = \Big((1-w)\Sigma(0)^{1/2} + w\,\Sigma(1)^{1/2}\Big)^2 = Q\,\mathrm{diag}\Big(\big((1-w)u_j + wv_j\big)^2\Big)Q^\top, \tag{6}$$

where we abbreviate the square-rooted spectra

$$u_j = e^{\ell_{0j}/2}, \qquad v_j = e^{(\ell_{0j}+\ell_{1j})/2}, \qquad t_j = e^{(\ell_{0j}+\bar{c}\,\ell_{1j})/2}. \tag{7}$$

A.2 Part (A): the affine predictor is exact

Take $a_1 = \bar{c}$ and $a_0 = 1 - \bar{c}$, which sum to one with $a_0 < 0$ — an affine, not convex, combination. Since $\log \Sigma(c) = Q(\ell_0 + c\ell_1)Q^\top$,

$$\hat{\Sigma} = \exp\big((1-\bar{c})\log\Sigma(0) + \bar{c}\log\Sigma(1)\big) = \exp\big(Q(\ell_0 + \bar{c}\,\ell_1)Q^\top\big) = \Sigma(\bar{c}), \tag{8}$$

using only $(1-\bar{c})\ell_0 + \bar{c}(\ell_0 + \ell_1) = \ell_0 + \bar{c}\ell_1$. Means are zero throughout, so $W_2\big(N(0,\hat{\Sigma}), N(0,\Sigma(\bar{c}))\big) = 0$.

A.3 Part (B): the sign lemma

Lemma 1 (sign lemma)**.** *For every $\bar{c} \geq 1$ and every coordinate j,*

$$(v_j - u_j)\,(t_j - v_j) \;\geq\; 0. \tag{9}$$

Proof. Fix j and split on the sign of ℓ_{1j}. If $\ell_{1j} \geq 0$ then $\ell_{0j} \leq \ell_{0j} + \ell_{1j} \leq \ell_{0j} + \bar{c}\,\ell_{1j}$ (the second inequality because $\bar{c} \geq 1$ and $\ell_{1j} \geq 0$), so by monotonicity of exp, $u_j \leq v_j \leq t_j$ and both factors are ≥ 0. If $\ell_{1j} \leq 0$ the same two inequalities reverse, giving $u_j \geq v_j \geq t_j$, so both factors are ≤ 0 and their product is again ≥ 0. If $\ell_{1j} = 0$ all three coincide and the product is 0. □

A.4 Part (C): the barycentric minimum sits at the endpoint

By (5)–(6), for $w \in \mathbb{R}$

$$f(w) \;:=\; W_2^2\big(\mathrm{Bar}_w, \Sigma(\bar{c})\big) = \sum_j \big((1-w)u_j + wv_j - t_j\big)^2 = \sum_j \big((w-1)p_j - q_j\big)^2, \tag{10}$$

where $p_j = v_j - u_j$ and $q_j = t_j - v_j$. Expanding,

$$f(w) - f(1) \;=\; (w-1)^2 \underbrace{\textstyle\sum_j p_j^2}_{=:S \geq 0} \;-\; 2(w-1)\underbrace{\textstyle\sum_j p_j q_j}_{=:P}. \tag{11}$$

Lemma 1 gives $p_j q_j \geq 0$ for every j, hence $P \geq 0$. For any $w \leq 1$ we have $w - 1 \leq 0$, so $-2(w-1)P \geq 0$, and $(w-1)^2 S \geq 0$ trivially. Therefore

$$f(w) \;\geq\; f(1) \qquad \text{for every } w \leq 1, \tag{12}$$

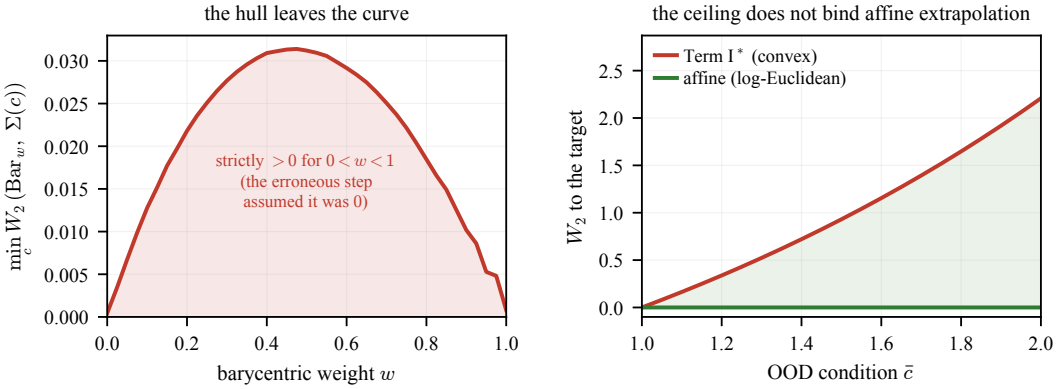

Figure 7: Left: the quantity the retracted step asserted was zero. For a commuting family with unequal ℓ_1 components, the Bures barycenter Bar_w is at strictly positive Bures distance from *every* member $\Sigma(c)$ of the curve, for all interior w — the hull is genuinely a two-dimensional object, not the curve segment. **Right:** the proposition itself, unaffected: the convex ceiling $\mathrm{Term\,I}^*$ grows with $\bar{c}$ while the affine predictor stays exactly on the target. Both panels are produced by `check_prop1.py` (d=6, seed 6).

and in particular for every w in the simplex $[0,1]$. Equivalently, the unconstrained minimiser of the convex quadratic f is $w^\star = 1 + P/S \geq 1$, so f is non-increasing on $[0,1]$.

Finally $f(1) = \sum_j q_j^2 = \sum_j (t_j - v_j)^2$. Since $\bar{c} > 1$ and $\ell_1 \neq 0$, there is a coordinate j_0 with $\ell_{1j_0} \neq 0$, whence $(\ell_{0j_0} + \bar{c}\,\ell_{1j_0}) \neq (\ell_{0j_0} + \ell_{1j_0})$ and, by injectivity of exp, $t_{j_0} \neq v_{j_0}$. Therefore

$$\mathrm{Term\,I}^*(\bar{c}) \;=\; \min_{w\in[0,1]} W_2\big(\mathrm{Bar}_w, \Sigma(\bar{c})\big) = W_2\big(\Sigma(1), \Sigma(\bar{c})\big) \;>\; 0. \tag{13}$$

Together with Part (A) this proves Proposition 1. □

A.5 The step we retract, and why it is false

An earlier version of this argument proceeded differently. It observed that for a single eigenvalue the square-root interpolation can be inverted, giving

$$c_w \;=\; \frac{2}{\ell_1}\,\log\!\big((1-w) + w\,e^{\ell_1/2}\big), \tag{14}$$

and then asserted $\mathrm{Bar}_w = \Sigma(c_w)$ with $c_w \in [0,1]$, concluding $\mathcal{B} = \{\Sigma(c) : c \in [0,1]\}$ — the hull *is* the curve segment — so that $\mathrm{Term\,I}^*$ is the distance from $\Sigma(\bar{c})$ to that segment.

This is false for $d \geq 2$. The division and logarithm in (14) are componentwise, so c_w is a *vector* $(c_{w,1}, \ldots, c_{w,d})$, one parameter per eigenvalue. It equals a single scalar only when all components of ℓ_1 coincide (or $d = 1$). Whenever they differ, no scalar c satisfies $\mathrm{Bar}_w = \Sigma(c)$, and the barycenter leaves the curve.

Numerically, for $d = 6$ with a random ℓ_1 (seed 6) and $w = 0.5$, the componentwise values (14) spread over $[0.482,\ 0.586]$ — a range of 1.04×10^{-1} — and

$$\min_{c\in\mathbb{R}} W_2\big(\mathrm{Bar}_{0.5},\ \Sigma(c)\big) \;=\; 3.12 \times 10^{-2} \;>\; 0. \tag{15}$$

Figure 7 (left) plots this residual across w; it vanishes only at the two endpoints, where the barycenter coincides with a training covariance. The effect grows with dimension: the same experiment gives 8.9×10^{-3} at d=2, 1.1×10^{-1} at d=10, and 2.3×10^{-1} at d=32.

The proof in Appendices A.A–C never uses the claim: it minimises f over the simplex directly. It is also strictly stronger, since (11) shows $f(w) \geq f(1)$ for all $w \leq 1$, not merely on $[0,1]$. We record the retraction here rather than silently substituting the corrected argument.

A.6 Numerical verification

`check_prop1.py` verifies all three parts over a grid $d \in \{2, 6, 10, 32\} \times \bar{c} \in \{1.25, 1.5, 2.0, 3.0\}$, checking that the affine predictor attains $W_2 = 0$ to machine precision (relative to the spectral scale,

since $\exp \circ \log$ round-off grows with it), that $\min_j (v_j - u_j)(t_j - v_j) \geq 0$, that $w^\star \geq 1$, that the brute-force minimum over a 4001-point grid on $[0, 1]$ is attained at $w = 1$, and that it agrees with $\sqrt{f(1)}$ to $\sim 10^{-14}$. All 16 configurations pass. The same script reports the refutation of §A.5 as a separate check.

B MACHINE-CHECKED FORMALISATION IN LEAN 4

The corrected proof is formalised in Lean 4 (Moura & Ullrich, 2021) against Mathlib (The mathlib Community, 2020) and is included in the supplementary material as `lean/BarycentricCeiling/`. It builds with `lake build` and contains no `sorry`; every theorem below depends only on the three standard axioms `propext`, `Classical.choice` and `Quot.sound`, as reported by `#print axioms`.

What is formalised, and what is assumed. Because the family commutes, the matrix layer carries no content: every quantity is a function of the spectra. The formalisation therefore works directly on square-rooted spectra $\mathtt{sq}\ \ell_0\ \ell_1\ c = (j \mapsto \exp((\ell_{0j} + c\,\ell_{1j})/2))$, and takes as *definitions* the two classical reductions used in (5)–(6): that the squared Bures metric between co-diagonalised SPD matrices is the squared Euclidean distance of their square-rooted spectra, and that the Bures barycenter of two commuting SPD matrices is the linear interpolation of their square roots (Bhatia, 2007; Agueh & Carlier, 2011). These two facts are the interface to the matrix layer and are *not* re-proved in Lean; everything above them is.

Table 2: Contents of `BarycentricCeiling/Basic.lean`. All eight results compile with no `sorry`.

Lean name	statement
`affine_log_combination`	the affine identity $(1-\bar{c})\ell_0 + \bar{c}(\ell_0 + \ell_1) = \ell_0 + \bar{c}\,\ell_1$
`affine_extrapolation_exact`	the log-Euclidean predictor has Bures cost exactly 0
`sign_lemma`	$0 \leq (v_j - u_j)(t_j - v_j)$ for all j, whenever $\bar{c} \geq 1$
`bary_sub_target`	$\mathrm{Bar}_w - t = (w-1)(v-u) - (t-v)$ coordinatewise
`bary_cost_min_at_one`	$f(1) \leq f(w)$ for every $w \leq 1$
`termIstar_pos`	$0 < f(1)$ when $\bar{c} > 1$ and $\ell_1 \neq 0$
`proposition_one`	the two halves combined: affine cost $= 0$, convex cost $\geq \mathrm{Term\,I}^* > 0$
`rotShare_mem_Icc`	`rot_share` $\in [0, 1]$ whenever the denominator is positive
`rotShare_pure_spectrum`, `rotShare_pure_frame`	pure spectral drift reads 0; pure frame drift reads 1

The Lean development also makes the retraction visible in the source: the module docstring states plainly that the "hull = curve" step is false for $d \geq 2$ and that the formal proof avoids it. Note that the formal statement of `bary_cost_min_at_one` is over all $w \leq 1$, which is the stronger fact (11) supports; the simplex restriction is a corollary.

C THE STEP-0 INSTRUMENT AND ITS CALIBRATION

Raw `rot_share` is a population quantity; on finite samples it is biased upward, because sample eigenvectors of a covariance estimated from n points in d dimensions are noisy even when the population eigenframe does not move. The instrument used for all real-data results therefore reports a *calibrated* verdict alongside the raw score, using two internal controls:

- a **positive control** (POS): a synthetic pure-spectral drift matched to the observed (n, d), which *should* read GO. Its measured headroom gives the optimistic anchor.
- a **negative control** (NEG): a synthetic pure-frame drift matched to the same (n, d), which *should* read NO-GO, giving the pessimistic anchor.

If the two anchors do not separate, the configuration is under-powered at that (n, d) and the instrument returns **UNRESOLVED** rather than a verdict — this is what COIL-20 does at n=20, d=16, with the controls only 10.8% apart. When they do separate, the observed headroom is rescaled against them and the verdict is GO, NO-GO, or AMBIGUOUS if the calibrated `rot_share` lands

between the anchors. Paul15 pseudotime is the one case that legitimately returns AMBIGUOUS with resolved controls (POS $+39.6\%$ vs NEG -10.3%; calibrated `rot_share` 0.246; calibrated headroom -116.1%).

The headroom proxy used on real data is the covariance-only headroom of the best SPD extrapolation over the barycentric reference, computed in a PCA subspace fitted on the *training* conditions only, so that no information from the held-out condition enters the feature map.

D REAL-DATA VALIDATION IN FULL

D.1 MNIST WITH NATURAL CONTINUOUS AXES

Digit '3', PCA to d=32, 7 training conditions $\times$ 300 images, held-out $\bar{c} \in \{1.25, 1.5, 2.0\}$. Axes are natural image transformations applied by us (hence the limitation noted in §5), split a priori into frame-like and spectrum-like.

Table 3: MNIST natural axes. Pooled $r(\texttt{rot_share}, \text{headroom proxy}) = -0.932$ over the 15 axis $\times$ $\bar{c}$ points. Verdicts are the instrument's calibrated output at $\bar{c} = 1.5$.

axis	kind	`rot_share`	headroom proxy
rotation	frame	0.944–0.985	$-32\% \cdots -64\%$
shear	frame	0.881–0.991	$-32\% \cdots -65\%$
darkening	spectrum	0.000	$+22.6\%$
decontrast	spectrum	0.000	$+11.0\%$
blur	spectrum	0.252–0.273	$+17\% \cdots +31\%$
calibrated verdicts at $\bar{c} = 1.5$: rotation **NO-GO** (-57.0%), shear NO-GO, darkening **GO** ($+87.7\%$), decontrast GO, blur GO			

D.2 PUBLISHED DATASETS, ROUND 1

Natural (not self-designed) continuous axes, matched samples per condition, PCA features, held-out extreme conditions.

Table 4: Published datasets, round 1. Pooled $r(\texttt{rot_share}, \text{headroom}) = -0.349$ over 14 rows; per dataset, COIL-20 -0.55, dSprites -0.57, 3DShapes $+0.35$.

dataset / axis	kind	`rot_share`	headroom proxy
COIL-20 azimuth pose	frame	0.858–0.949	$-33\% \cdots -70\%$
dSprites orientation	frame	0.882–0.933	$-14\% \cdots -102\%$
dSprites scale	spectrum	$\mathbf{0.098}$	$\mathbf{+7.6\%}$
3DShapes orientation	frame	0.851–0.920	$-5\% \cdots -39\%$
3DShapes scale	spectrum	0.840–0.843	$-6\% \cdots -44\%$
3DShapes object hue	spectrum	0.424–0.819	$-31\% \cdots -42\%$

Frame axes read correctly on all three datasets and dSprites scale reads correctly as spectral. The 3DShapes spectrum axes are *misclassified* as frame-dominated: on multi-object colour scenes the pixel-covariance frame/spectrum split is confounded by object–background interaction. COIL-20's own calibration honestly reports UNRESOLVED at n=20, d=16.

D.3 ROUND 2: COIL-100 POSE AND EMNIST

Natural object pose on a published dataset is thus correctly and *calibratedly* flagged as frame-dominated, and the discrete-scope boundary is independently confirmed on a real image dataset.

D.4 ROUND 3: UTKFACE AGE AND SINGLE-CELL PSEUDOTIME (EXPLORATORY)

Honest reading. The frame-detection half of `rot_share` transfers to published datasets (COIL-20/COIL-100/dSprites/3DShapes/UTKFace); the spectrum-detection half transfers to grayscale

Table 5: Round 2. COIL-100 closes COIL-20's under-powering; EMNIST is the discrete-scope control.

dataset	outcome
COIL-100 azimuth pose (n=100/condition)	raw `rot_share` 0.956–0.959, raw headroom $-39.8/-41.9/-57.0\%$; controls now **separate** (POS $+14.4\%$ vs NEG -16.5%); calibrated verdict at $\bar{c}$=1.5 is **NO-GO**, calibrated headroom -100.4%, calibrated `rot_share` 0.964
EMNIST letters (26 classes, discrete)	$r(\texttt{rot_share}, \text{Term I}^*) = -0.01$ over the held-out classes

Table 6: Round 3, recorded as exploratory observations rather than claims.

dataset / axis	outcome
UTKFace age	reads **frame-like**, not spectral: `rot_share` $0.85/0.85/0.77$ with headroom $+3.1/-1.5/-19.9\%$ at $\bar{c} = 1.167/1.333/1.5$. Ageing changes the eigenframe of face covariance, not merely its spectrum — the axis had been pre-registered as "spectrum-ish?" and the instrument disagreed.
Paul15 dpt pseudotime	a genuine frame/spectrum **mix**: raw `rot_share` $0.42/0.42/0.32$ with headroom $-53.8/-20.3/-58.3\%$; at $\bar{c}$=1.5 controls resolve (POS $+39.6\%$ vs NEG -10.3%), calibrated headroom -116.1%, calibrated `rot_share` $0.246 \rightarrow$ **AMBIGUOUS**

single-object scenes (MNIST, dSprites) but not to multi-object colour scenes (3DShapes); and on single-cell pseudotime the instrument reports the drift as a frame/spectrum mix. This is recorded as a limitation, not hidden.

Protocol per dataset. Fix the axis and the OOD $\bar{c}$ *before* looking; estimate $(\hat{\mu}_c, \hat{\Sigma}_c)$ per condition in a PCA/autoencoder subspace fitted on training conditions only; run the instrument with calibration controls; record `rot_share`, the headroom proxy and the verdict; pool and report the correlation. Success criterion, declared in advance: negative $r(\texttt{rot_share}, \text{headroom})$ and correct verdicts on the pre-declared frame vs spectrum axes.

E SCOPE: CONTINUOUS CONDITIONS ONLY

A 26-letter point-cloud cross-check with discrete, unordered conditions gives $r(\texttt{rot_share}, \text{slack}) = +0.250$ and $r(\texttt{rot_share}, \text{gap}) = -0.004$ (n=26): the continuous-axis prediction does **not** transfer to unordered class identity. This is independently confirmed on real images by EMNIST letters, $r(\texttt{rot_share}, \text{Term I}^*) = -0.013$ (n=26 held-out classes).

We report this as a boundary, not a refutation. The instrument's documented scope is *ordered/continuous* conditions — dose, time, angle, morph parameter — and the result is consistent with that scope: with unordered labels there is no "direction" along which to extrapolate, so the entire convex-vs-affine question is ill-posed. The null result should be cited whenever `rot_share` is recommended.

F E3: TRAINED-NETWORK PROTOCOL AND FULL RESULTS

F.1 PROTOCOL

Families spanning the `rot_share` spectrum, all with fixed means where the family name says so: `dilate_fixed_mean` (`rot_share` = 0.00), `ai_geodesic_fixed_mean` (0.28 at d=10), `shear_translate` (0.72), `rotate_fixed_mean` (1.00), plus `nonlinear_rotate_fixed_mean` and `rotate_translate` for the six-family battery. `translate_iso` is excluded because its `rot_share` is undefined (the covariance does not move).

Each network is a small MLP trained from scratch with a straight-line (rectified-flow) conditional flow-matching objective (Liu et al., 2023; Lipman et al., 2023), 6000 steps, on $c \in [0, 1]$ only, with

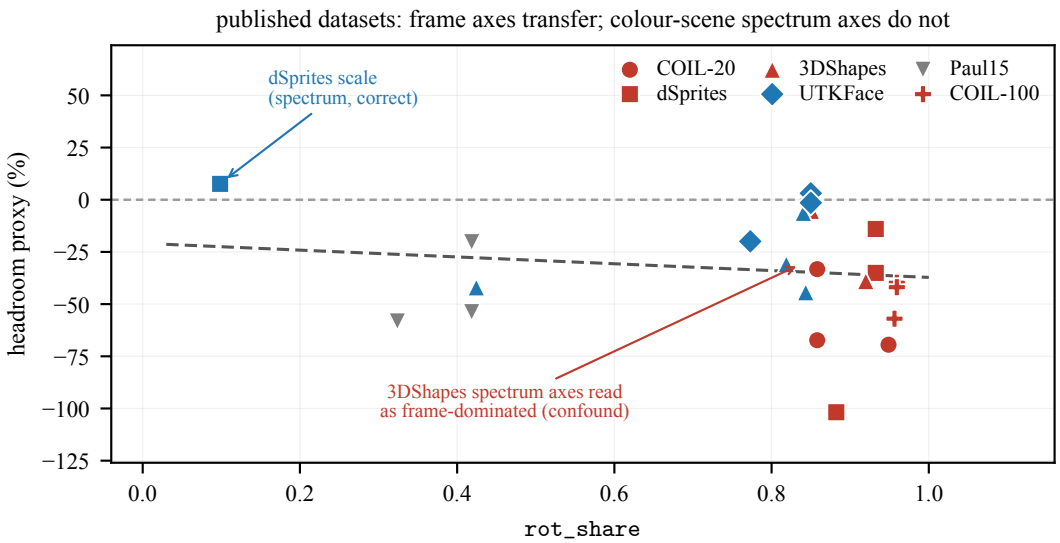

Figure 8: All published-dataset rows in one view. The frame axes (red) cluster at high `rot_share` with negative headroom on every dataset; dSprites scale (blue, far left) is the clean spectral case. The confounded 3DShapes spectrum axes are the cluster called out at high `rot_share`, which is why the pooled correlation is only -0.349 despite the frame side being unanimous.

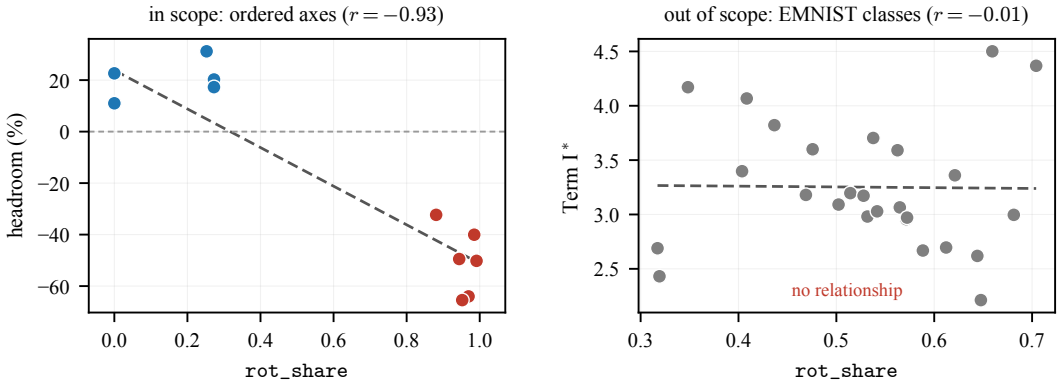

Figure 9: The documented scope boundary. **Left:** on ordered axes the diagnostic predicts headroom. **Right:** across 26 EMNIST letter classes — unordered categories — `rot_share` carries no information about $\mathrm{Term\,I}^*$.

a 256-point interpolated moment grid as the training condition distribution. Sampling uses RK4. Scoring uses $\mathrm{hfrac} = 1 - W_2(\mathrm{pred})/\mathrm{Term\,I}^*$, the same convention as the oracle battery, so the columns are directly comparable. At d=50 the training sampler uses a Cholesky factorisation rather than batched `eigh`, which is exact for these SPD covariances and roughly $50\times$ faster on the GPU used.

The four conditioning schemes are those of §3.3 and Figure 3. The oracle reference column is **best-of-3** over {LE, BW, AI} per family; an earlier internal version of this table used LE only and consequently disagreed with the oracle battery on three of four families. That discrepancy was traced to this single undisclosed methodological choice and is recorded here so the columns cannot be misread: $\mathrm{Term\,I}^*$ itself, which hfrac is measured against, is method-independent.

F.2 Replication

The original single-seed table reproduced **row-for-row** on different hardware (RTX 5060 Ti → RTX 3060). The oracle pipeline (56/56 tests, E0, E0b, figure) is deterministic in its exact-moment quantities, and these reproduce bit-for-bit across machines: the median shape-only out-of-hull headroom 45.1%,

the d=50 survival check -4.9%, $r = -0.836$ for E0b and $r = -0.972$ for the calibrated figure, together with the falsifier control (covariance gain on the constant-Σ family, $+0.0000$).

One number is *not* bit-reproducible and we flag it rather than round it away: the n=200 finite-sample survival check. It is the only quantity in the battery estimated from sampled points rather than exact moments, and although its draw is seeded, it passes through a multi-restart optimiser whose result depends on the BLAS/LAPACK build. We record -14.1% on the machine used for the paper's result JSONs and -12.7% on a second machine. The check is comfortably negative either way, so the pre-registered verdict on the oracle battery is **PARTIAL** on both (the rule required $\geq 50\%$ median headroom *plus* both survival checks, and the finite-sample check fails regardless).

F.3 Per-seed and per-family detail

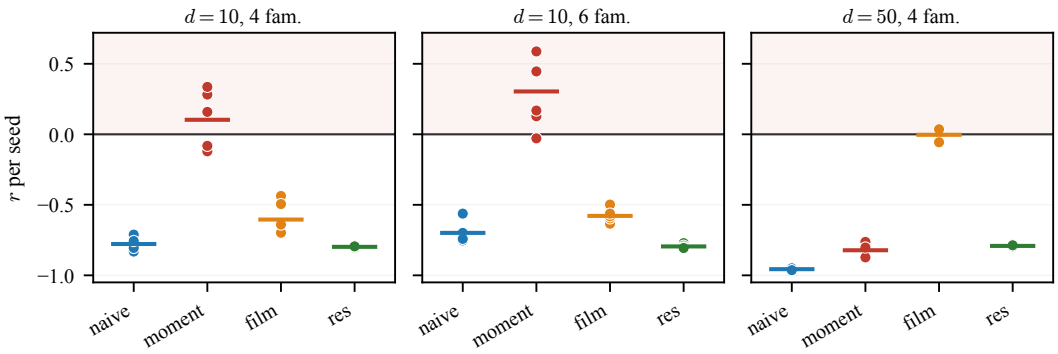

Figure 10: Every seed, every variant. The raw moment head is not merely noisy: at d=10 no seed is strongly negative, and several are clearly positive. The analytic-residual head is the tightest cluster in all three settings.

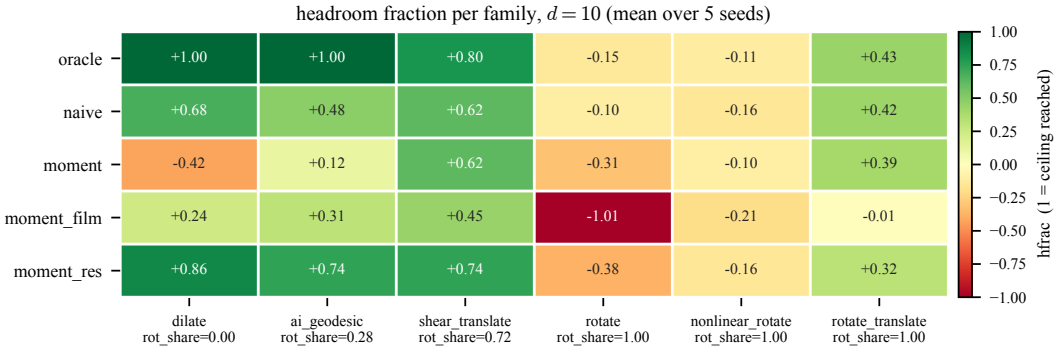

Figure 11: Headroom fraction per family at d=10, families ordered by `rot_share`. The diagnostic's prediction is the left-to-right green-to-red gradient. It is visible in the oracle, naive and `moment_res` rows and broken in the `moment` row, whose `dilate` entry (-0.42) is the single cell most responsible for the sign flip: the easiest family for the theory is the one the raw moment head fails hardest on.

Per-seed correlations at d=10 with four families are naive $\{-0.711, -0.762, -0.831, -0.806, -0.757\}$, moment $\{+0.282, -0.121, -0.082, +0.159, +0.336\}$, `moment_film` $\{-0.703, -0.437, -0.698, -0.640, -0.495\}$ and `moment_res` $\{-0.796, -0.805, -0.795, -0.799, -0.795\}$. With six families the sign flip is *stronger* ($+0.304$ vs $+0.103$), and the `moment_res` range is $[-0.772, -0.806]$. At d=50, naive carries the diagnostic almost perfectly (-0.956, range $[-0.950, -0.963]$), the raw moment head no longer flips but degrades (-0.822), FiLM collapses to chance (-0.003), and `moment_res` is stable at -0.791, above the oracle's own -0.673 on those four families.

Where the residual head still costs headroom. On the pure-rotation family `moment_res` sits at $\approx -38\%$ (d=10) and $\approx -32\%$ (d=50) versus the oracle's -14.7% and -6.0%. The architectural fix restores the *diagnostic*; it does not make rotation extrapolable, which is exactly what the diagnostic predicts.

G E3-REAL: FULL RESULTS ON PUBLISHED IMAGES

dSprites × 3 axes and 3DShapes × 3 axes; PCA to d=32; CFM trained on the actual pixels; matched samples per condition (4000 train / 4000 held out); 3 seeds; 3 held-out $\bar{c}$ per axis, i.e. 18 rows.

Table 7: E3-real with log-Euclidean test-time conditioning. Mean generated-vs-truth W_2 (lower is better).

group	naive	moment (raw)	moment_film	moment_res
dSprites (9 rows)	0.083	0.086	5.64	**0.026**
3DShapes orientation	1.34–3.46	1.79–7.19	1.75–17.5	**1.26–3.02**
3DShapes scale (far $\bar{c}$)	1.6–3.6	explodes	nan	explodes
3DShapes object hue	**2.6–4.3**	6.2–11.5	5.9–27.4	4.9–7.2

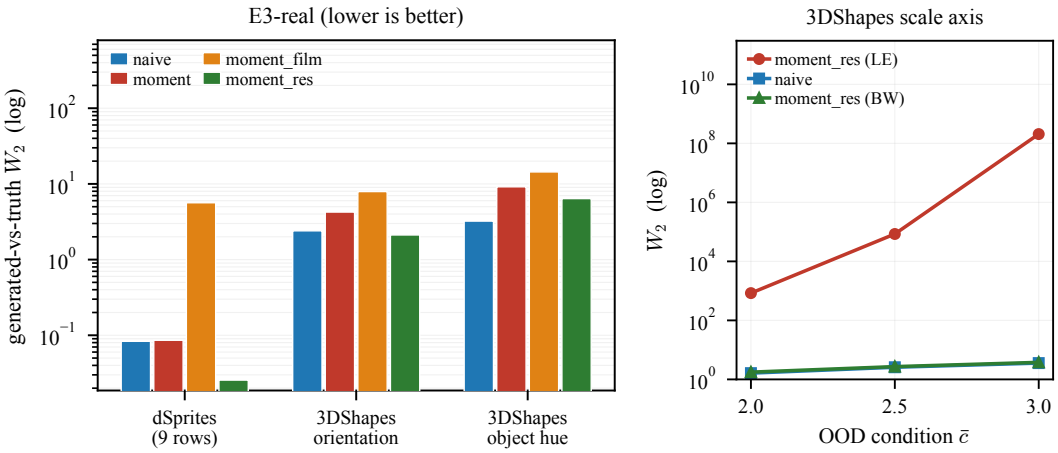

Figure 12: Left: E3-real W_2 by group (log scale; the diverging 3DShapes scale rows are excluded here and tabulated separately in Table 8). **Right:** the far-OOD failure and its fix. Log-Euclidean test-time extrapolation drives the moment heads through eight orders of magnitude; the Bures–Wasserstein geometry keeps them within a few percent of the scalar baseline.

`moment_res` has the lowest W_2 on **11 of 18 rows**; naive wins the other 7, mostly the extreme 3DShapes rows. Three honest conclusions:

1. **On dSprites the synthetic finding replicates cleanly.** The analytic-residual head is ∼3.3× closer to the true OOD distribution than the naive scalar head, on all three natural axes.
2. **The** hfrac**/`rot_share` correlation does not transfer as-is.** On dSprites the barycentric ceiling is nearly reachable ($\mathrm{Term\,I}^* \approx 0.002$, two orders of magnitude below the W_2 scale), so hfrac degenerates numerically; on 3DShapes the Step-0 confound of Appendix D propagates into the trained models.
3. **The explosion is a flaw of the test-time SPD extrapolation, not of the head.** At far $\bar{c}$ the LE-extrapolated covariance on the 3DShapes scale axis becomes near-singular and both moment heads diverge while the scalar head stays finite. Switching test-time extrapolation to Bures–Wasserstein removes the instability completely (Table 8).

Table 8: Bures–Wasserstein test-time conditioning removes the far-OOD explosion on the 3DShapes scale axis.

3DShapes scale, $\bar{c}$	naive	moment_res (LE test)	moment_res (BW test)
2.0	1.63	840	**1.76**
2.5	2.58	84 754	**2.71**
3.0	3.57	206 366 098	**3.78**

With BW test conditioning the analytic-residual head is stable everywhere and wins all three orientation rows (1.22/1.82/2.33 vs naive 1.34/2.40/3.46); naive still wins the colour-scene spectrum axes (scale, hue), where Appendix D already showed the pixel-covariance split is confounded.

Implementation note. The 3DShapes HDF5 file uses gzip chunks of shape $(15000, 4, 4, 1)$, so per-image random reads cost ≈ 0.5 s each; the loader reads the full array into RAM (5.9 GB) and slices contiguously, which takes 0.5 s for the whole dataset. Result JSONs are written incrementally with resume support.

H THE DEMOTED FEW-SHOT CORRECTION

An earlier line of work attempted a k-shot correction: given a handful of samples at the OOD condition, correct the barycentric reference before measuring headroom. We report it here as a negative result rather than a contribution.

The signal was noisy and *sign-unstable across repetition counts* — the direction of the correction's effect changed depending on how many repeats were averaged, which is the signature of a statistic dominated by estimation variance rather than by the effect being measured. Its only surviving predictor was the absolute k=0 gap, and that was weak ($r \approx 0.7$ on $n \approx 8$–12 runs, i.e. a confidence interval wide enough to include much weaker relationships). Few-shot correction of the barycentric reference remains plausible, but it did not yield a stable signal in our hands and we leave it to future work rather than reporting an underpowered positive.