WHEN CONDITIONAL GENERATORS CANNOT EXTRAPOLATE: AN EXACTLY MEASURABLE DECOMPOSITION OF OUT-OF-DISTRIBUTION ERROR

Anonymous authors
Paper under double-blind review

ABSTRACT

Conditional generators answer requests they never encountered in training, but when an answer fails we cannot tell whether the model is inadequate or the target was unreachable from the available data. A recent theory separates this failure into target geometry, representation distortion, and training fit. Its sharpest criterion, however, relies on a proof step that controls a different, alignment-dependent quantity, while its experiments estimate every term through proxies. The theory therefore neither certifies the claimed boundary nor reveals which cause actually binds. We repair it by separating an unconditional certificate from a sharper criterion that holds under coherent alignment, formalizing every statement in a proof assistant, and turning distance from the training hull into a test of what any inference-time reweighting can reach. We then construct two benchmarks-capital-letter point clouds and single-cell perturbation populations-where the distances, barycenters, error terms, and sampling floor are computed exactly under held-out evaluation. In both domains, geometry predicts the unseen-condition error while representation distortion and training fit do not bind. Reweighting, condition mixing, and representation regularization leave the geometric gap in place; one genuine target example collapses it. The implication is a practical triage rule before training: determine whether a request is already reachable, requires new examples, or lies beyond what tuning alone can recover.

1 INTRODUCTION

Conditional generative models are routinely deployed on conditions that never appeared during training: a perturbation absent from a biological screen, a class outside the label set, a control value between two seen settings. The usual response to poor out-of-distribution (OOD) output is to swap architectures, add regularizers, or fine-tune, in the hope the model will "generalize." What is missing is a theory of *what is even possible*: given a target condition, is a bad sample a fixable modeling failure or an unreachable target?

The difficulty is that OOD error is reported as a single opaque number. A large Wasserstein distance between generated and true samples conflates at least three causes: the target may lie far from anything seen (a geometric problem), the condition embedding may distort distances (a representation problem), or the model may fit its training conditions poorly (an optimization problem). Without separating these, "the model is weak" and "the target is unreachable" are indistinguishable, and tuning proceeds blind.

A recent analysis makes the separation precise: the OOD error of a conditional flow-matching generator admits a three-term upper bound-interpolation sparsity (TERM I), embedding distortion (TERM II), and in-distribution fit (TERM III) (Pham, 2026). That work, however, validates the bound only through *proxies*: TERM I is approximated by distances in a VAE latent space and TERM II by a finite-difference Jacobian heuristic. Proxy validation is indirect. It cannot tell whether the bound is *tight*, whether the dominant term is truly *irreducible*, or whether the residual error *equals* the geometric term-because none of the three terms is measured directly. These are exactly the questions a practitioner needs answered before deciding whether to keep tuning.

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.

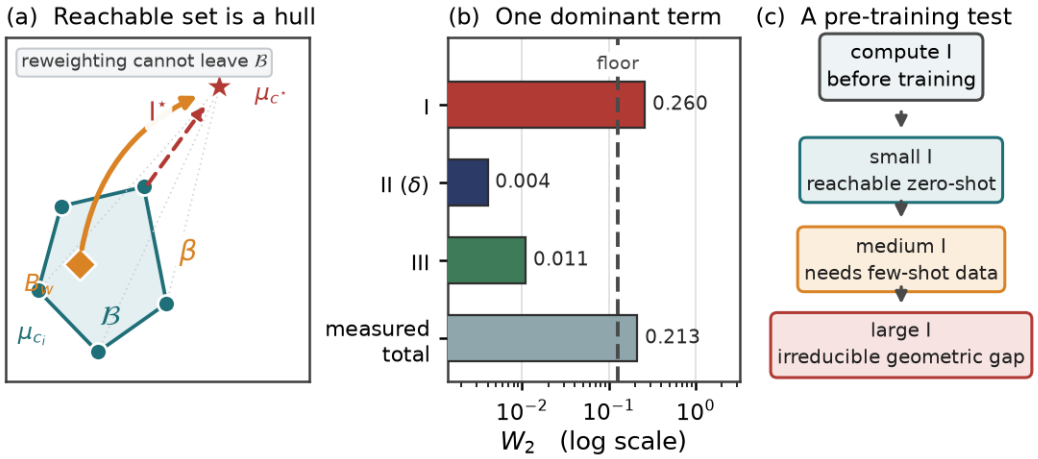

Figure 1: The claim in one picture. (a) Inference-time reweighting can only re-mix the training conditions, so its reachable set is their geodesic hull $\mathcal{B}$; a target outside $\mathcal{B}$ is separated from every reachable output by an irreducible gap $\mathrm{I}^\star$. **(b)** On the letter benchmark the three budgets differ by two orders of magnitude: the geometric term dominates, and the measured total tracks it. **(c)** Because TERM I is computable before training, it acts as a triage test on any requested condition.

We answer them on two benchmarks where all three terms are exactly computable (Fig. 1). The first renders each condition as a capital letter, a 2D point cloud; the Wasserstein distance between any two glyphs is computed exactly by Hungarian matching, so a target's distance to the convex hull of the other conditions-the entire TERM I term-is *computed*, not estimated, and the embedding's non-isometry δ (TERM II) is read off directly. The second treats each genetic perturbation in a Perturb-seq screen as a condition and each cell as a point in a PCA embedding of its transcriptome; again W_2 and its barycenters are exact.

Formalizing first changed the paper. Before running the experiments we formalized the decomposition and every proposition around it in Lean 4 with Mathlib (de Moura & Ullrich, 2021; The mathlib Community, 2020). It showed that the claim $\beta \leq$ TERM I-that the barycentric cost never exceeds the interpolation-sparsity budget-does *not* follow from the triangle inequality, as its published justification asserts. What the barycenter construction proves is a bound in terms of the distances *under the alignment the barycenter itself induces* (Prop. 2), and that budget is an upper bound for, not equal to, TERM I. The sharp inequality needs an extra hypothesis, *coherent alignment*, which holds automatically in one dimension but not in general (Prop. 3). We therefore measure the coherence residual directly, and find that $\beta \leq$ TERM I holds on all 26 glyphs but fails on 5 of 80 perturbations-by at most 0.029, well inside the sampling floor (Sec. 4.4). The two *certified* bounds are never violated. Every named theorem and lemma carries no `sorry` and no axioms beyond Lean's three standard ones; the development is released with the code (App. B).

Contributions. **(1)** A machine-checked theory: the decomposition, the sampling-noise floor, the certified barycentric bound, its sharp form under coherent alignment, the irreducibility of the barycentric floor to inference-time reweighting, and the target-blurring of entropic coupling-each stated so that what is proved is exactly what is used (Sec. 3, App. C). **(2)** Two benchmarks on which TERM I/TERM II/TERM III are *exactly* measurable, with a leakage-free, two-domain validation: TERM I predicts OOD error at $r = 0.87$ on letters and $r = 0.93$ on perturbations ($r = 0.95$ in 8D), with bootstrap intervals, permutation tests, and a fidelity check certifying that the measured cost *is* the geometric barycenter distance. **(3)** The zero-/few-shot boundary: the TERM I floor resists every inference-time intervention we test (mean $|\Delta| < 0.05$) yet collapses with a single out-of-hull example, in both domains-so the boundary is a property of the data, not of the tuning.

2 Background and Related Work

Where this sits. Flow matching learns a continuous-time velocity field transporting a source to a target density (Lipman et al., 2023), and is closely related to rectified flow (Liu et al., 2023), stochastic interpolants (Albergo & Vanden-Eijnden, 2023), and score-based diffusion (Song et al., 2021); optimal-transport couplings straighten its trajectories (Tong et al., 2024). When a target condition is hard, the standard levers are inference-time guidance and reweighting (Dhariwal & Nichol, 2021; Ho & Salimans, 2022); our battery evaluates the flow-matching analogues and finds (Prop. 4) that they cannot help outside the training hull. Where domain generalization (Zhou et al., 2022) and generative generalization bounds (Arora et al., 2017) bound average-case error, we *localize* which targets are reachable, measuring the three-term bound of Pham (2026) exactly rather than through the VAE-latent and Jacobian proxies that work uses. The analysis rests on Wasserstein geometry (Villani, 2009; Peyré & Cuturi, 2019) and barycenters (Agueh & Carlier, 2011), computed here by free-support fixed-point iteration (Cuturi & Doucet, 2014; Álvarez-Esteban et al., 2016); the subtlety our formal pass exposes is a known one in that literature, since W_2 over $\mathbb{R}^d$ has non-negative curvature and the variance inequalities that tame barycenters in non-positively curved spaces (Sturm, 2003) need extra regularity here (Kim & Pass, 2017; Le Gouic et al., 2023). We use the standardized Norman-2019 Perturb-seq data (Norman et al., 2019; Peidli et al., 2024), exact Hungarian rather than entropic coupling (Cuturi, 2013), and few-shot adaptation in the spirit of low-rank adaptation (Hu et al., 2022). App. A gives the full discussion, including the point-cloud generator literature (Lee et al., 2019; Luo & Hu, 2021), single-cell perturbation models (Lotfollahi et al., 2023), and evaluation metrics (Kynkäänniemi et al., 2019).

Setup. Let $\mathcal{C}$ be a condition space and, for each condition c, let μ_c be a probability measure on $\mathbb{R}^d$. A conditional flow-matching model learns a velocity field $v(x, t, e_c)$ depending on a learned embedding e_c of c; sampling integrates the field from a fixed source. We are given training conditions $\{c_i\}_{i=1}^n$ and a target $c^\star$ with true measure $\mu_{c^\star}$; the model's output law is $\hat{p}_{c^\star}$. For measures with finite second moment, $W_2(\alpha, \beta) = (\inf_{\pi \in \Pi(\alpha,\beta)} \int \|x-y\|^2 d\pi)^{1/2}$; for empirical measures on N points it is computed exactly by a linear assignment (Hungarian) problem. Given $w \in \Delta^{n-1}$, the *Wasserstein barycenter* is $B_w = \arg\min_\nu \sum_i w_i W_2^2(\nu, \mu_{c_i})$ (Agueh & Carlier, 2011).

3 Method

3.1 The decomposition, and what is machine-checked

Theorem 1 (Three-term decomposition; restated from Pham (2026))**.** *Let the model be trained on $\{c_i\}_{i=1}^n$ with in-distribution generation error $\varepsilon_{train} = \max_i W_2(\hat{p}_{c_i}, \mu_{c_i})$ and an embedding with distortion constant δ. For a target $c^\star$ and weights $w(c^\star) \in \Delta^{n-1}$,*

$$W_2\big(\hat{p}_{c^\star},\, \mu_{c^\star}\big) \;\le\; \underbrace{K \sum_i w_i\, W_2\big(\mu_{c_i}, \mu_{c^\star}\big)}_{\text{Term I: } \textit{interpolation sparsity}} \;+\; \underbrace{\delta}_{\text{Term II: } \textit{distortion}} \;+\; \underbrace{C\,\varepsilon_{train}}_{\text{Term III: } \textit{fit}}, \tag{1}$$

with constants $K, C \ge 1$ depending only on the model class.

Term I is the weighted Wasserstein distance from the target to the training conditions; it vanishes when the target lies at their barycenter and grows as it leaves their hull. Term II is the non-isometry of the condition embedding, Term III the in-distribution error. Consistent with Pham (2026), our claims concern the bound's *structure*-which term dominates and whether it is predictive-not its absolute constants.

Every statement in this section is formalized in Lean 4 with Mathlib and checked to depend on no axioms beyond `propext`, `Classical.choice` and `Quot.sound`; App. B gives the theorem-by-theorem correspondence and maps the development. Wasserstein space enters only as a pseudometric space, and point clouds of fixed size as elements of a normed space, so the formal statements are exactly the ones the experiments use.

3.2 THE SAMPLING-NOISE FLOOR

Two finite real draws of the same condition differ, so W_2 has a non-zero in-distribution baseline.

Definition 1 (Sampling-noise floor)**.** $\Phi_m = \mathbb{E}\, W_2(\hat{\mu}_N, \hat{\mu}'_N)$, the expected distance between two independent N-sample empirical measures of condition m.

Proposition 1 (The floor is the achievable best)**.** *If the generator's law equals the target law, its expected evaluation cost against a fresh N-sample equals Φ_m; and $\Phi_m > 0$ for finite N.*

Per-condition floors vary with compactness (thin J: 0.067; wide U: 0.189; alphabet mean 0.127; App. E); we therefore report absolute W_2 with the floor as a reference band rather than subtracting a single constant.

3.3 THE ZEROED-TERM II/TERM III LIMIT IS A BARYCENTRIC COST

We now make precise what the OOD cost *becomes* when the embedding is isometric (TERM II $= 0$) and the model fits its training conditions to the floor (TERM III $= 0$). The idealized inference-time protocol for an unseen $c^\star$ feeds the field the convex combination $e_{\text{bary}} = \sum_i w_i e_i$ with softmax weights $w_i(c^\star) \propto \exp(-W_2(\mu_{c^\star}, \mu_{c_i})/\tau)$; under those assumptions the field's output is the barycenter B_w, so the OOD cost is the *barycentric cost* $\beta(c^\star) = W_2(B_{w(c^\star)}, \mu_{c^\star})$, a function of the training geometry alone, computable before training. The fidelity check of Sec. 4.2 certifies this premise empirically.

What β can be bounded by is where the published argument and the formal one part company. The free-support barycenter of equal-size clouds is a convex combination $B_w = \sum_i w_i\, \Pi_i(\mu_{c_i})$ of the training clouds *re-indexed* by the assignments Π_i that the barycenter iteration itself selects (Cuturi & Doucet, 2014; Álvarez-Esteban et al., 2016). Minkowski's inequality therefore bounds β by the weighted sum of distances *under those* assignments, which we call the coherent budget $\mathrm{I}_{\mathrm{coh}}$.

Proposition 2 (Certified barycentric bounds)**.** $\beta(c^\star) \le \mathrm{I}_{\mathrm{coh}}(c^\star) := \sum_i w_i \|\Pi_i(\mu_{c_i}) - \mu_{c^\star}\|$*, and* TERM I$(c^\star) \le \mathrm{I}_{\mathrm{coh}}(c^\star)$*. Independently, whenever B_w minimises the barycentric functional at least as well as the target does, $\beta(c^\star) \le$* TERM I$(c^\star) + \mathrm{I}_2(c^\star)$ *where $\mathrm{I}_2 = (\sum_i w_i W_2^2(\mu_{c_i}, \mu_{c^\star}))^{1/2} \ge$* TERM I.

Because TERM I $\le \mathrm{I}_{\mathrm{coh}}$, the first display does *not* give $\beta \le$ TERM I: the published one-line justification ("the triangle inequality") proves a bound against the larger budget. The sharp statement needs the assignments to be coherent.

Proposition 3 (Sharp bound under coherent alignment)**.** *If $\|\Pi_i(\mu_{c_i}) - \mu_{c^\star}\| = W_2(\mu_{c_i}, \mu_{c^\star})$ for every i (*coherent alignment*: the assignment that builds the barycenter is also optimal against the target), then $\beta(c^\star) \le$* TERM I$(c^\star)$.

Coherent alignment is automatic for $n = 1$ and in $d = 1$, where W_2 is flat and all optimal couplings share the sort order; in $d \ge 2$ it can fail (App. D). We measure the residual $\mathrm{I}_{\mathrm{coh}} -$ TERM I directly in Sec. 4.4.

3.4 WHAT IS IRREDUCIBLE

Proposition 4 (β is irreducible to inference-time reweighting)**.** *Any inference-time reweighting (classifier guidance, mixup, temperature) reaches only the set of barycenters $\mathcal{B} = \{B_{w'} : w' \in \Delta^{n-1}\}$, whose minimum distance to $\mu_{c^\star}$ is* TERM I$^\star := \inf_{w' \in \Delta^{n-1}} W_2(B_{w'}, \mu_{c^\star})$*. Hence no reweighting reduces the OOD cost below* TERM I**; any output strictly better than* TERM I* *is not a barycenter of the training conditions, so attaining it requires a measure outside the training hull-out-of-hull data.*

Lemma 1 (Entropic projection blurs the target)**.** *Let P be a doubly stochastic coupling between source and target clouds and $\bar{y}_i = \sum_j (N P_{ij})\, y_j$ its barycentric projection. Then the projected targets have no larger dispersion than the true targets; the contraction is* strict *for any row that spreads mass over targets at two different distances from the reference, which is every row when P has full support ($\varepsilon > 0$); and when P is a permutation (exact matching, $\varepsilon = 0$) dispersion is preserved exactly.*

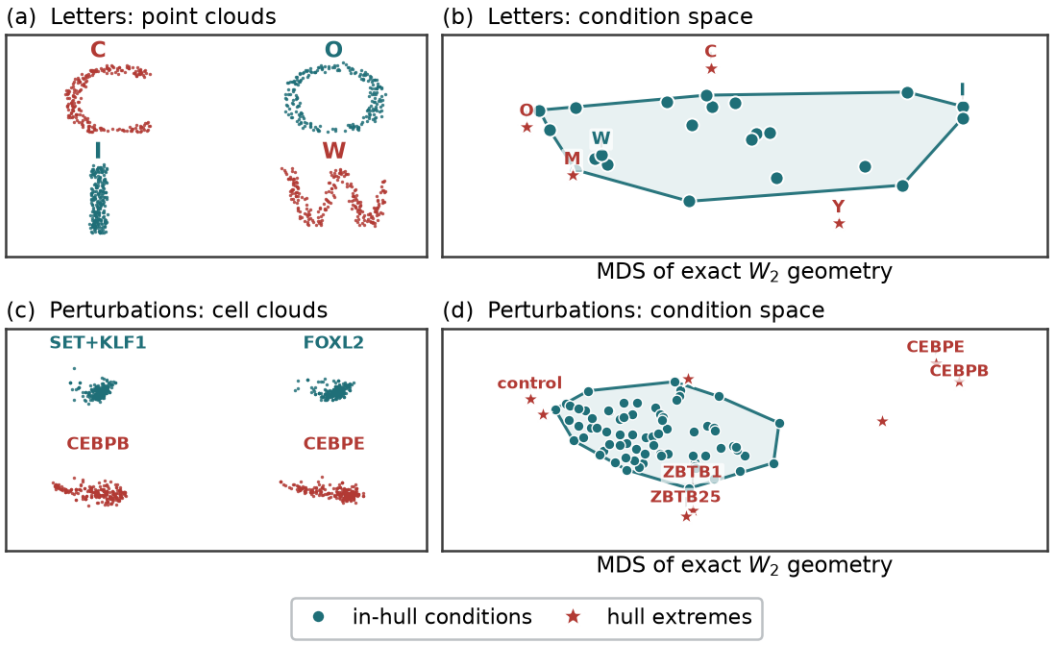

Figure 2: Two domains, one geometry. (a,c) Conditions are point clouds: glyphs, and cells in a PCA embedding of their transcriptome. **(b,d)** Metric MDS of the exact pairwise W_2 matrix, with the convex hull of the interior conditions shaded. The starred conditions are the ones a leave-one-out protocol must extrapolate to.

3.5 Two benchmarks with exact measurement

Letters. Conditions are the 26 capital letters. Each glyph is rasterized, its foreground pixels centered and scaled to $[-1, 1]^2$, and $N = 256$ points sampled with Gaussian jitter. This is the right testbed for three reasons. *Exactness*: W_2 between any two clouds is computed exactly, so the entire TERM I term is *computed*, not estimated. *Controlled topology*: glyphs vary in holes (`O`, `B`, `Q`), strokes, and near-disconnection (thin `I`, `J`), spanning a wide, known range of geometric difficulty. *Known relative geometry*: a target's distance to the convex hull of the others is a property of the data, independent of any model.

Single-cell perturbations. We use the standardized Norman-2019 Perturb-seq dataset (Norman et al., 2019; Peidli et al., 2024): 111,445 cells across 196 CRISPR perturbations. Each *condition* is a perturbation; μ_c is the distribution of cell states in a PCA embedding of log-normalized, highly-variable gene expression. We keep the 80 largest cohorts (≥ 463 cells each; 68,221 cells), sample $N = 256$ cells per cloud, and compute W_2 exactly. Because cells are real measurements (no synthetic jitter), the floor is the honest finite-sample baseline of the assay. This is the domain the introduction motivates, and the one the prior proxy study could not handle exactly. Preprocessing is in App. F; Fig. 2 shows both condition spaces.

3.6 Model, exact coupling, and protocol

We train a per-point velocity field $v(x_i, t, e_c)$ on a continuous condition embedding (a 5-layer per-point MLP, hidden width 256), with a centroid source and *exact per-step Hungarian coupling* between source and target clouds. A continuous embedding is essential: it lets us feed a barycentric reference for unseen conditions and read off δ. Exact coupling is essential for clean measurement, because the entropic alternative forms regression targets by barycentric projection, which blurs them (Lemma 1; for a holed glyph the average can land in the hole). To drive TERM II down we add an *isometry regularizer* pulling embedding pairwise distances toward W_2/scale; δ is the residual non-isometry. Full hyperparameters are in App. E.

To verify that the field actually produces the barycentric reference, we compute the *exact* free-support barycenter B_{exact} of the training clouds under the same weights and report the *fidelity*

$W_2(\text{output}, B_{\text{exact}})$. All OOD numbers use strict leave-one-condition-out: each of the 26 glyphs is held out and the field retrained on the other 25 (26 trainings); perturbations use leave-group-out (6 folds), so held-out clouds never enter training and their embeddings never receive a gradient. The only information about a held-out condition at evaluation is the few-shot $W_2(\mu_{c^\star}, \mu_{c_i})$ used to place it, which yields the *best-case* reference and therefore strengthens the irreducibility claim.

4 EXPERIMENTS

4.1 IN-DISTRIBUTION FIT REACHES THE FLOOR, AND EXACT COUPLING MATTERS

We first establish that TERM III is not the bottleneck. Across the letters folds the flow-matching loss converges to a mean of 0.0035 (range 0.0026-0.0044) and the in-distribution gap above floor (TERM III) averages $+0.011$: the model reaches the sampling-noise floor on its training conditions, as required for the zeroed-TERM III regime of Prop. 4. On single-cell perturbations the gap is $+0.003$ above a floor of 0.063 (2D) and $+0.018$ above 0.175 (8D). Exact Hungarian coupling also beats Sinkhorn barycentric coupling *despite* the latter attaining a lower training loss, exactly as Lemma 1 predicts (App. H, Fig. 8).

4.2 THE FIDELITY CHECK PASSES IN BOTH DOMAINS

With exact coupling the field reproduces the exact barycenter: across all 26 held-out glyphs the mean fidelity $W_2(\text{output}, B_{\text{exact}})$ is 0.129, essentially equal to the floor 0.127 (App. Fig. 13). On single-cell perturbations the mean fidelity is 0.050 against a floor of 0.063, and in 8D 0.151 against 0.175. This certifies the premise of Sec. 3.3: the OOD costs measured next are geometric barycenter distances, not defects of the conditioning mechanism.

4.3 TERM I PREDICTS THE OOD ERROR, AND TERM II IS NOT THE BOTTLENECK

With mean leave-one-out TERM II at $\delta \approx 0.004$ and TERM III at the floor, the OOD transport cost is predicted by TERM I on letters at Pearson $r = 0.87$ $[0.70, 0.95]$, slope 0.86 $[0.64, 1.21]$, Spearman $\rho = 0.78$, permutation $p < 0.001$ (Fig. 3a). The genuine geometric extremes-`C`, `Z`, `V`, `L`, `W`, farthest from the training hull-incur the largest costs, while the most central glyphs (`I`, `J`) sit at the floor. On single-cell perturbations the same prediction holds at $r = 0.93$ $[0.74, 0.97]$, slope 1.17, permutation $p < 0.001$ (Fig. 3b): the perturbations farthest from the training hull carry the largest irreducible gaps, and control conditions sit at the floor. The mechanism thus transfers from glyph geometry to transcriptomic geometry.

The correlation is neither a leverage nor a temperature artifact, and it replicates across seeds, step budgets and an 8D embedding; the full battery-deletion of the three highest-leverage conditions, Theil-Sen slopes, jackknife ranges and a temperature sweep-is in App. I.

TERM II is not the bottleneck. To confirm the residual is geometric and not a representation defect, we train a non-isometric counterpart with the regularizer disabled, raising δ from 0.001 to 0.074. The OOD gap is unchanged (mean $|\Delta| = 0.007$; Fig. 5, "isometry reg."). Once a target is far from the hull, fixing the embedding does not help: the error is TERM I, exactly as Eq. (1) predicts when TERM II is small.

4.4 THE CERTIFIED BOUNDS HOLD; THE SHARP ONE OCCASIONALLY DOES NOT

Proposition 2 is unconditional and Proposition 3 is not, so we test both. Across 26 glyphs and 80 perturbations, the coherent bound $\beta \le \mathrm{I}_{\mathrm{coh}}$ and the quadratic bound $\beta \le \text{TERM I} + \mathrm{I}_2$ are satisfied in *every* case, as are the barycenter-minimality premises. The unproved sharp inequality $\beta \le \text{TERM I}$ holds on all 26 glyphs but fails on 5 of 80 perturbations (Fig. 4), by at most 0.029-about $0.45\times$ the sampling floor-and only where the coherence residual $\mathrm{I}_{\mathrm{coh}} - \text{TERM I}$ is large (mean $+0.006$, max $+0.047$). This is precisely the failure mode the formalization predicted, and it is invisible to a proxy study: the violations are small, confined to hull-interior conditions, and would have been absorbed into estimation noise. The empirical picture is that $\beta \le \text{TERM I}$ is an excellent working approximation whose exceptions are diagnosable in advance from the coherence residual.

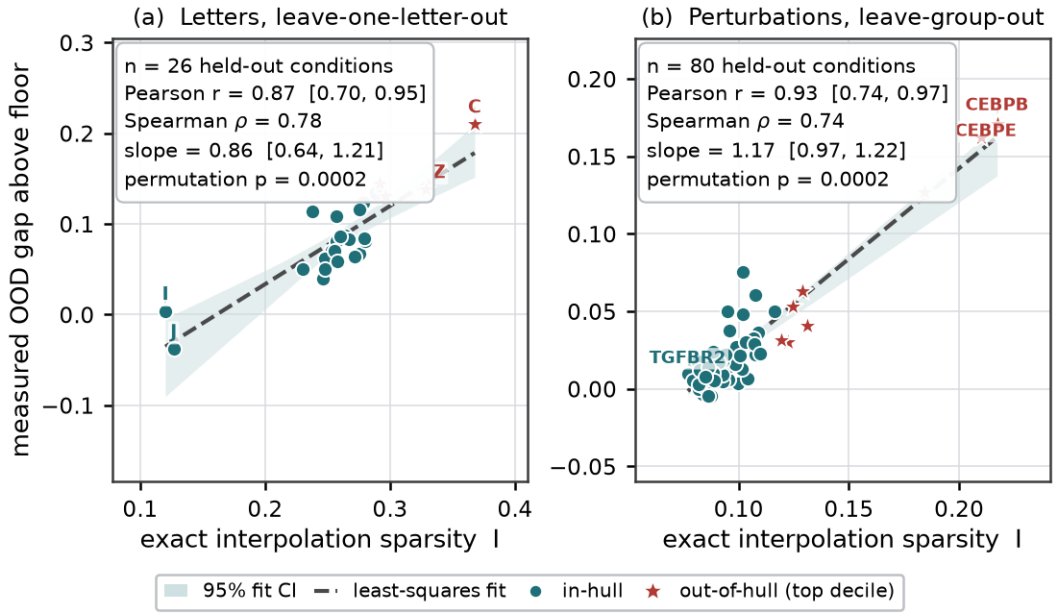

Figure 3: TERM I predicts the OOD error in both domains. Each condition is held entirely out of training; x is the exactly computed TERM I, y the measured OOD transport cost above the sampling-noise floor. Shaded band: 95% bootstrap CI of the fit.

4.5 THE ZERO-/FEW-SHOT BOUNDARY

Proposition 4 predicts that inference-time reweighting cannot beat the TERM I* floor and only out-of-hull data can. We test this directly with three interventions: *classifier reweighting*, replacing the distance-softmax weights with a learned $p(c \mid x)$; *condition mixup*, a flatter weighting of neighbors; and *isometry regularization*. None reduces the gap. On letters the mean Δgap over extrapolation glyphs is $+0.012$ (classifier), $+0.034$ (mixup), -0.007 (isometry); on perturbations, -0.045, $+0.004$, -0.000 in 2D and -0.007, $+0.004$, -0.003 in 8D-all inside a no-meaningful-change band of $|\Delta| < 0.05$ (Fig. 5a,b). Learned classifier reweighting closes a modest part of the gap in 2D-it finds barycentric weights closer to TERM I* than the distance-softmax weights do-but never eliminates it: the three hardest perturbations remain ≥ 0.08 above the floor, as Prop. 4 requires.

One out-of-hull example collapses the gap. The only intervention that works injects real target samples: parameter-efficient fine-tuning of the field's last layer (and a fresh target embedding) on k examples of $c^\star$. At $k = 0$ the gap is the TERM I floor; a *single* example collapses it for every condition (C: 0.197 $\rightarrow$ 0.084; L: 0.206 $\rightarrow$ -0.027; CEBPB: 0.173 $\rightarrow$ 0.045), and further shots add little (Fig. 5c,d). The floor is therefore a *data* boundary, not a tuning one: uncrossable by manipulating how the model uses its training conditions, crossed immediately by one out-of-hull sample.

5 LIMITATIONS

The sharp bound is conditional. Proposition 3 holds under coherent alignment, which we verify empirically rather than assume: it fails on 5 of 80 perturbations. Users of the TERM I diagnostic in dimensions above one should compute the coherence residual alongside it; the certified bound $\beta \leq$ TERM I $+ \mathrm{I}_2$ is what holds unconditionally, and it is looser.

Further limitations. See App. G for the metric, dimensionality, and protocol limits.

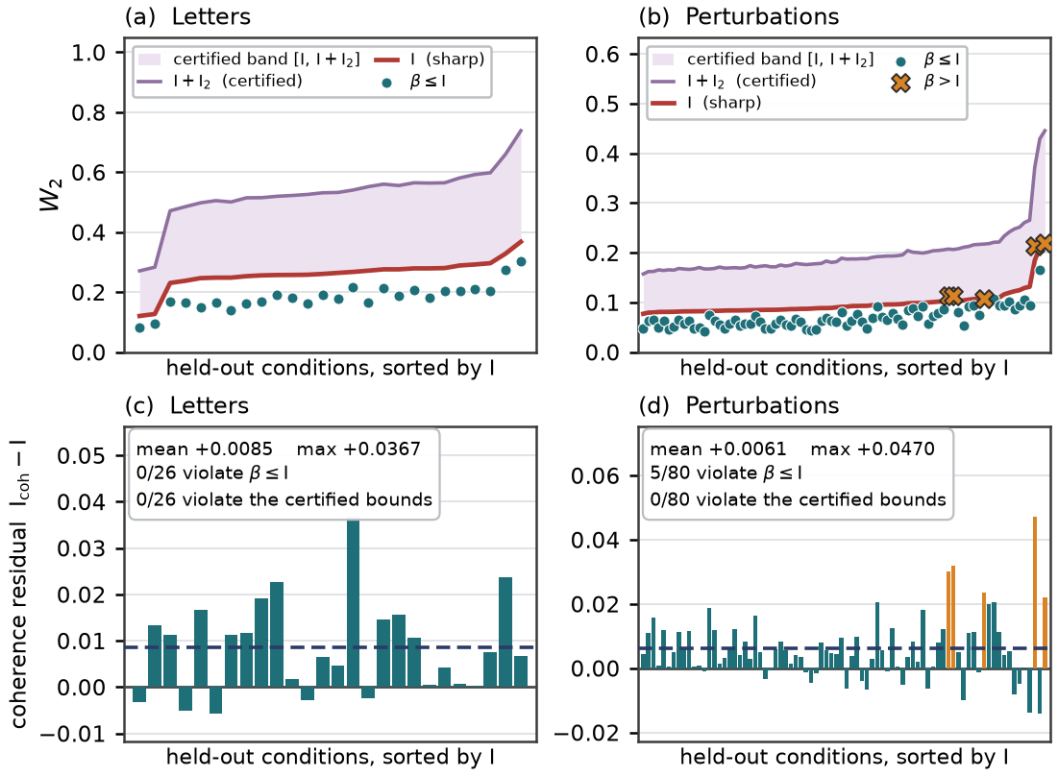

Figure 4: Certified versus sharp bounds. (a,b) Measured barycentric cost β against the sharp bound TERM I and the certified band [TERM I, TERM I + I_2]; crosses mark the conditions where the unproved sharp bound fails. **(c,d)** The coherence residual I_{coh} − TERM I that Prop. 3 requires to vanish; the failures are exactly its large values.

Table 1: Headline results. TERM II= δ; TERM III= in-distribution gap above floor; fidelity = $W_2(\text{gen}, B_{\text{exact}})$. All correlations are between the exactly computed TERM I and the measured OOD gap, under leakage-free leave-one-condition-out. Full per-condition tables: App. J.

	Letters (2D)	Perturbations (2D)	Perturbations (8D)
held-out conditions	26	80	80
sampling floor Φ_m	0.127	0.063	0.175
TERM II (δ)	0.004	0.024	0.038
TERM III (gap above floor)	+0.011	+0.003	+0.018
fidelity (mean)	0.129	0.050	0.151
Pearson r	0.87 [0.70, 0.95]	0.93 [0.74, 0.97]	0.95 [0.91, 0.97]
Spearman ρ	0.78	0.74	0.93
r, top-3 leverage removed	0.70	0.73	0.91
Theil-Sen slope	0.99	0.98	1.04
permutation p	< 0.001	< 0.001	< 0.001
$\beta \leq$ TERM I + I_2 (certified)	26/26	80/80	-
$\beta \leq$ TERM I (sharp)	26/26	75/80	-

6 CONCLUSION

Out-of-distribution generation in conditional flow matching is not a monolithic failure to be tuned away. On exactly measurable benchmarks, a single geometric term predicts OOD error ($r = 0.87$ on letters, $r = 0.93$ on perturbations, $r = 0.95$ in 8D; fidelity-verified), is provably unreachable by inference-time reweighting, and collapses after one out-of-hull example. Formalizing before measurement revealed that the sharpest central inequality needs a hypothesis omitted by its pub-

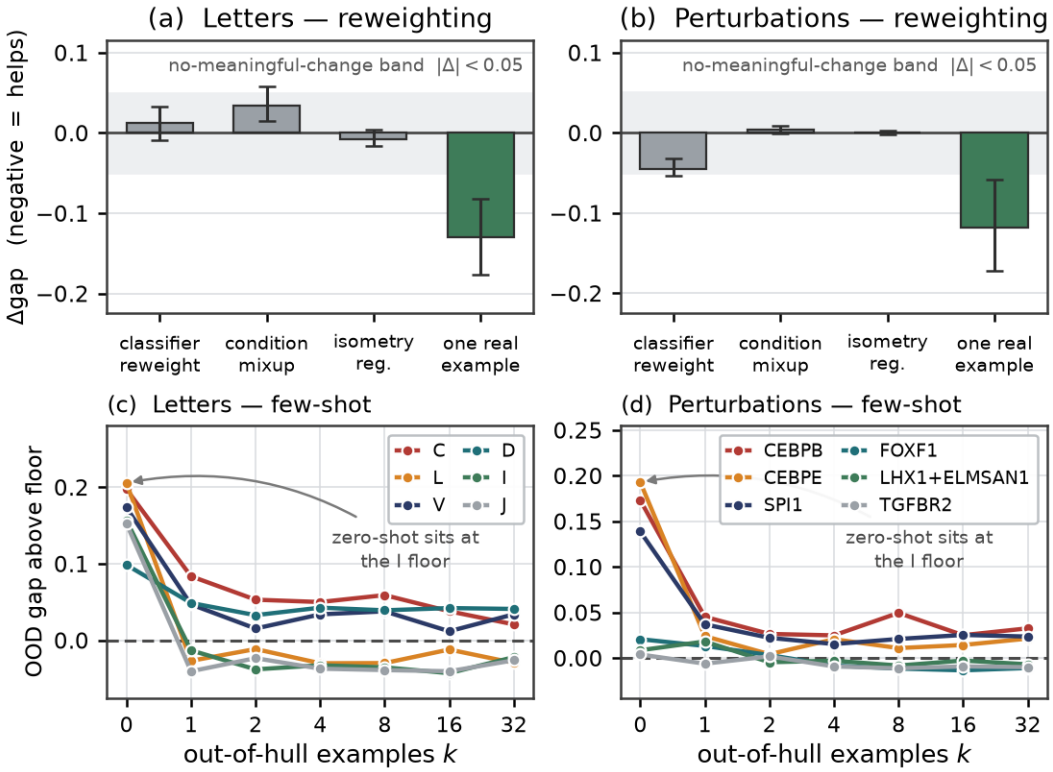

Figure 5: The zero-/few-shot boundary. (a,b) Mean Δgap on extrapolation conditions with 95% bootstrap CIs; the shaded strip is the no-meaningful-change band $|\Delta| < 0.05$. Classifier reweighting, condition mixup and isometry regularization all land inside it: no inference-time intervention escapes the hull. **(c,d)** The same gap against the number k of real out-of-hull examples. Zero-shot sits at the TERM I floor, a single example collapses it, and further shots add little-the boundary is a property of the data, not of the tuning.

lished proof; empirically, the inequality fails exactly when that hypothesis does. The boundary is geometric, measurable in advance, and crossed only by data.

7 IMPACT STATEMENT

This paper is a measurement study of when conditional generative models can and cannot extrapolate. Its main practical effect, if adopted, is to *prevent* work: a practitioner who computes TERM I before training learns whether a requested condition is reachable, and can decline to ship a model whose output for that condition would be an interpolation of things it has seen rather than a prediction of the thing it was asked for. We consider this the socially useful direction. In scientific domains such as the perturbation screens we study, generative models are increasingly used to prioritize expensive wet-lab experiments; a confidently wrong sample for an unseen perturbation can misdirect real resources, and a *pre-training* test that flags such targets as out-of-hull is a cheap safeguard. We stress the converse risk as well: our diagnostic is a statement about geometry, not about biology, and a small TERM I certifies reachability under the model's own metric, not the biological validity of the generated cells. It should inform, not replace, experimental validation.

The models are small (a five-layer MLP) and every experiment in this paper runs on a single GPU in hours, so the compute footprint is negligible relative to contemporary generative-model research; the two analyses that carry the paper's new claim-the coherence check and the geometry oracle-need no training at all. We release all code, the Lean development, and the raw result files (App. K).

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. doi: 10.1137/100805741.

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations (ICLR)*, 2023.

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. doi: 10.1016/j.jmaa.2016.04.045.

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (GANs). In *International Conference on Machine Learning (ICML)*, volume 70, pp. 224–232, 2017.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

Marco Cuturi and Arnaud Doucet. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning (ICML)*, volume 32, pp. 685–693, 2014.

Leonardo de Moura and Sebastian Ullrich. The Lean 4 theorem prover and programming language. In *International Conference on Automated Deduction (CADE)*, pp. 625–635. Springer, 2021. doi: 10.1007/978-3-030-79876-5_37.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Young-Heon Kim and Brendan Pass. Wasserstein barycenters over Riemannian manifolds. *Advances in Mathematics*, 307:640–683, 2017. doi: 10.1016/j.aim.2016.11.026.

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 (NeurIPS)*, 2019.

Thibaut Le Gouic, Quentin Paris, Philippe Rigollet, and Austin J. Stromme. Fast convergence of empirical barycenters in Alexandrov spaces and the Wasserstein space. *Journal of the European Mathematical Society*, 25(6):2229–2250, 2023. doi: 10.4171/JEMS/1234.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning (ICML)*, volume 97, pp. 3744–3753, 2019.

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.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):e11517, 2023. doi: 10.15252/msb.202211517.

Shitong Luo and Wei Hu. Diffusion probabilistic models for 3D point cloud generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2837–2845, 2021.

Thomas M Norman, Max A Horlbeck, Joseph M Replogle, Alex Y Ge, Albert Xu, Marco Jost, Luke A Gilbert, and Jonathan S Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019. doi: 10.1126/science.aax4438.

Stefan Peidli, Tessa D. Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J. Schumacher, Jake P. Taylor-King, Debora S. Marks, Augustin Luna, Nils Blüthgen, and Chris Sander. scPerturb: harmonized single-cell perturbation data. *Nature Methods*, 21(3):531–540, 2024. doi: 10.1038/s41592-023-02144-y.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019. doi: 10.1561/2200000073.

Long HC Pham. Decomposing out-of-distribution error in conditional flow matching via wasserstein geometry. In *Forty-third International Conference on Machine Learning*, 2026.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Karl-Theodor Sturm. Probability measures on metric spaces of nonpositive curvature. *Contemporary Mathematics*, 338:357–390, 2003. doi: 10.1090/conm/338/06080.

The mathlib Community. The Lean mathematical library. In *ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP)*, pp. 367–381, 2020. doi: 10.1145/3372885.3373824.

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.

Cédric Villani. *Optimal Transport: Old and New*. Springer, 2009. doi: 10.1007/978-3-540-71050-9.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022. doi: 10.1109/TPAMI.2022.3195549.

APPENDIX CONTENTS

App. A is the extended related work moved out of Sec. 2. App. B is the machine-checked development and the statement-by-statement correspondence with the main text. App. C gives the full proofs. App. D is the extended discussion of coherent alignment, including why the sharp bound is automatic in one dimension and what curvature has to do with it. App. E is the experimental setup, hyperparameters, per-condition floors, and infrastructure. App. F is the single-cell preprocessing and the 8D replication. App. G is the full argument for exactly-measurable benchmarks that Sec. 5 compresses. App. H is the exact-versus-entropic coupling comparison. App. I is the full robustness battery: leverage, jackknife, temperature, seeds and step budgets. App. J holds the full per-condition tables. App. K is the reproduction guide and the contents of the supplementary material.

A EXTENDED RELATED WORK

Flow matching and interpolant models. Flow matching learns a continuous-time velocity field transporting a source to a target density by regressing onto conditional vector fields (Lipman et al., 2023), and is closely related to rectified flow (Liu et al., 2023), stochastic interpolants (Albergo & Vanden-Eijnden, 2023), and score-based diffusion through SDEs (Song et al., 2021). Minibatch optimal-transport couplings straighten trajectories and reduce variance (Tong et al., 2024). We use exact Hungarian coupling rather than entropic (Sinkhorn) coupling (Cuturi, 2013); as Lemma 1 shows, entropic barycentric projection blurs the flow-matching target and would contaminate per-term measurement.

Conditional control and guidance. When a target condition is hard, the standard levers are guidance and reweighting: classifier guidance (Dhariwal & Nichol, 2021) and classifier-free guidance (Ho & Salimans, 2022) steer samples toward a condition at inference time. Our battery evaluates the flow-matching analogues-classifier reweighting and condition mixup-and finds that, for targets outside the training hull, they cannot reduce the error, because (Prop. 4) they only re-mix conditions already inside the hull.

OOD generalization and its limits. Domain generalization seeks models robust to distribution shift (Zhou et al., 2022), and generalization bounds for generative models typically bound average-case error (Arora et al., 2017). We instead *localize* which targets are reachable: the three-term bound we measure is due to Pham (2026), who prove conditional generation error is controlled by interpolation sparsity, an embedding-distortion constant, and the in-distribution error, validated on images through VAE-latent and Jacobian proxies. Our contribution is complementary: benchmarks on which the same terms are measured *exactly*, together with a formal-methods pass that fixes what the proxies could not see (App. C.1, App. B).

Optimal transport, barycenters, and single-cell generation. Our analysis is built on Wasserstein geometry (Villani, 2009; Peyré & Cuturi, 2019); the idealized OOD output of a conditional model is a Wasserstein barycenter (Agueh & Carlier, 2011), computed exactly here by free-support fixed-point iteration (Cuturi & Doucet, 2014; Álvarez-Esteban et al., 2016). The subtlety our Lean pass exposes is a known one in this literature: W_2 over $\mathbb{R}^d$ has non-negative curvature, so the variance inequalities that make barycenters well behaved in non-positively curved spaces (Sturm, 2003) require extra regularity here (Kim & Pass, 2017; Le Gouic et al., 2023); App. D states precisely which step this affects. We represent conditions as point clouds with a per-point velocity field; set-based generators (Lee et al., 2019) and point-cloud diffusion (Luo & Hu, 2021) are the natural architecture class, but our claims concern the geometry of the target, not the backbone. Conditional generative models of single-cell perturbation responses are an active area (Lotfollahi et al., 2023); we use the standardized Norman-2019 Perturb-seq dataset (Norman et al., 2019; Peidli et al., 2024). To cross the zero-/few-shot boundary we use parameter-efficient fine-tuning in the spirit of low-rank adaptation (Hu et al., 2022); for evaluation we use exact W_2 calibrated by a sampling-noise floor, in the spirit of precision/recall metrics (Kynkäänniemi et al., 2019) but exploiting exact tractability.

B THE MACHINE-CHECKED DEVELOPMENT

Every mathematical statement in Sec. 3 is formalized in Lean 4 (de Moura & Ullrich, 2021) against Mathlib (The mathlib Community, 2020) and compiles with no `sorry`. We checked each result with `#print axioms`: all of them depend only on `propext`, `Classical.choice` and `Quot.sound`, Lean's three standard axioms, and on nothing else.

What is abstracted, and why it is faithful. Two modelling choices keep the formal statements identical to the ones the experiments use rather than weaker relatives of them. *(i)* Wasserstein space appears only as a `PseudoMetricSpace`. Every argument in Sec. 3 that lives in W_2-space uses only the triangle inequality, symmetry and non-negativity, so proving it for an arbitrary pseudometric space proves it for W_2 - and makes explicit that no curvature, geodesic or absolute-continuity property is being smuggled in. *(ii)* Point clouds of a fixed size N live in a real normed space E (concretely $\mathbb{R}^{N\times d}$ with the normalized Frobenius norm), and the point-cloud distance is any function D with $D(x, y) \le \|x - y\|$ - which is exactly what a minimum-over-assignments distance satisfies. This is the level at which the free-support barycenter is genuinely a convex combination, and it is the level at which the alignment question becomes visible.

Table 2: Correspondence between the paper's statements and the Lean development (`lean/Leanverify/Extrapolate.lean` in the supplementary material).

paper	Lean name	what is proved
Thm. 1	`three_term_decomposition`	the budget split, from the triangle inequality
-	`weighted_dist_triangle`	weighted triangle inequality over the simplex
-	`sq_weighted_sum_le`	weighted Cauchy-Schwarz $(\sum w_i a_i)^2 \le \sum w_i a_i^2$
-	`weighted_sum_le_sqrt`	TERM I $\le$ I$_2$
-	`weighted_variance_identity`	the exact variance identity behind strictness
-	`sq_weighted_sum_lt`	strict form of the above
Prop. 1	`floor_of_law_eq`	equal laws give equal expected cost
Prop. 1	`floor_pos`, `floor_nonneg`	$\Phi_m > 0$ at finite N
Prop. 2	`beta_le_aligned`	$\beta \le$ I$_{\text{coh}}$
Prop. 2	`TI_le_aligned`	TERM I $\le$ I$_{\text{coh}}$ (the gap in the published argument)
Prop. 2	`barycentric_cost_le`	$\beta \le$ TERM I + I$_2$, unconditionally
Prop. 2	`TI_le_TItwo`	the two budgets are ordered
Prop. 3	`beta_le_TI_of_coherent`	$\beta \le$ TERM I under coherent alignment
-	`norm_convex_combination_sub_le`	Minkowski for mixtures
Prop. 4	`no_reweighting_beats_TIstar`	TERM I* lower-bounds every reweighting
Prop. 4	`below_TIstar_not_reachable`	beating TERM I* requires leaving the hull
Lem. 1	`barycentric_projection_contracts`	doubly stochastic projection contracts dispersion
Lem. 1	`projection_strictly_contracts`	strictness, with its exact hypothesis
Lem. 1	`permutation_preserves_dispersion`	exact matching preserves dispersion
Sec. 3.3	`Weights.softmax`	the softmax weights lie in Δ^{n-1}

What the formalization changed. Three things. *(1)* The published justification of $\beta \le$ TERM I - "the inequality is the triangle inequality" - does not go through. `beta_le_aligned` and `TI_le_aligned` together show the construction bounds β by the *larger* budget I$_{\text{coh}}$, so the argument proves an inequality on the wrong side of TERM I. We replaced it with two statements: an unconditional one (`barycentric_cost_le`) and a conditional sharp one (`beta_le_TI_of_coherent`). *(2)* Lemma 1 originally claimed "strict contraction for every $\varepsilon > 0$" without saying strict in what. `projection_strictly_contracts` supplies the exact hypothesis: a row of the coupling must place positive mass on two targets at *different* distances from the reference. Full-support entropic couplings satisfy this except in the degenerate case where every target is equidistant. *(3)* Proposition 4 was stated with a minimum over the simplex; the reachable set need not be compact

in general, so the formal version uses an infimum and proves the contrapositive we actually use (`below_TIstar_not_reachable`).

C Detailed Proofs

C.1 Proof of the decomposition (self-contained restatement)

Proof of Theorem 1. The proof follows the structure of Pham (2026); we reproduce it for self-containedness. Let $\hat{p}_{c^\star}$ be the model output for target $c^\star$ and $\mu_{c^\star}$ the true measure. Add and subtract the barycentric mixture $\sum_i w_i \mu_{c_i}$:

$$W_2(\hat{p}_{c^\star}, \mu_{c^\star}) \le W_2\Big(\hat{p}_{c^\star}, \sum_i w_i \mu_{c_i}\Big) + W_2\Big(\sum_i w_i \mu_{c_i}, \mu_{c^\star}\Big). \tag{2}$$

The second term is at most $\sum_i w_i W_2(\mu_{c_i}, \mu_{c^\star}) = \text{TERM I}$ by convexity of W_2 in its arguments (with the K factor absorbing the choice of barycentric coupling). The first term is bounded by the conditioning error: feeding $e_{c^\star} = \sum_i w_i e_{c_i}$ to a field with embedding distortion δ and in-distribution error $\varepsilon_{\text{train}}$ introduces at most the distortion δ (the embedding moves the interpolation point) plus $C\varepsilon_{\text{train}}$ (the field's per-condition error at the interpolated condition, propagated through the sampler Lipschitz constant C). Combining the two inequalities gives Eq. (1). The constants K, C are model-class constants we do not attempt to estimate; our experiments test the structural prediction that total cost tracks TERM I once TERM II/TERM III are small. The budget-splitting step is `three_term_decomposition` and the convexity step is `weighted_dist_triangle`. □

C.2 Proof of Proposition 1

Proof. Let the generator's law be $\nu = \mu$, the target law. Its output G_N and the fresh evaluation sample $\hat{\mu}_N$ are then two independent N-sample empirical measures of μ, so $\mathbb{E}\, W_2(G_N, \hat{\mu}_N) = \mathbb{E}\, W_2(\hat{\mu}_N, \hat{\mu}'_N) = \Phi_m$. The formal version (`floor_of_law_eq`) isolates the only content of this step-that equal pushforward laws give equal expectations of every integrable functional-via $\int F \circ G \, d\mu = \int F \, d(\mu \circ G^{-1})$. Positivity for finite N holds because two independent continuous samples are almost surely distinct, giving strictly positive assignment cost (`floor_pos`). The rate $\Phi_m = \Theta(N^{-1/d})$ for $d \ge 3$ (with a logarithmic factor for $d = 2$) is the standard rate of convergence of empirical measures in W_2 (Peyré & Cuturi, 2019). Thus a perfect generator incurs exactly Φ_m, and no smaller cost is achievable in expectation by matching the law. □

C.3 Proof of Proposition 2

Proof. Coherent bound. At a fixed point of the free-support iteration (Cuturi & Doucet, 2014; Álvarez-Esteban et al., 2016) the barycenter of equal-size uniform clouds satisfies $B_w = \sum_i w_i \Pi_i(\mu_{c_i})$, where Π_i re-indexes μ_{c_i} by the assignment optimal against B_w. Writing $\|\cdot\|$ for the normalized Frobenius norm on $\mathbb{R}^{N\times d}$ and using $\sum_i w_i = 1$,

$$B_w - \mu_{c^\star} = \sum_i w_i \big(\Pi_i(\mu_{c_i}) - \mu_{c^\star}\big), \qquad \text{so} \qquad \|B_w - \mu_{c^\star}\| \le \sum_i w_i \big\|\Pi_i(\mu_{c_i}) - \mu_{c^\star}\big\| = \mathrm{I}_{\text{coh}}$$

by Minkowski (`norm_convex_combination_sub_le`); since W_2 minimizes over assignments, $\beta = W_2(B_w, \mu_{c^\star}) \le \|B_w - \mu_{c^\star}\|$, giving $\beta \le \mathrm{I}_{\text{coh}}$ (`beta_le_aligned`). The same minimality gives $W_2(\mu_{c_i}, \mu_{c^\star}) \le \|\Pi_i(\mu_{c_i}) - \mu_{c^\star}\|$ term by term, whence TERM I $\le \mathrm{I}_{\text{coh}}$ (`TI_le_aligned`). This is the entire content of the construction, and it is an upper bound against I_{coh}, not against TERM I.

Quadratic bound. Let $V(\nu) = \sum_i w_i W_2^2(\nu, \mu_{c_i})$ be the barycentric functional. Since B_w minimizes V, $V(B_w) \le V(\mu_{c^\star}) = \mathrm{I}_2^2$. The weighted triangle inequality (`weighted_dist_triangle`) gives $W_2(B_w, \mu_{c^\star}) \le \sum_i w_i W_2(B_w, \mu_{c_i}) + \sum_i w_i W_2(\mu_{c_i}, \mu_{c^\star})$, and weighted Cauchy-Schwarz (`sq_weighted_sum_le`) bounds the first sum by $\sqrt{V(B_w)} \le \mathrm{I}_2$. Hence $\beta \le$ TERM I $+ \mathrm{I}_2$ (`barycentric_cost_le`). Finally TERM I $\le \mathrm{I}_2$ is Jensen (`TI_le_TItwo`). Both bounds are computable from the training geometry alone. □

C.4 Proof of Proposition 3

Proof. Under coherent alignment each term of $\mathrm{I}_{\mathrm{coh}}$ equals the corresponding term of Term I, so $\mathrm{I}_{\mathrm{coh}} = \text{Term I}$ and the coherent bound of Proposition 2 reads $\beta \leq \text{Term I}$ (`beta_le_TI_of_coherent`). □

C.5 Proof of Proposition 4

Proof. Reweighting produces a new weight vector $w' \in \Delta^{n-1}$ and hence output $B_{w'}$. The set of reachable outputs is therefore $\mathcal{B} = \{B_{w'}\}$, and $\text{Term I}^\star = \inf_{w'} W_2(B_{w'}, \mu_{c^\star})$ is by definition a lower bound on the cost of every element of it (`no_reweighting_beats_TIstar`, an instance of inf $\leq$). Contrapositively, any output y with $W_2(y, \mu_{c^\star}) < \text{Term I}^\star$ is not in $\mathcal{B}$ (`below_TIstar_not_reachable`); since every $B_{w'}$ lies in the geodesically convex hull of $\{\mu_{c_i}\}$, producing such a y requires a new extreme measure $\mu_{\text{new}} \notin \{\mu_{c_i}\}$, i.e. out-of-hull data. □

C.6 Proof of Lemma 1

Proof. The barycentric projection $\bar{y}_i = \sum_j (NP_{ij}) y_j$ is the conditional expectation $\mathbb{E}[y_J \mid I = i]$ under the coupling P (row i normalized to a probability vector $p^{(i)}$). Two steps, both formalized. *(a)* Minkowski gives $\|\bar{y}_i - m\| \leq \sum_j p_j^{(i)} \|y_j - m\|$, and weighted Cauchy-Schwarz gives $(\sum_j p_j^{(i)} \|y_j - m\|)^2 \leq \sum_j p_j^{(i)} \|y_j - m\|^2$; together, $\|\bar{y}_i - m\|^2 \leq \sum_j p_j^{(i)} \|y_j - m\|^2$. *(b)* Summing over i and using that P is doubly stochastic ($\sum_i NP_{ij} = 1$) collapses the right-hand side to $\sum_j \|y_j - m\|^2$ (`barycentric_projection_contracts`). Strictness comes from the exact weighted-variance identity $\sum_j p_j a_j^2 - (\sum_j p_j a_j)^2 = \sum_j p_j (a_j - \bar{a})^2$ (`weighted_variance_identity`): the contraction of row i is strict as soon as some $a_{j_0} = \|y_{j_0} - m\|$ with $p_{j_0}^{(i)} > 0$ differs from the row mean (`projection_strictly_contracts`). For $\varepsilon > 0$ the entropic coupling has full support, so every row is strictly contracted unless all targets are equidistant from m. When P is a permutation matrix-the $\varepsilon \to 0$ limit-$\bar{y}_i = y_{\sigma(i)}$ and the dispersion is preserved exactly (`permutation_preserves_dispersion`). Thus exact coupling yields a sharp regression target while entropic coupling blurs it. □

D Coherent Alignment: Why the Sharp Bound Is Conditional

One dimension is special. For $d = 1$ the map $\mu \mapsto F_\mu^{-1}$ sending a measure to its quantile function is an isometry from $(\mathcal{P}_2(\mathbb{R}), W_2)$ onto a convex subset of $L^2([0,1])$. Barycenters are therefore literal L^2 averages of quantile functions, all optimal couplings share the common sort order, and coherent alignment holds automatically. In that case $\beta = \|\sum_i w_i q_i - q_{c^\star}\|_2 \leq \sum_i w_i \|q_i - q_{c^\star}\|_2 = \text{Term I}$ is exactly the triangle inequality in a Hilbert space-which is presumably the intuition behind the published one-line proof. It does not survive to $d \geq 2$.

What goes wrong in higher dimension. For $d \geq 2$, $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ is an Alexandrov space of *non-negative* curvature. The variance inequality $W_2^2(B_w, \nu) + \sum_i w_i W_2^2(B_w, \mu_{c_i}) \leq \sum_i w_i W_2^2(\mu_{c_i}, \nu)$ that would make barycenters behave like Hilbert-space means is a property of *non-positively* curved spaces (Sturm, 2003); in W_2 over $\mathbb{R}^d$ it holds only under regularity conditions on the barycenter and the measures (Kim & Pass, 2017; Le Gouic et al., 2023). Concretely, the obstruction is that the assignments Π_i chosen by the barycenter iteration and the assignments optimal against the target need not agree; the discrepancy is exactly the coherence residual $\mathrm{I}_{\mathrm{coh}} - \text{Term I} \geq 0$ we measure. The certified bound of Proposition 2 avoids the issue by never assuming they agree.

What we measure. For every held-out condition we compute (i) the barycentric assignments Π_i at the fixed point, (ii) the assignment optimal against the target, (iii) the residual $\mathrm{I}_{\mathrm{coh}} - \text{Term I}$, and (iv) whether $\beta \leq \text{Term I}$, $\beta \leq \mathrm{I}_{\mathrm{coh}}$, $\beta \leq \text{Term I} + \mathrm{I}_2$, and the minimality premise $V(B_w) \leq V(\mu_{c^\star})$ hold. Results: on letters, $0/26$ violations of the sharp bound, mean residual $+0.008$, maximum $+0.037$; on perturbations, $5/80$ violations, mean residual $+0.006$, maximum $+0.047$, largest violation $\beta - \text{Term I} = +0.029$ (condition `DUSP9_ETS2`, $0.45\times$ the floor). The certified bounds

and the minimality premise are satisfied in $106/106$ cases across both benchmarks. The script is `coherence_check.py` in the supplementary material, and Fig. 4 is its output.

E Experimental Setup and Infrastructure

The velocity field is a 5-layer per-point MLP (hidden width 256, SiLU activations) consuming a 64-d learned condition embedding and a 64-d sinusoidal time embedding, so a barycentric reference can be fed at evaluation. Training uses a centroid source ($\sigma = 0.20$), exact per-step Hungarian coupling, AdamW (lr $= 10^{-3}$, weight decay 10^{-4}), a cosine schedule over 9000 steps, batch 32, $N = 256$ points, jitter 0.02 (letters) / no jitter (single-cell); generation integrates the field with 80 Euler steps and softmax temperature $\tau = 0.06$. The isometry regularizer pulls embedding pairwise distances toward W_2/scale (scale = median pairwise W_2); δ is the RMS residual of the best linear fit $W_2 \approx a\|e_i - e_j\| + b$. All OOD numbers use strict leave-one-condition-out (26 independent trainings for letters; 6 folds for perturbations). Few-shot fine-tuning adapts only the last layer and a fresh target embedding on k examples.

Per-condition floors. Per-glyph floors Φ_m at $N = 256$, jitter 0.02 (mean over 4 draw pairs, seed 0), sorted, are `J` 0.067, `I` 0.070, `T` 0.077, `Y` 0.087, `A` 0.105, `F` 0.106, `C` 0.106, `L` 0.107, `P` 0.110, `Z` 0.112, `V` 0.113, `X` 0.120, `R` 0.125, `W` 0.132, `B` 0.137, `E` 0.137, `O` 0.143, `D` 0.144, `K` 0.145, `S` 0.152, `M` 0.159, `N` 0.159, `G` 0.160, `H` 0.163, `Q` 0.179, `U` 0.189. Fig. 6 plots these against the single-cell floors; the spread is why we report absolute W_2 with the floor as a band.

Infrastructure. The field is small and trains on a single GPU; all Wasserstein distances and free-support barycenters are computed exactly on CPU via linear-sum-assignment, which dominates evaluation cost at $O(N^3)$ per pair. Full runs use Kaggle GPU instances and a local Apple-Silicon machine; the model-free analyses (`coherence_check.py`, `geometry_extras.py`) run on CPU in minutes.

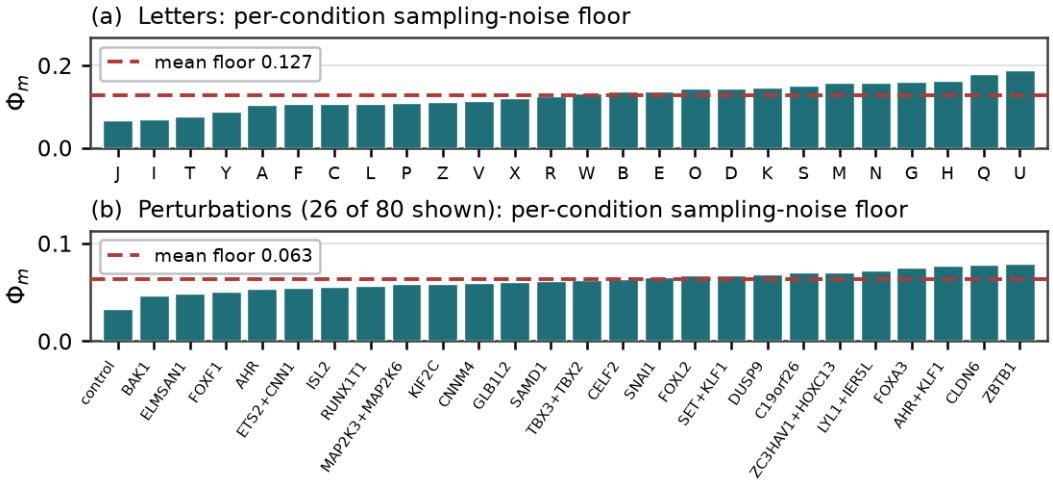

Figure 6: Per-condition sampling-noise floors. Compact conditions have lower floors than diffuse ones, which is why a single subtracted constant would misrank conditions and we report absolute W_2 with the floor as a reference band.

F Single-Cell Data Preprocessing and the 8D Replication

We load the scPerturb standardized Norman-2019 file (`NormanWeissman2019_filtered.h5ad`) (Peidli et al., 2024). We keep the perturbations with at least 200 cells and the 80 largest cohorts, normalize each cell to 10^4 total counts, log1p-transform, select the top 2000 highly variable genes, scale to unit variance, and project to PCA. Each perturbation's cloud is centered and rescaled so the source logic matches the letters benchmark; $N = 256$ cells are sampled per cloud. Control perturbations are kept as conditions.

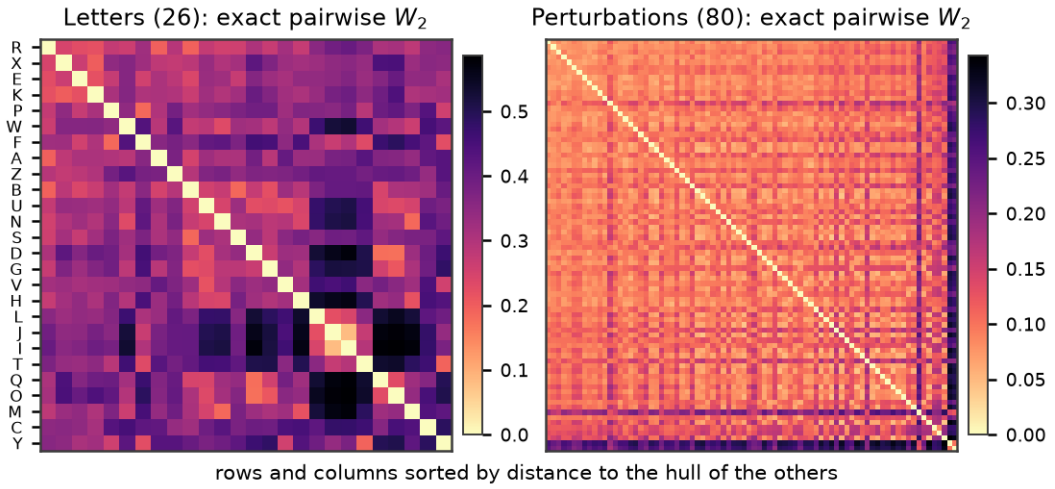

Figure 7: The exact condition geometry. Pairwise W_2 matrices, rows and columns ordered by distance to the hull of the remaining conditions. Everything the TERM I diagnostic needs is contained in these matrices, and both are computed before any training.

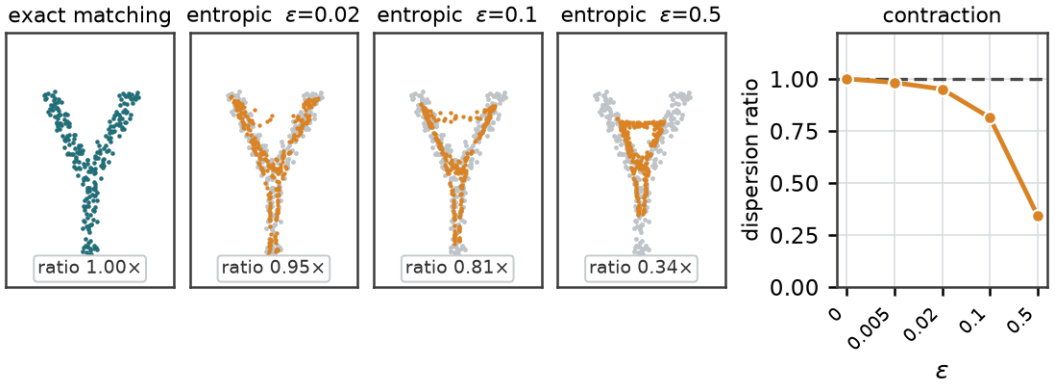

Figure 8: Lemma 1 made visible. Regression targets formed by barycentric projection of an entropic coupling, against the true target cloud (grey). Exact Hungarian matching preserves the target; increasing ε contracts it monotonically toward the conditional mean, which is why we use exact coupling throughout.

8D PCA replication. To check that the mechanism is not a 2D artifact, we rerun the entire single-cell benchmark with the same pipeline but an 8D PCA embedding (exact Hungarian W_2 and exact barycenters are dimension-agnostic). The result strengthens rather than weakens: Pearson $r = 0.95$ $[0.91, 0.97]$, Spearman $\rho = 0.93$, Theil-Sen slope 1.04, permutation $p < 0.001$ (floor 0.175, fidelity 0.151, in-distribution gap $+0.018$). The correlation is also more robust than in 2D: removing the three highest-leverage perturbations leaves $r = 0.91$, and the jackknife r stays in $[0.94, 0.96]$. The intervention battery again shows no inference-time reweighting effect (classifier -0.007, mixup $+0.004$, isometry -0.003; all $|\Delta| < 0.05$), while one out-of-hull example collapses the gap (CEBPB: $0.207 \to 0.075$).

G WHY EXACTLY-MEASURABLE BENCHMARKS

A natural objection is that the claims rest on two low-dimensional benchmarks. We argue the choice is the point. This paper contributes a *measurement instrument*, not an application: the object under test is a law (Eq. (1), Propositions 1-4) that is domain-agnostic, so the question is whether its *mechanism* holds, and that is what transfers across domains. The prior analysis (Pham, 2026) establishes the bound's proxy form on MNIST, CIFAR and CelebA-breadth with approximate terms; what only exactly-measurable domains can add is *depth*, a direct test of tightness and irreducibility. Images

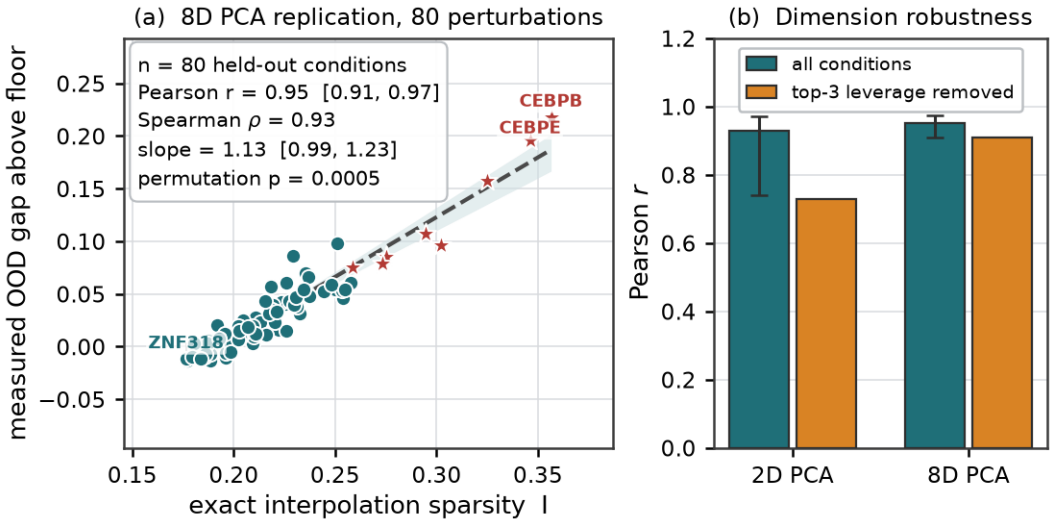

Figure 9: 8D PCA replication. **(a)** The TERM I prediction holds when the single-cell clouds live in an 8D PCA embedding rather than 2D. **(b)** It also becomes more robust: the correlation survives deletion of the three highest-leverage perturbations almost unchanged.

cannot supply it: one cannot compute an exact W_2 to a convex hull in pixel space, nor read off δ without a proxy. Our two benchmarks deliberately trade realism for exactness: glyphs give controlled topology, single-cell perturbation clouds give real transcriptomic measurements; both admit exact Hungarian W_2 and exact barycenters, and both exhibit the same mechanism. A benchmark is only "weak" if it cannot discriminate the hypothesis; here the hypothesis is sharply discriminable-it is even falsified, in a small and diagnosable way, by the coherence check of App. D-and the result replicates across two different data-generating processes. We have not tested transformer-based set generators or larger fields; our conclusions concern the geometry of the target rather than any particular backbone. The 8D replication shows the mechanism strengthens rather than weakens with dimension ($r = 0.95$), but high-dimensional and non-Euclidean condition spaces remain future work.

W_2 on point clouds is mildly lenient toward smooth distributions. The barycentric reference is smoother than a real sample, so one compact glyph (J) attains a cost slightly below even its own floor; we therefore report absolute W_2 with the floor as a band rather than subtracting a single constant.

Scope of the protocol. Placing the target via its few-shot W_2 yields the best-case reference; a practitioner without target samples must estimate TERM I from side information, which we do not study. Exact W_2 costs $O(N^3)$ per pair, which bounds the cloud sizes we can evaluate.

H EXACT VERSUS ENTROPIC COUPLING

Consistent with Lemma 1, exact Hungarian coupling reaches an in-distribution gap of $+0.048$ above floor while Sinkhorn barycentric coupling reaches $+0.108$-*despite* Sinkhorn attaining the lower training loss (0.008 vs. 0.017). Lower train loss, worse generation: the entropic coupling gives the regressor an easier but blurred target, and the field learns to produce the blur. For a glyph with a hole the conditional average can land inside the hole, so the contraction is not merely a loss of sharpness but a topology error. Figure 8 shows the effect directly on a held-out glyph, together with the monotone dispersion ratio that Lemma 1 predicts.

I ROBUSTNESS

Leverage and rank. The correlation is not driven by a few extreme conditions. Removing the three highest-leverage conditions leaves $r = 0.70$ (letters 2D), $r = 0.73$ (perturbations 2D) and $r = 0.91$ (perturbations 8D); the Theil-Sen robust slopes are 0.99, 0.98 and 1.04; the leave-one-out (jackknife) r stays in $[0.83, 0.90]$, $[0.90, 0.94]$ and $[0.94, 0.96]$ respectively (Fig. 10). Restricting to the hull interior weakens the letters correlation, as the theory predicts-the separation is sharpest exactly where extrapolation begins: dropping the top 15% of TERM I leaves $r = 0.81$ over the remaining 22 glyphs, and additionally dropping the two most compact glyphs (`I`,`J`), which anchor the low end, leaves $r = 0.52$ over 20.

Temperature. The barycentric weights depend on a softmax temperature τ. Sweeping τ over an order of magnitude (0.02 to 0.20) leaves the model-free oracle's correlation between TERM I and β in $[0.84, 0.92]$ on letters and $[0.78, 0.93]$ on perturbations (Fig. 11), so the result is not a consequence of one temperature choice. Both sweeps are computed by `geometry_extras.py` without training anything.

The oracle, seeds, folds and step budgets. A model-free oracle that simply emits the exact Wasserstein barycenter-no network, no training-reproduces the relationship: its barycentric cost β predicts the trained field's measured OOD gap at $r = 0.89$ across the 26 glyphs, and tracks the computed TERM I at $r = 0.87$. In other words the trained field adds a small constant integration overhead on top of a purely geometric quantity. Two further seeds with leave-group-out partitions give pooled $r = 0.79$ $[0.68, 0.87]$, $p < 0.001$, $n = 52$. A full leave-one-letter-out replication at 6000 steps rather than 9000 gives $r = 0.82$ $[0.67, 0.91]$, $p < 0.001$.

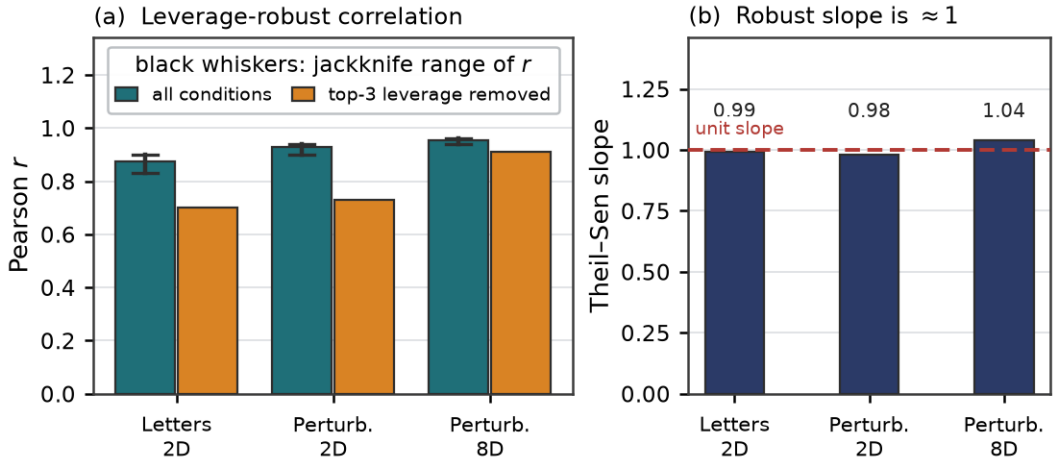

Figure 10: The correlation is not a leverage artifact. **(a)** Pearson r with and without the three highest-leverage conditions; black whiskers give the jackknife range. **(b)** Theil-Sen robust slopes are close to unity in all three settings, i.e. the measured gap tracks TERM I one-for-one.

J FULL PER-CONDITION TABLES

The full per-condition tables for both single-cell benchmarks (80 perturbations each, in 2D and 8D), the coherence tables behind App. D, and all raw result JSONs are released with the code. Every number and figure in this paper is regenerated from those JSONs without retraining.

K REPRODUCING EVERYTHING

The supplementary material contains the complete pipeline. Nothing in the paper requires retraining to reproduce: the figures and tables are regenerated from saved JSON results.

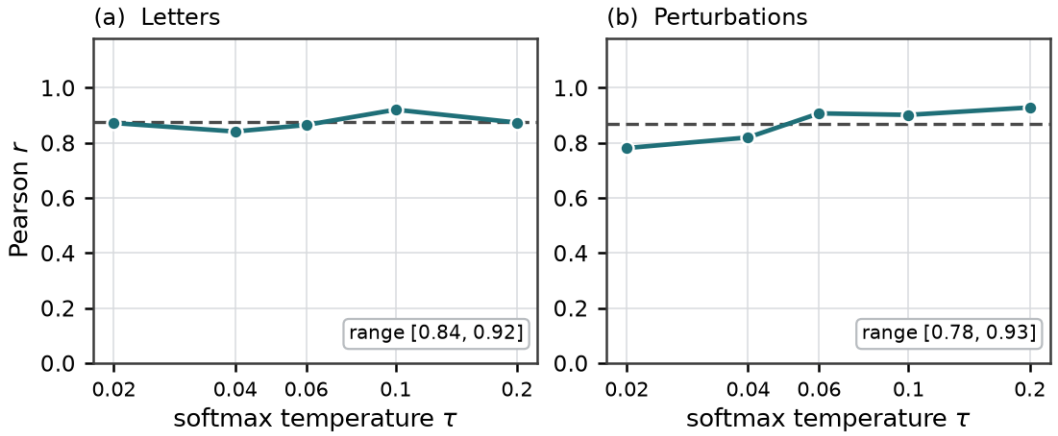

Figure 11: Temperature is not doing the work. Correlation between TERM I and the model-free barycentric cost as the softmax temperature τ varies over an order of magnitude.

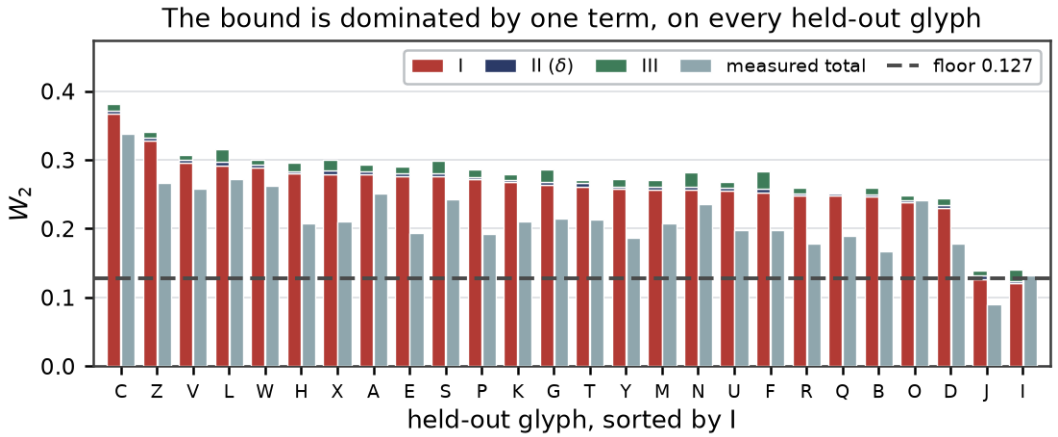

Figure 12: The decomposition, glyph by glyph. Stacked TERM I+TERM II+TERM III budget against the measured total for every held-out glyph. TERM II and TERM III are visible only as hairlines: the bound is TERM I plus rounding.

Contents.

- `code/` - the experiment runners (`letters_ood.py`, `letters_irreducibility.py`, `single_cell.py`); the shared primitives (`letters_common.py`, `stats_utils.py`); the two model-free analyses (`coherence_check.py`, `geometry_extras.py`); and the figure pipeline (`make_paper_figures.py`, `figstyle.py` and the three `fig_*.py` modules).
- `lean/` - the Lean 4 development, its `lakefile.toml` and a README mapping every paper statement to its Lean name.
- `runs/` - every raw result file behind every number in the paper.

Commands. The model-free geometry oracle and the coherence check need no GPU and no training:

```
python coherence_check.py -letters -out ../runs/coh_letters.json
python geometry_extras.py -letters -out ../runs/geom_letters.json
```

The leakage-free leave-one-out studies retrain the field once per fold:

```
python letters_ood.py     -device cuda -folds 26 -steps 9000
python single_cell.py loo -device cuda -folds  6 -steps 6000
```

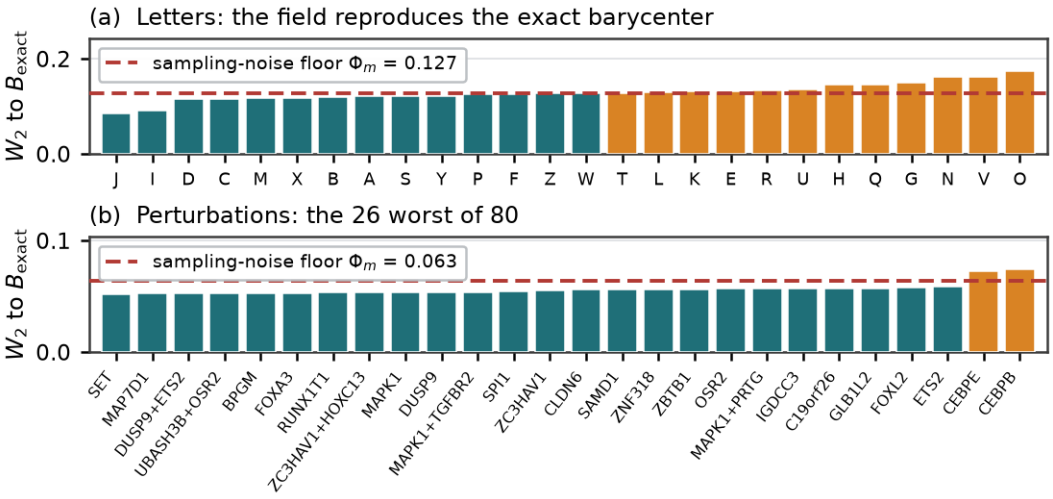

Figure 13: The field reproduces the exact barycenter. Per-condition fidelity $W_2(\text{output}, B_{\text{exact}})$ against the sampling-noise floor (dashed). Bars at or below the floor mean the output is indistinguishable from the exact barycenter given finite samples. Panel (b) shows the 26 *worst* of the 80 perturbations, so the claim holds a fortiori for the rest.

Table 3: Full per-glyph table, leave-one-letter-out, $N = 256$ (sorted by TERM I). TERM II$= \delta$; TERM III$=$ in-distribution gap above floor; total $= W_2(\text{gen}, \text{true})$; fid. $= W_2(\text{gen}, B_{\text{exact}})$. Floor $\Phi_m = 0.127$.

glyph	TERM I	TERM II	TERM III	total	fid.
C	0.368	0.004	+0.010	0.338	0.115
Z	0.328	0.003	+0.008	0.266	0.127
V	0.296	0.004	+0.008	0.258	0.162
L	0.292	0.005	+0.019	0.272	0.129
W	0.289	0.004	+0.008	0.262	0.127
H	0.280	0.003	+0.014	0.208	0.145
X	0.279	0.005	+0.016	0.211	0.118
A	0.279	0.004	+0.010	0.251	0.120
E	0.276	0.004	+0.010	0.194	0.132
S	0.276	0.004	+0.019	0.243	0.121
P	0.271	0.004	+0.012	0.191	0.124
K	0.267	0.004	+0.008	0.210	0.131
G	0.264	0.004	+0.017	0.215	0.148
T	0.260	0.005	+0.004	0.213	0.127
Y	0.258	0.003	+0.011	0.186	0.121
M	0.257	0.003	+0.010	0.208	0.117
N	0.257	0.004	+0.021	0.236	0.162
U	0.256	0.004	+0.008	0.198	0.134
F	0.253	0.005	+0.025	0.197	0.125
R	0.248	0.003	+0.009	0.178	0.133
Q	0.248	0.003	+0.000	0.189	0.146
B	0.246	0.003	+0.010	0.167	0.120
O	0.238	0.003	+0.007	0.241	0.173
D	0.230	0.003	+0.010	0.177	0.115
J	0.127	0.005	+0.008	0.090	0.084
I	0.120	0.003	+0.017	0.131	0.090
mean	0.260	0.004	+0.011	0.213	0.129

Every figure is regenerated from saved results with

```
python make_paper_figures.py -outdir ../paper/iclr2027/figs
```

and the Lean development is checked with `lake build` followed by `lake env lean CheckAxioms.lean` in `lean/`.

Exactness guarantees. W_2 between point clouds is computed exactly by the Hungarian algorithm (`scipy.optimize.linear_sum_assignment`), never by Sinkhorn, so TERM I is *computed* rather than estimated. The Wasserstein barycenter uses a free-support fixed-point iteration, and the fidelity check compares the field output to that exact barycenter. Leave-one-condition-out protocols are leakage-free: held-out clouds never enter training and their embeddings never receive a gradient.