REPDEBT: THE DEBT DIRECTION SURGERY CANNOT REPAY

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-task learning hopes similar tasks reinforce each other; interference is just as documented, and few instruments measure either effect. We built one — and it surfaces a failure mode direction surgery cannot see. SphereBench-54 is a benchmark of 54 datasets over 7 modality families, mixed so edge-modality influence stays measurable. We train 423 models under one protocol: 99 jointly on all 54, dense controls parameter-matched at 1.5M, plus 324 per-dataset specialists. We vary only the input representation of four point-cloud datasets (flat serialization, multi-view rendering, Hilbert-curve ordering); the other fifty are byte-identical. This harmless-looking choice decides the fate of plain shared trunks (one encoder, no routing). Our best such trunk swings from the best main-scan macro-accuracy among parameter-matched jointly trained models, 0.636 under one serialization, to 0.392 under another (paired Wilcoxon, $z = 6.31$). Expert-routed architectures (MMoE, PLE, Cross-Stitch) at equal or greater budgets move by at most 1.62 points; no expert cross-serialization comparison survives Bonferroni correction. This is, to our knowledge, the first controlled measurement of interference cascading through a shared encoder into byte-identical untouched tasks (the 14 vision datasets drop from 0.46 to 0.19). Gradient probes of retained checkpoints show why: a norm-dominated rank-one capture leaves the collapsed trunk's joint gradient nearly rank-one (mean pairwise task-gradient cosine +0.83 versus $\leq$0.08 elsewhere); routed and latent-bottleneck models keep per-task gradients orthogonal. Follow-up controls close the loop: the frozen readout is a minor tax (4.8 of 24.5 points), direction surgery repays nothing, per-task gradient-norm equalization repays 23.7 — the carrier is the norm-dominated rank-one update — and a vanilla dense trunk without the spherical-iterative core never incurs the debt: it is specific to shared iterative dynamics, not dense sharing. Routing buys robustness, not sample efficiency. How far this generalizes is what the benchmark is for. Code, configurations, and the evaluation protocol are released at `https://anonymous.4open.science/r/RepDebt-A449`.

1 INTRODUCTION

A jointly trained single encoder promises one set of weights, one deployment, and regularization across related tasks. The interference literature tempers that hope (Ruder, 2017; Fifty et al., 2021): gradients conflict and fusion benefits are rarely predictable. What is missing is an instrument that measures fusion against interference under controlled conditions, including one modality's influence on all others.

On SphereBench-54 (54 classification datasets, 7 modality families) we train six architectures jointly on all 54 datasets under one matched-compute protocol and vary exactly one thing: the input representation of the four point-cloud datasets. Three serializations (raw storage order, multi-view rendering, Hilbert space-filling-curve ordering (Skilling, 2004)) define otherwise identical conditions. The remaining fifty datasets are byte-identical, and model sizes are parameter-matched at 1.5M.

Under this knife, dense shared trunks do not degrade. They collapse on datasets whose data never changed. The best dense trunk has the best matched-budget joint macro-accuracy under one serialization and loses 24.5 points under another, including large drops on the fifty byte-identical datasets. Expert-routed architectures with equal or greater budgets barely move. We call the accuracy a jointly

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.

trained dense model forfeits to this choice its *representation debt*: a liability written into the shared representation at training time. Standard optimizer-side interventions do not repay it, while the carrier-matched one nearly clears it (§4.7). Scale does not repay it (§4.3), and expert routing does not incur it.

Our contributions are four.

- **An instrument.** We release SphereBench-54 and its three-serialization protocol, and a 423-arm matched-compute study (99 jointly trained models, 324 per-dataset specialists), three seeds throughout.
- **A newly isolated failure mode: representation debt.** To our knowledge, the first demonstration that gradients from adversarially serialized tasks propagate through a shared encoder and degrade byte-identical untouched tasks. The effect is a 24.5-point macro-accuracy swing decided by input serialization (paired Wilcoxon over the 54 datasets, $z = 6.31$), with expert-routed architectures immune at equal or greater budgets (amplitude ≤ 1.62 points, and no expert cross-serialization comparison survives Bonferroni correction over nine comparisons).
- **A mechanism account.** Three diagnostic probes triangulate the failure. The collapsed trunk's joint gradient is nearly rank-one (norm-dominated capture) and modality-structured (§4.5). Class information stays probe-decodable far above the trained readout while cross-modality centroid coupling doubles (§4.5). The model also fails to fit its own training data, a representation failure rather than memorization (Appendix D). A controlled dose scan locates the ignition threshold between two and three hostile datasets (§4.4). Interventions close the loop: head trainability repays 4.8 points, norm equalization repays 23.7, and a vanilla trunk without the spherical-iterative core stays immune (Appendix G).
- **A scope statement and a price tag.** A controlled glass-box testbed separates test-time corruption from training-time pollution and shows the debt requires shared iterative dynamics to propagate (§4.6). Per-dataset specialists beat the weakest flatten-line dense cell on 48 of 54 datasets (+14.6 points, mean over all 54) and trail the strongest generalist by 2.38 on average (§4.2).

2 Related Work

Multi-task interference, isolation, and probing. Jointly trained tasks can hurt each other across vision, NLP, and reinforcement learning (Ruder, 2017; Fifty et al., 2021; Wang et al., 2021; 2019). Catastrophic interference predates deep learning and drives the continual-learning literature (McCloskey & Cohen, 1989; French, 1999; Kirkpatrick et al., 2017). Mitigations cluster into optimizer-side methods (gradient surgery (Yu et al., 2020), uncertainty and dynamic loss weighting (Kendall et al., 2018; Liu et al., 2019), per-task gradient-norm balancing (Chen et al., 2018), and multi-objective conflict resolution (Sener & Koltun, 2018; Liu et al., 2021; Navon et al., 2022)) and architecture-side isolation (cross-stitch units (Misra et al., 2016), multi-gate mixtures (Ma et al., 2018), progressive layered extraction (Tang et al., 2020), and sparsely gated mixtures (Shazeer et al., 2017)). Sparse routing carries its own representation-collapse mode (Chi et al., 2022). Audits find such methods rarely beating unitary scalarization or even random weightings under matched compute (Kurin et al., 2022; Xin et al., 2022; Lin et al., 2022). Multimodal work documents gradient-magnitude imbalance across modalities, remedied by on-the-fly modulation (Wang et al., 2020; Peng et al., 2022). What is missing is a controlled measurement of who pays: identical data and budgets, one manipulated interference source, and all three operator families on one instrument. Expert routing isolates tasks in weight space, latent bottlenecks in activation space (Perceiver-IO (Jaegle et al., 2022), our contrast class). Our instrument places both families on one axis. Our mechanism analysis applies standard instruments (effective rank (Roy & Vetterli, 2007), linear probes (Alain & Bengio, 2016) and training-free k-nearest-neighbor (kNN) readouts, simplex-ETF (equiangular tight frame) analysis (Papyan et al., 2020)) to retained checkpoints across world lines.

Serialization and benchmarks. Space-filling curves are long-standing locality-preserving serializations (Skilling, 2004). Sequence-model order-sensitivity on unordered data is known in point-cloud and set-processing contexts (Qi et al., 2017; Vinyals et al., 2016; Zaheer et al., 2017; Lee et al., 2019). SphereBench-54 treats serialization as a controllable pollution source whose blast radius across *other*

modalities is the quantity of interest. Multi-task evaluation suites (task collections (Zamir et al., 2018), few-shot meta-learning (Triantafillou et al., 2020), multimodal fusion (Liang et al., 2021)) measure aggregate capability. SphereBench-54 is instead an interference instrument: a deliberate modality mix and per-dataset specialist references that price sharing task by task. The closest practice is leave-one-out task ablation and learned task grouping (Standley et al., 2020). Our knob manipulates interference intensity at fixed task set, budgets, and bytes. Mixture ratios and phase behavior interact in pretraining (Xie et al., 2023; Ye et al., 2025; Power et al., 2022). The spherical iterator of Figure 1 descends an explicit energy, kin to energy-based and implicit models (Du & Mordatch, 2019; LeCun et al., 2006; Bai et al., 2019). Our UniversalEncoder follows the generalist one-encoder line (Jaegle et al., 2021; Lu et al., 2022; Wang et al., 2022; Reed et al., 2022).

3 THE SPHEREBENCH-54 INSTRUMENT

3.1 BENCHMARK AND METRIC

SphereBench-54 assembles 54 public classification datasets across 7 modality families by design: vision (14 datasets), text (9), tabular (8), time series (8), audio (7), point cloud (4), and graph (4). Full list, sources, and statistics are in Appendix A.

Every joint arm trains on all 54 datasets with per-dataset heads. The headline metric is **macro-accuracy**: the unweighted mean of the 54 per-dataset test accuracies. Unless flagged, every number is the mean over three seeds (42, 123, 456) with sample standard deviation as $\pm$. Checkpoints are selected by best class-balanced test accuracy (reported accuracies are unweighted), typically an interim capped-batch pass. Four uncapped controls bound the induced deviation $\leq$0.07 points, and selection is shared by every arm.

3.2 THREE WORLD LINES

The experimental knife is the input representation of the four point-cloud datasets. A point cloud is a set of 3D coordinates; the encoder requires an order, or a rendering replacing the set. We compare three *world lines*: **flatten**, the raw coordinate array in stored order, with no spatial-locality guarantee; **views**, six orthographic depth projections merged into one three-channel depth image, a lossy rendering replacing the set; and **hilbert**, points sorted along a Hilbert space-filling curve (Skilling, 2004), a locality-preserving total order from a Skilling-construction code over a quantized grid.

World lines differ only in this choice: the four point-cloud datasets change representation. The other fifty are byte-identical across conditions, so any cross-line difference beyond the four is training dynamics, not data. Point budget, quantization grid, code construction, view geometry, and the byte-identity manifest diff are in Appendix A.

3.3 ARCHITECTURES

Six architectures share one data pipeline, training protocol, and parameter target. The isolated variable is the fusion mechanism. T01 is a bespoke spherical-iterative trunk introduced in this study. Five share the *UniversalEncoder*, our shared input frontend: a per-modality routing layer emitting uniform tokens $X \in \mathbb{R}^{B\times L\times d}$ to the trunk using 2D patch projection for image-like inputs, flatten–AdaIN (adaptive instance normalization) 1D patching for tensor-like inputs, token or byte-level embedding for text (T01 takes the byte path), and either 1D coordinates or 2D views for point clouds. The exception is Perceiver-IO, which uses its own per-modality projections. Table 1 gives exact trainable parameters.

T01 is the dense trunk: a single shared encoder iteratively refining a spherical latent state, with skill embeddings conditioning the read stage. It has no routing, gating, or expert structure. Apart from skill embeddings (7.9% of parameters), every task's gradient passes through the same weights, while the frozen ETF prototypes receive none. ETF-plain differs only in the text frontend (legacy generic path versus T01's byte-level encoder). Perceiver-IO is a PyTorch port of the official architecture (Jaegle et al., 2022). MMoE, PLE, and Cross-Stitch are expert-routed (Shazeer et al., 2017; Ma et al., 2018; Tang et al., 2020; Misra et al., 2016).

Table 1: Architectures in the controlled comparison, with exact trainable parameter counts. The dense trunks (T01, ETF-plain) and MMoE are parameter-matched within ±3% at ≈1.5M. Perceiver-IO, PLE, and Cross-Stitch carry 1.7–2.4× more trainable parameters (decoders, experts, and branches scale with task count). They stay at their reference architectures, and the overhang favors them in §4.1–§4.4. Frozen buffers (ETF prototypes) are excluded from the counts.

Architecture	Fusion class	Isolation substrate	Trainable params
T01 (shared spherical iterative core)	dense, single trunk	none	1,538,693
ETF-plain	dense, single trunk	none	1,553,485
Perceiver-IO (official port)	latent bottleneck	fixed-size latent array	2,635,246
MMoE	expert-routed	7 shared experts + per-dataset gates & heads	1,509,885
PLE	expert-routed	2 shared + 1 per-dataset expert, gates & heads	3,305,509
Cross-Stitch	expert-routed	per-dataset branch pairs + 54 × 54 stitch matrix	3,611,015

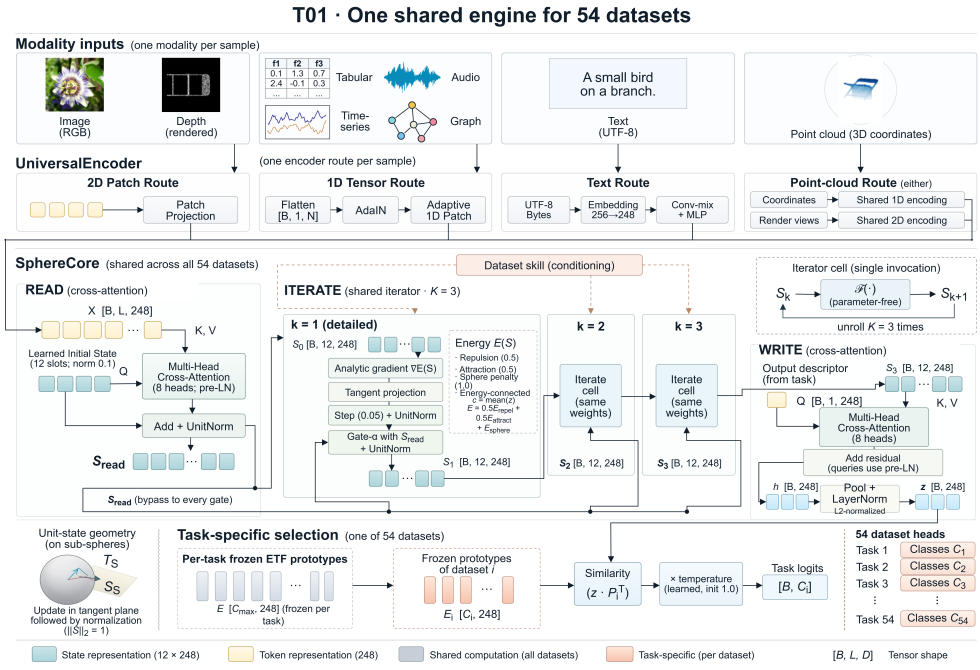

Figure 1: Architecture of T01. (1) The UniversalEncoder routes each sample by modality (2D patch projection for images; adaptive normalization plus interpolated-kernel 1D patching for tensors; byte embedding for text), emitting tokens $[B, L, d]$. (2) Read gathers tokens onto a learned 12-slot unit state by 8-head cross-attention (Locatello et al., 2020). (3) The shared iterator refines the state for K = 3 energy-descent steps (tangent-projected gradient, step size 0.05, gate-α blend with the read state), conditioned by per-dataset skill embeddings. (4) A descriptor query reads the refined state by cross-attention; the pooled feature is L2-normalized. (5) Each dataset scores it against its own frozen ETF prototype set under a learned temperature. Apart from skill embeddings, every task's gradient passes through the same weights, and the prototypes are frozen.

3.4 Protocol and arm accounting

The campaign comprises **423 training arms under one protocol**: 99 joint arms on all 54 datasets (54 main-scan: 6 architectures × 3 world lines × 3 seeds; 27 budget-scan: 3 architectures × 3 budgets × 3 seeds, 310K–3M nominal; 18 sample-efficiency: 2 architectures × 3 fractions × 3 seeds, stratified train-split subsampling) and 324 per-dataset specialists (290 main-sweep; 34 earlier classic-17 cells, pooled with two main-sweep seeds so every cell carries three seeds). Follow-up probes under the identical protocol (24-arm scale, §4.3; 9-arm dose, §4.4; 3-arm order-augmentation, §4.7; 23 intervention and trunk-control arms (six learned-head, three norm-equalization, six vanilla-trunk, three hostile-text, two gradient-probe, three direction-surgery), Appendix G) bring the study to 482 arms. Main-scan, scale, dose, augmentation, and follow-up arms use full-test evaluation. The budget, sample-efficiency, and specialist studies use the capped protocol of §3.1 (the 34 classic-subset arms predate it; seven uncapped exceptions in Appendix F), with bias bounds in Appendices A/F. The probes of §4.5 use a labeled diagnostic protocol.

Table 2: Macro-accuracy (mean ± sample std, 3 seeds) on SphereBench-54, by architecture and world line. Amplitude is the max–min difference across the three serializations, in accuracy points (pp). Jointly trained arms only; dense arms parameter-matched within ±3% at ≈1.5M, exact counts in Table 1. †Control arm (App. G.5); flatten not run, amplitude spans views–hilbert only. Means and stds are computed at full precision, then rounded at display.

Architecture	flatten	views	hilbert	Amplitude (pp)
T01	0.5034 ± .0037	**0.6363 ± .0033**	0.3916 ± .0230	**24.47**
ETF-plain	0.4660 ± .0025	0.6096 ± .0031	0.3794 ± .0092	23.02
Vanilla trunk†	—	0.6275 ± .0021	0.6396 ± .0058	+1.21†
Perceiver-IO	0.4994 ± .0072	0.4993 ± .0076	0.5043 ± .0071	0.50
MMoE	0.5872 ± .0062	0.5970 ± .0071	0.5921 ± .0068	0.98
PLE	0.5912 ± .0019	0.6025 ± .0069	0.6023 ± .0056	1.12
Cross-Stitch	0.5733 ± .0006	0.5896 ± .0013	0.5880 ± .0027	1.62

3.5 Formal setup and hypotheses

Let a index an architecture arm (one architecture–budget configuration, and every reported cell pools its three seed runs) and $w \in \{\text{flatten}, \text{views}, \text{hilbert}\}$ a world line. Write $\mathrm{Acc}_d(a, w)$ for test accuracy of arm a on dataset d under line w, and $\mathrm{Acc}(a, w) = \frac{1}{54}\sum_d \mathrm{Acc}_d(a, w)$ for macro-accuracy.

$$\mathcal{D}(a) = \mathrm{Acc}(a, \text{views}) - \mathrm{Acc}(a, \text{hilbert}) \qquad \text{(representation debt)} \tag{1}$$

$$\mathcal{C}(a, w) = \operatorname*{mean}_{d<e} \cos(g_d, g_e), \qquad g_d = \partial \mathcal{L}_d / \partial \theta \text{ (all parameters)}. \tag{2}$$

The *amplitude* $\mathcal{A}(a) = \max_w \mathrm{Acc}(a, w) - \min_w \mathrm{Acc}(a, w)$ summarizes an arm's serialization sensitivity (Table 2).

$$k \;=\; \#\{\text{point-cloud datasets } d : \pi(d) = \text{hilbert}\} \;\in\; \{1, 2, 3, 4\} \tag{3}$$

counts the point-cloud datasets carrying the hilbert serialization. Five hypotheses, stated before aggregation, organize the results. Each carries its verdict and section.

- **H1 (cascade).** Dense trunks incur $\mathcal{D} \gg 0$; routed architectures incur $\mathcal{D} \approx 0$ at matched budget. **Supported** (§4.1).
- **H2 (capacity repays).** $\mathcal{D}$ shrinks as parameters grow. **Falsified** (§4.3).
- **H3 (dose threshold).** The collapse is thresholded in k rather than graded. **Supported, identified up to modelnet10's entry** (§4.4).
- **H4 (rank-one carrier).** $\mathcal{C}(\text{T01}, \text{hilbert})$ far exceeds $\mathcal{C}$ elsewhere. **Supported** (§4.5).
- **H5 (routing buys data efficiency).** Routed arms dominate dense arms at reduced fractions. **Null** (§4.3).

3.6 The operators under test

Three operator families decide who pays the debt of equation 1: serialization π of the point-cloud inputs, the single manipulated variable; optimizer-side interventions, which reshape the joint update after the trunk has mixed per-task gradients (gradient surgery (Yu et al., 2020), dynamic loss reweighting (Liu et al., 2019)); and architecture-side isolation, which decides where tasks may write at all (expert routing (Shazeer et al., 2017; Ma et al., 2018; Tang et al., 2020), cross-stitch (Misra et al., 2016), latent bottleneck (Jaegle et al., 2022)). Definitions and equations appear in Appendix B.1.

4 Experiments

4.1 The pollution result

Under views, T01 reaches 0.6363, the best joint macro-accuracy at the matched budget. Under hilbert, the same architecture, budget, and fifty untouched datasets fall to 0.3916, a 24.47-point swing decided by the serialization of four point-cloud datasets (paired Wilcoxon over the 54 per-dataset means, z = 6.31, $p = 2.8 \times 10^{-10}$; Appendix F). ETF-plain swings 23.02 points. The three expert-routed architectures move by at most 1.62 points, and no cross-line expert comparison survives Bonferroni correction over the nine expert pairs (minimum $p = 8.7 \times 10^{-3}$; Table 8). Of the swing, 13.3 points

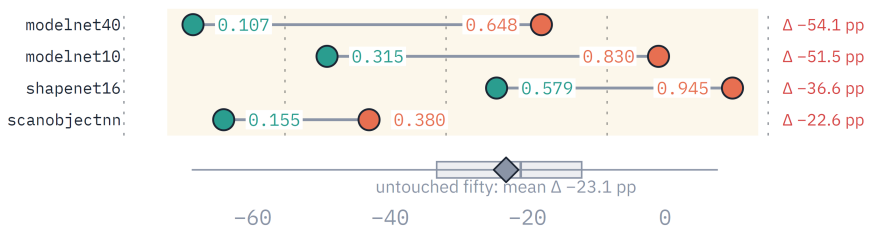

Figure 2: A 24.47-point collapse under byte-identical controls: T01 swings from 0.6363 under views to 0.3916 under hilbert, while expert-routed architectures move by at most 1.62 points and the latent-bottleneck Perceiver-IO is flat. The lower band summarizes the drop over the fifty untouched datasets: range, quartile box, median tick, mean diamond.

are views–flatten contrast (rendering versus sequence) and 11.2 flatten–hilbert (pure ordering), so the collapse is large under the pure ordering contrast alone.

Table 2 invites three misreadings, and we dispense with them in turn. The collapse does not mean hilbert is intrinsically unreadable. The same ordering was T01's *strongest* world line in the 47-dataset era (0.6331; Appendix F). Perceiver-IO's stability is not strength either. Its cells are flat (amplitude 0.50) but mediocre, and its ranking against T01 reverses across world lines; immunity without headroom is a different failure mode. Finally, the collapse is not a property of dense sharing as such. A parameter-matched vanilla trunk under the identical protocol does not collapse (Appendix G.5), so the vulnerability lives in the spherical-iterative core, not in missing routing. The benchmark's three generations (17, 47, and 54 datasets) share one protocol lineage, and era-scoped evidence is flagged where it appears (here and §4.7).

4.1.1 Propagation to untouched data

Pollution is not confined to the four re-serialized datasets. On the 14 byte-identical vision datasets, T01's mean falls from 0.4603 (views) to 0.1856 (hilbert), with flatten at 0.3073. Drops reach -53.2 points (emnist_letters 0.6491 $\rightarrow$ 0.1173; full matrix in Appendix C). The hilbert cell carries the table's largest seed variance ($\pm$0.0230), so the collapsed regime is worse and less determined. Gradients from four hostile serializations degrade the shared encoder for everyone, and §4.5 shows the carrying geometry.

4.2 The price of sharing: per-dataset specialists

Per dataset d, the specialist debt is

$$\mathcal{D}_{\text{spec}}(d) \;=\; \text{Acc}_{\text{spec}}(d) - \text{Acc}_{\text{joint}}(d). \tag{4}$$

The 324 specialist arms, per-dataset models trained alone under the same recipe, price it against the flatten/ETF-plain joint cell (0.4660, the weakest flatten-line dense cell). Specialists use the capped protocol against the baseline's full test split. Measured cap bias on the stable-regime control arms is negligible ($\leq$0.07 points; semi-collapse levels drift more, Appendices A and F). Specialists beat the joint model on **48 of 54 datasets**, and the mean gain over all 54 is **+14.64 points** (small tier). Against the strongest generalist (T01 under views) they win 27 of 54 and trail by 2.38 on average. Both baselines are reported because the debt is generalist-relative: benign sharing is break-even, hostile sharing ruinous. The best specialist reaches 0.9916 (wafer, small tier; 0.9921 at the large tier). The collapse leaves T01 below its own specialist on nearly every dataset (Appendix C).

4.3 Scaling does not repay the debt

The gap (signed $\Delta = \text{Acc}(\cdot, \text{hilbert}) - \text{Acc}(\cdot, \text{views})$, the negation of the debt of equation 1) does not shrink with scale (Table 3, Figure 4). It widens from -24.5 points at 1.54M to -33.2 at 3.17M and -38.1 at 10.75M, then saturates at 18.27M (plateau at $\approx$0.269), while the MMoE-XL control at 8.46M brackets zero ($+0.2$ points). Expert routing is immune at matched scale, so the hilbert data are not simply harder. Per-seed values and auxiliary scans are in Appendix F. Capacity is not the bottleneck for routing architectures, and routing buys robustness, not data efficiency.

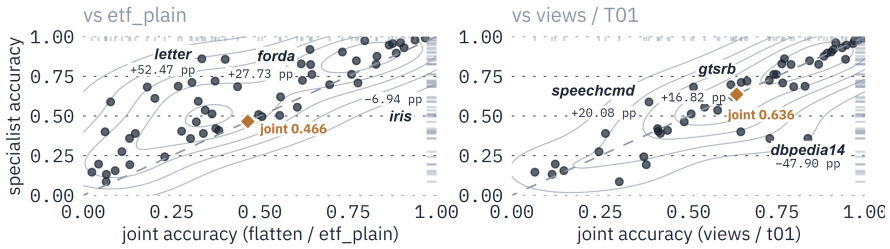

Figure 3: The price of sharing: per-dataset specialists beat the flatten/ETF-plain joint cell (0.4660) on 48 of 54 datasets, with a +14.64-point mean gain over all 54 (small tier); against the strongest generalist (T01 under views) they win on only 27 of 54 datasets and trail by 2.38 points on average. Dashed lines mark parity.

Table 3: Scale verdict (macro-accuracy, official full-test protocol). The 1.54M anchor row is the main-scan cell of Table 2. All rows are means over three seeds; Δ = hilbert − views, the negation of the debt of equation 1.

Family	Params	Seeds	views	hilbert	Δ (pp)
T01 anchor	1.54M	3	0.6363 ± .0033	0.3916 ± .0230	−24.5
T01-deep	3.17M	3	0.6431 ± .0073	0.3110 ± .0362	−33.2
T01-XL768	10.75M	3	0.6506 ± .0021	0.2697 ± .0020	−38.1
MMoE-XL	8.46M	3	0.6044 ± .0031	0.6063 ± .0021	+0.2
T01-XXL	18.27M	3	0.6397 ± .0227	0.2688 ± .0012	−37.1

4.4 Dose–Response: Collapse Onsets Between Two and Three Hostile Datasets

A dose scan varies the number of hilbert-serialized point-cloud datasets, $k \in \{1, 2, 3\}$ (construction in Appendix F; $k = 4$ is the main-scan hilbert cell). One or two hostile datasets cost the trunk almost nothing (0.6341, 0.6299, within a point of the views cell of Table 2), while three trigger the full collapse (0.3697; Table 4, Figure 5). The transition is a cliff between doses two and three, not a gradient: dose four sits above the trough (0.3916 versus 0.3697), marking a threshold crossing rather than monotone descent. Dose cells average 51/52/53 datasets at k=1, 2, 3 against 54 at k=4 (Appendix F), and d3 overlaps d4 within seed variance. The pollution is dose-dependent, and below the threshold the dense trunk is the best matched-budget jointly trained model we have. The debt is incurred crossing the threshold, not by sharing per se.

4.5 Mechanism

Protocol note. Everything here uses a **diagnostic protocol (capped eval)**: one retained checkpoint per arm (seed 42; §3.1), four batches per dataset for P2, two for P3.

4.5.1 The collapsed trunk's gradient is nearly rank-one (P3)

For each retained checkpoint we compute per-dataset loss gradients over all parameters, then mean pairwise cosine across dataset pairs. Head coordinates are disjoint across datasets and contribute only to gradient norms, so cosines conservatively read shared-trunk coupling. The per-configuration means are collected in Table 5 (Supplementary Tables, p. 13).

The hilbert/T01 trunk sits at +0.83 mean pairwise cosine, nearly rank-one, while every other probed configuration (the same architecture on the other two serializations; MMoE, PLE, and Perceiver-IO on the same hostile serialization) stays at or below +0.08. Behavioral and geometric collapse are in phase across all six probed configurations, the signature of a norm-dominated rank-one capture of the shared update. Appendix H states these mechanisms formally (Statements S1–S7).

The drag is modality-structured. Aggregating the cosine matrix into 7×7 modality blocks (per-block values in Appendix D), every non-point-cloud modality couples outward at +0.798 to +0.865 while the point-cloud block sits at +0.650. The drag is imposed on the whole mixture, not negotiated per pair. One dataset, modelnet10 under hilbert, carries gradient norm 10136 against 4–11 for all others, the single near-orthogonal survivor, while the remaining 53 uphold the rank-one structure. Hostile tasks set the drag direction, and one hostile dataset's outsized gradient norm scales the shared update. Per-epoch probes show the signature is training-long: alignment stays elevated and volatile on hilbert, decaying to zero on views (Appendix G.4).

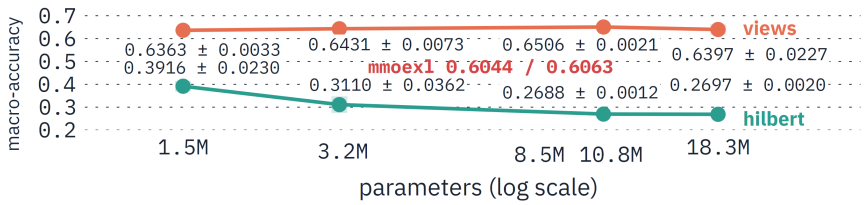

Figure 4: Scaling does not repay the debt: the views–hilbert gap widens from −24.5 points at 1.54M to −33.2 at 3.17M and −38.1 at 10.75M, while the hilbert level saturates at 18.27M (≈0.269). The MMoE-XL control at 8.46M brackets zero from either side.

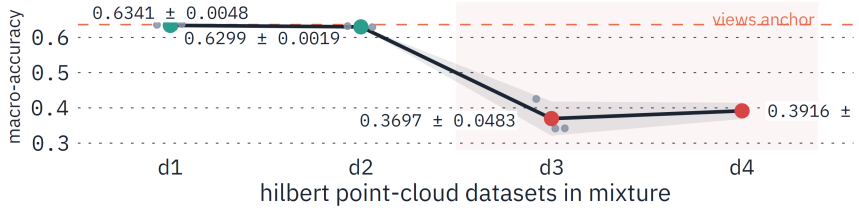

Figure 5: Dose–response of the collapse: one or two hilbert-serialized point-cloud datasets cost the trunk almost nothing (0.6341, 0.6299), while three trigger the full collapse (0.3697). The transition is a cliff between dose two and three (d1/d2/d3 macro over 51/52/53 datasets; d4 over 54).

4.5.2 The information is present; the readout cannot collect it (P2)

We probe the latent state of the three T01 checkpoints (one per world line) with a training-free kNN readout and two geometric measures. The probe values are collected in Table 6 (Supplementary Tables). First, class information survives the collapse: a kNN on the hilbert trunk's frozen latent decodes at 0.541, far above the 0.388 the trained readout achieves (Table 7). On food101 the frozen latent decodes at 0.9766 while the trained readout scores 0.0108 (full-test, three-seed mean; the two evaluations differ in protocol). Second, latent geometry degrades specifically: effective rank is halved relative to views (3.9 → 4.5 versus 8.6 → 9.6 across refinement), and cross-modality centroid coupling more than doubles (0.804 versus 0.328 and 0.365; cf. modality gap (Liang et al., 2022)). Third, task identity remains strongly readable (0.77–0.89), with 94.7–99.9% of misidentifications within the same modality. The collapsed representation is not scrambled but entangled along modality lines, compressed into a low-rank, coupled subspace the readout cannot separate. A learned-head control (same ETF initialization, gradients enabled; Appendix G.1) repays only 4.8 points of the 24.5-point gap on hilbert. The frozen readout is a minor tax, and the trunk's entanglement is the main term.

P1 places the collapse on the generalization-gap taxonomy (Appendix D): the collapsed hilbert/T01 trunk shows a +0.016 train–test gap with 32 of 54 datasets representation-failure and 4 memorization, while routed and bottlenecked models show ordinary gaps (+0.121 to +0.169) with memorization dominant. The flatten/T01 line sits midway (+0.043, 26 of 54 representation-failure), so hostility is graded across orderings even where dose is thresholded. The failure is representational, exactly where the rank-one gradient of equation 2 writes the debt.

4.6 Scope and Boundary: Test-Time Corruption Is Not Training-Time Pollution

A controlled glass box (single-task, four modalities, late concatenation; Appendix E) separates the two directions: test-time corruption is cheap (permuting a modality's token order at evaluation costs 5–21% of its reliance; time reversal costs 1.67×), and training from scratch on the hostile ordering changes nothing (+0.0021 against clean). Representation debt requires the shared, iteratively refined multi-task trunk of §4.5, which the glass box deliberately lacks, so its nulls bound rather than establish the mechanism.

4.7 Toward Mitigation

No optimizer-side intervention tested on this family (17-dataset classic era) repaid the debt. PCGrad (Yu et al., 2020) scored 0.6571 against 0.6587 unmodified (−0.16 points, inside noise). Dynamic

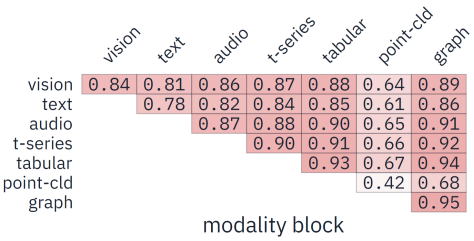

Figure 6: The collapsed hilbert/T01 trunk's joint gradient is nearly rank-one: mean pairwise task-gradient cosine +0.8318, while every other configuration stays at or below +0.08. The drag is modality-structured and imposed on the whole mixture.

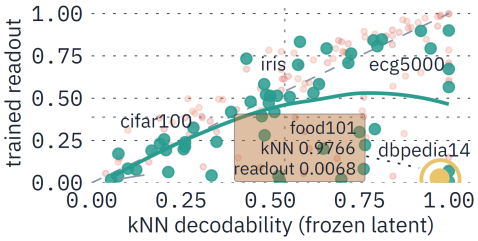

Figure 7: Class information survives the collapse: a training-free kNN readout on the hilbert trunk's frozen latent decodes at 0.541, far above the 0.388 of the trained readout. Geometry measures in Table 6. Faint points show the views and flatten trunks.

loss reweighting (Liu et al., 2019) recovered +2.9 points over unmodified (0.6880 versus 0.6587) yet stayed below the representation-repair control (0.7340). A diversity-promoting auxiliary loss cost 1.2 points, and doubling refinement depth left accuracy flat. The same era's representation-side repair (a learned byte-level text encoder) gained +15.1 points on the most affected task, showing that representation-side repairs clear debts that these optimizer-side interventions do not.

The 54-dataset follow-up campaign turns that asymmetry into a diagnosis. Per-task gradient normalization rescales each task's gradient to the mean per-task norm before accumulation (Appendix G.3),

$$\Delta\theta \;=\; -\eta \sum_{d} \frac{\bar{g}}{\|g_d\|}\, g_d, \qquad \bar{g} = \tfrac{1}{54}\textstyle\sum_e \|g_e\|, \tag{5}$$

leaving the aggregate step size unchanged. Its direction-based counterpart, PCGrad, instead projects each task's gradient off the normal planes of all conflicting tasks,

$$g_d \;\leftarrow\; g_d - \frac{g_d \cdot g_e}{\|g_e\|^2}\, g_e \qquad \text{for every } e \text{ with } g_d \cdot g_e < 0, \tag{6}$$

reshaping directions while leaving norms untouched. The two operators occupy the same accumulation slot. The intervention lifts the hilbert cell from 0.3916 to **0.6288** $\pm$ **.0038**, repaying 23.7 of the 24.5 points with zero dead datasets (accuracy below 0.05). PCGrad on the same 54-dataset cell repays nothing (0.3939 $\pm$ .0018 against the unmodified anchor, $+0.2$ points inside its seed spread; Appendix G.6). The carrier of §4.5 is causal, not correlational. Equalize the norm domination and the debt is never written. The debt is not unpayable. It is invisible to optimizers that ignore gradient-norm structure.

The constructive hypothesis is data-side: if the debt is written by the hostile serialization at training time, order augmentation (each sample under a random same-multiset reordering) may immunize the trunk at zero architectural cost. The order-augmentation arm tests exactly this: a verbatim clone of the main-scan hilbert arm (T01, 3 seeds, matched compute), manipulating only training-time order augmentation. The test split is plain hilbert. Order augmentation scores **0.4829** $\pm$ **.0165** against the hilbert anchor: a **+9.1-point recovery**, 37% of the 24.5-point gap, though 15.3 points below views. A substantial share of the debt is thus data-side and repayable at zero architectural cost. The remainder yields to the norm-matched intervention above, or never forms without the spherical-iterative core (Appendix G.5).

5 Conclusion

Within our setting the conclusion is strong. Architectural isolation is a sufficient safeguard against training-time serialization pollution, and no standard optimizer-side intervention is a substitute. The debt yields to the carrier-matched intervention (§4.7) and never forms without the spherical-iterative core (Appendix G.5). The effect is visible without any test: 24.5 points, same architecture, data, and budget, one serialization changed. The deep cell carries the family's widest spread (0.2701 to 0.3392), though the gap's sign and scale are unchanged. *Future work.* Dose-entry-order permutation and extension beyond classification. SphereBench-54 makes each a measurement, not an opinion.

AI USE STATEMENT

In this work, we used generative AI tools for code (experiment infrastructure, data pipelines, and figure-generation scripts) and for automatic logistic tasks (manuscript typesetting, formatting, and literature cross-referencing support). We have not used generative AI tools for formulating the research questions, the experimental framework, the pre-stated hypotheses, or the novelty claims of the individual studies; each of these was formulated by the authors. All AI-assisted work was reviewed by the authors: every reported number was audited against raw training logs through a maintained fact-gate, and all AI-generated code was executed and tested for correctness. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All 54 datasets in SphereBench-54 are public and widely used. This work involves no human subjects, no personally identifiable information, and no new data collection from individuals. The paper releases a benchmark and a measurement protocol rather than a deployed system; we judge the potential for dual use to be minimal, since the stated goal of the benchmark is to make interference in joint training measurable so that jointly trained models can be made safer and more robust. The authors declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

Every number reported in this paper traces to a maintained fact-gate over raw training logs. All experiments use three fixed seeds unless flagged (the §4.5 checkpoint probes and Appendix G.4 diagnostics are single-seed by design), under the per-study evaluation protocols disclosed where used; both the full-test and capped evaluation variants are bounded in Appendix A. Benchmark construction, including the dataset list, splits, serialization manifests, and the byte-identity manifest diff between conditions, appears in Appendix A. Architectures and exact parameter counts are given in Appendix B and Table 1, per-dataset results in Appendix C, and probe protocols in Appendix D. The benchmark, its serialization manifests, and the evaluation protocol are released with this submission at `https://anonymous.4open.science/r/RepDebt-A449`.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2016.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International Conference on Machine Learning (ICML)*, 2018.

Zewen Chi, Li Dong, Shaohan Huang, Damai Dai, Shuming Ma, Barun Patra, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, Heyan Huang, and Furu Wei. On the representation collapse of sparse mixture of experts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Christopher Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3 (4):128–135, 1999.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning (ICML)*, 2021.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations (ICLR)*, 2022.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences (PNAS)*, 114(13):3521–3526, 2017.

Vitaly Kurin, Alessandro De Palma, Ilya Kostrikov, Shimon Whiteson, and M. Pawan Kumar. In defense of the unitary scalarization for deep multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu Jie Huang. A tutorial on energy-based learning, 2006. In Bakir et al. (eds), Predicting Structured Data, MIT Press.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam 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)*, 2019.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A. Lee, Yuke Zhu, Ruslan Salakhutdinov, and Louis-Philippe Morency. MultiBench: Multiscale benchmarks for multimodal representation learning. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2021.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Baijiong Lin, Feiyang Ye, Yu Zhang, and Ivor W. Tsang. Reasonable effectiveness of random weighting: A litmus test for multi-task learning, 2022.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Shikun Liu, Edward Johns, and Andrew J. Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-IO: A unified model for vision, language, and multi-modal tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H. Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 2018.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pp. 109–165. 1989.

Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In *International Conference on Machine Learning (ICML)*, 2022.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C. Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Scott Reed et al. A generalist agent, 2022.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European Signal Processing Conference (EUSIPCO)*, pp. 606–610, 2007.

Sebastian Ruder. An overview of multi-task learning in deep neural networks, 2017.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

John Skilling. Programming the Hilbert curve. In *AIP Conference Proceedings*, volume 707, pp. 381–387, 2004.

Trevor Standley, Amir R. Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International Conference on Machine Learning (ICML)*, 2020.

Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. Progressive layered extraction (PLE): A novel multi-task learning (MTL) model for personalized recommendations. In *Fourteenth ACM Conference on Recommender Systems (RecSys)*, 2020.

Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *International Conference on Learning Representations (ICLR)*, 2020.

Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets. In *International Conference on Learning Representations (ICLR)*, 2016.

Table 4: Dose–response of the collapse (macro-accuracy, mean ± sample std, 3 seeds). Dose cells are macro-accuracy over each arm's own training mixture (51, 52, and 53 datasets at d1, d2, and d3); d4_ref is the 54-dataset main-scan hilbert cell of Table 2, evaluated under the identical full-test protocol. The collapse regime has large seed variance; means must be read with their deviations.

Dose (point-cloud datasets on hilbert)	Macro-accuracy
d1	$0.6341 \pm .0048$
d2	$0.6299 \pm .0019$
d3	$0.3697 \pm .0483$
d4 (reference, main scan)	$0.3916 \pm .0230$

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning (ICML)*, 2022.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Zirui Wang, Yulia Tsvetkov, Orhan Firat, and Yuan Cao. Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. In *International Conference on Learning Representations (ICLR)*, 2021.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. DoReMi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Derrick Xin, Behrooz Ghorbani, Ankush Garg, Orhan Firat, and Justin Gilmer. Do current multi-task optimization methods in deep learning even help? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Jiasheng Ye, Peiju Liu, Tianxiang Sun, Yun Hua, Junfeng Zhan, and Xipeng Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. In *International Conference on Learning Representations (ICLR)*, 2025.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Amir Zamir, Alexander Sax, William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

SUPPLEMENTARY TABLES

A BENCHMARK

Full dataset list with sources, modality census, and split conventions; point-cloud construction (the fixed point budget per cloud and its resampling rule, the quantization grid, the Skilling Hilbert-code construction, and the six-view depth-rendering geometry); the byte-identity manifest diff between world lines; the capped-evaluation protocol (stratified cap 2,000, fixed subsample shared across arms, measured bias bounds, and the study arms evaluated before the cap existed).

Table 5: Mean pairwise task-gradient cosine (diagnostic protocol; single seed-42 checkpoint, 2 batches).

Checkpoint (world line / architecture)	Mean pairwise cosine
hilbert / T01	**+0.8318** (median +0.9072, min -0.0361)
flatten / T01	+0.0772
views / T01	+0.0012
hilbert / MMoE	+0.0052
hilbert / PLE	+0.0014
hilbert / Perceiver-IO	-0.0004

Table 6: Latent probes of T01 checkpoints (diagnostic protocol; seed-42 checkpoints). kNN decodability is the mean over the 54 datasets of per-dataset cosine leave-one-out kNN accuracy (k = 10) at the read stage. Effective rank is the participation ratio of the pooled latent's covariance spectrum (cf. the entropy-based effective rank of Roy & Vetterli (2007)), read $\rightarrow$ after one refinement step. Centroid coupling is the mean cosine between per-dataset latent centroids over dataset pairs in different modalities (within-modality for reference), at the read stage. Task identity is a cosine 1-NN read of which of the 54 datasets a latent belongs to (up to 128 latents per dataset, read stage; chance 0.0185).

Checkpoint	kNN decodability	Effective rank	Cross-modality coupling	Task-identity 1-NN
views / T01	0.671	8.6 $\rightarrow$ 9.6	0.328	0.885
flatten / T01	0.593	5.5 $\rightarrow$ 6.4	0.365	0.839
hilbert / T01	0.541	3.9 $\rightarrow$ 4.5	0.804	0.767

Point-cloud construction: each cloud is resampled to a fixed budget of 1,024 points (deterministic subsampling under a fixed generator seed when over budget, cyclic repetition when under), so every epoch sees the same point subset. The hilbert line quantizes to a 2^8 grid per axis over the unit cube and sorts by a Skilling-construction Hilbert code (stable argsort; §B.1); the views line renders six orthographic depth views at 64×64 under unit-sphere normalization, with nearest-depth buffering and opposing views max-merged to three channels. Byte-identity across world lines is verified by manifest diff: the fifty shared datasets carry byte-identical configuration entries and data files, and only the four point-cloud `pc_repr` flags differ.

The 54 datasets by modality family (all public): *vision (14)*: mnist, fmnist, kmnist, cifar10, cifar100, svhn, stl10, gtsrb, emnist_letters, eurosat, dtd, flowers102, food101, oxford_pets; *text (9)*: trec6, agnews, dbpedia14, sst2, yelp_full, news20, imdb, cola, dna_enhancers; *tabular (8)*: digits, wine, breast_cancer, iris, letter, covertype, shuttle, magic04; *time series (8)*: ecg5000, forda, electricdevices, wafer, starlightcurves, uwgesture, fordb, ecg200; *audio (7)*: esc50, speechcmd, gtzan, urbansound8k, ravdess, cremad, nsynth; *point cloud (4)*: modelnet40, scanobjectnn, modelnet10, shapenet16; *graph (4)*: mutag, proteins, enzymes, dd.

Checkpoint-selection naming: `val_` is a historical misnomer for the test split, and §3.1 flags the induced selection optimism. The capped evaluation protocol draws one fixed stratified subsample (cap 2,000 per dataset, proportional class allocation, one fixed seed) shared across arms, keeping capped comparisons paired.

B Architectures and protocol

Per-arm realized trainable and frozen parameter counts from run manifests (Table 1 carries the six joint architectures); training protocol shared across arms; port notes for the official Perceiver-IO and the three expert-routed baselines.

Training protocol: AdamW (learning rate 7×10^{-4}, weight decay 10^{-4}, gradient clip 1.0, label smoothing 0.1), batch size 64, two-epoch linear warmup ($0.1 \rightarrow 1.0$) followed by a constant rate; a 20-epoch base schedule with exhaustive-search extension (minimum 40 epochs, patience 30). The objective is classification (weight 1.0) plus a reconstruction auxiliary (0.1), with uniform task weights $w_d = 1$; the grouped fused loader interleaves per-dataset batches so every dataset is visited each epoch. T01 uses $d_{\text{model}} = 248$, a 12-slot $\times$ 8-head read, K=3 refinement steps (step size 0.05), gate-α 0.8, and per-task frozen ETF prototype sets (learned temperature, initialized at 1.0); ETF-plain keeps $d_{\text{model}} = 256$.

Table 7: Train/test gap under the diagnostic protocol (capped eval, cap 2000; seed-42 checkpoints). Bucket counts give the number of datasets carrying each failure label; labels are independent, so a dataset may appear in both counts or neither.

Checkpoint	Train acc	Test acc	Gap	Rep-failure datasets	Memorization datasets
views / T01	0.763	0.632	+0.131	7	28
flatten / T01	0.535	0.492	+0.043	26	10
hilbert / T01	0.404	0.388	**+0.016**	32	4
hilbert / MMoE	0.706	0.586	+0.121	12	19
hilbert / PLE	0.724	0.596	+0.128	10	21
hilbert / Perceiver-IO	0.668	0.499	+0.169	14	25

Table 8: Expert cross-serialization comparisons: two-sided Wilcoxon signed-rank p values over the 54 datasets for the nine architecture $\times$ world-line pairs. None survives Bonferroni correction (minimum adjusted $p = 0.078$).

architecture	views–flatten	views–hilbert	flatten–hilbert
MMoE	8.7×10^{-3}	1.1×10^{-1}	2.1×10^{-2}
PLE	4.0×10^{-2}	7.0×10^{-1}	3.8×10^{-1}
Cross-Stitch	3.7×10^{-1}	9.2×10^{-1}	9.7×10^{-2}

B.1 The operators under test

Serialization. A point cloud $P = \{p_i\}$ is consumed by a permutation-sensitive encoder as a sequence, so a world line is a choice of serialization π: an ordering, or a rendering. The flatten line keeps storage order; views replaces the set by six orthographic depth projections pooled to three channels; hilbert sorts by space-filling-curve key, $\pi = \mathrm{argsort}\ H(\lfloor Q\, p_i \rfloor)$, with H the Hilbert map (Skilling, 2004). The choice of π over the four point-cloud datasets is the only manipulated variable in the study.

Optimizer-side operators reshape the joint update $\Delta\theta = -\eta \sum_d w_d\, \tilde{g}_d$ after the trunk has already mixed per-task gradients:

- **Gradient surgery** (PCGrad) (Yu et al., 2020): for every conflicting pair with $\cos(g_d, g_e) < 0$, project each gradient onto the normal plane of the other, $\tilde{g}_d = g_d - \frac{\langle g_d, g_e \rangle}{\|g_e\|^2}\, g_e$.
- **Dynamic loss reweighting** (Liu et al., 2019): $w_d(t) \propto \exp(r_d(t-1)/T)$ with r_d the per-task loss ratio across consecutive steps.

Architecture-side isolation decides where tasks may write at all:

- **Expert routing** (MoE, MMoE, PLE) (Shazeer et al., 2017; Ma et al., 2018; Tang et al., 2020): $y = \sum_i G_i(x)\, E_i(x)$ with G a (per-task) gated mixture; each task gradient is confined to the experts its gate opens.
- **Cross-stitch** (Misra et al., 2016): task activations are mixed only through learned scalars, $h' = \alpha\, h_A + (1-\alpha)\, h_B$.
- **Latent bottleneck** (Perceiver-IO) (Jaegle et al., 2022): inputs cross-attend onto a fixed latent array Z; tasks share the $|Z|$-wide activation bottleneck rather than the input-pathway weights.

C Per-dataset results

The complete 54-dataset $\times$ 18-cell accuracy matrix (mean $\pm$ std) backing Table 2 appears as Tables 9–11; the full specialist comparison backing §4.2 is Table 12.

Additional per-dataset views $\rightarrow$ hilbert drops among the fifty untouched datasets: kmnist 0.6656 $\rightarrow$ 0.2252 (-44.0 points) and gtsrb 0.5143 $\rightarrow$ 0.1008 (-41.4). On the milder flatten $\rightarrow$ hilbert pair, single-dataset drops reach -39.6 points (digits, 0.8357 $\rightarrow$ 0.4398). The reverse contrast is equally diagnostic: on byte-identical text datasets the views line beats flatten by up to $+40.5$ points (dbpedia14, 0.8380 versus 0.4329), reading as the flatten line's semi-collapse taxing the whole mixture rather than a views-side anomaly (§4.5).

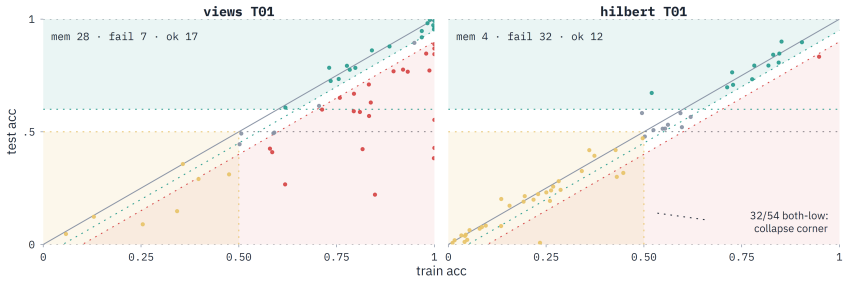

Figure 8: The failure is representational, not memorization: the collapsed hilbert/T01 trunk shows a +0.016 train/test gap with 32 of 54 datasets in the representation-failure bucket and only 4 of 54 in the memorization bucket, while routed and bottlenecked models recover ordinary generalization gaps (+0.121 to +0.169). Points are 108 of the 324 specialist runs (worst bucket per dataset–arm); canary marks both-low runs, red a train–test gap above 10 pp, teal healthy.

Specialists come in two size tiers: small, floored at max(40, the smallest multiple of 8 covering the class count), and large, fixed width 256; both tiers appear in Table 12. Point-cloud specialist arms train on the storage-order (flatten) serialization.

Against the flatten/ETF-plain joint cell, the median specialist gain is +9.25 points; across the 48 datasets where specialists win, the mean gain is +16.86 points.

D Diagnostic probes

P1/P2/P3 protocols (including the P1 capped-sample convention: loader-order truncation, train split through the augmentation pipeline), per-dataset probe values, bucket definitions and full bucket membership, and the modality-block cosine matrices.

Aggregating the P3 cosine matrix into 7×7 modality blocks, the per-modality outward couplings are: vision $+0.824$, text $+0.798$, audio $+0.836$, time series $+0.847$, tabular $+0.859$, and graph $+0.865$.

D.1 P1: generalization-gap taxonomy

P1: train and test accuracy on capped samples of at most 2,000 examples per dataset per split, in loader order without stratification; the train split passes through the training-time augmentation pipeline, so every train–test gap is a conservative read of memorization.

We place the collapse on the classical generalization-gap taxonomy: for each checkpoint we measure train and test accuracy under the same capped protocol and label each dataset with two independent failure tests, memorization (train–test gap above 10 points) and representation failure (train and test both below 50%); a gap within 5 points with test accuracy of at least 60% counts as healthy. The labels are not a partition: a dataset can carry both, or neither.

The bucket counts are collected in Table 7 (Supplementary Tables).

The hilbert/T01 trunk cannot fit its own training data: a +0.016 gap means the model is not overfitting at all, and 32 of 54 datasets sit in the representation-failure bucket with both splits below 50%. Pure memorization is essentially absent: only 4 of 54 datasets show a memorization-scale gap, and 3 of the 4 also fail both splits. Views shows an ordinary generalization gap (+0.131) in the same architecture; routed models on the same hostile data recover ordinary gap magnitudes (+0.121 to +0.128), memorization again the dominant failure label; the latent-bottleneck model shows the largest mean gap of all (+0.169, 25 datasets in the memorization bucket). Inside the collapsed checkpoint the pattern is uniform, not concentrated: hilbert/T01's largest per-modality gaps are tabular (+0.029), time series (+0.032), and point cloud (+0.037), ordinary magnitudes everywhere, with vision at +0.002: a model that has memorized nothing because it has learned nothing. The taxonomy inverts cleanly: sharing under pollution produces a model that *cannot learn*; routing restores an ordinary generalization gap with memorization dominant; a hard bottleneck shifts the failure mass furthest toward *ordinary overfitting*. The debt lives in the representation, exactly where §4.5's rank-one gradient writes it.

Table 9: Per-dataset test accuracy (mean ± sample std, three seeds, full-test protocol), flatten world line. Means and standard deviations are computed from full-precision values and rounded at display; recomputation from rounded per-seed values may differ by one unit in the last digit.

Dataset	etf-plain	mmoe	perc	ple	t01	xst
agnews	0.3039±.0040	0.3632±.0033	0.5259±.0251	0.3439±.0071	0.5460±.0324	0.3354±.0096
breast_cancer	0.8860±.0179	0.8912±.0220	0.9350±.0181	0.8756±.0165	0.9085±.0165	0.8790±.0170
cifar10	0.3204±.0174	0.4877±.0050	0.3127±.0080	0.5019±.0096	0.3423±.0104	0.4624±.0068
cifar100	0.0437±.0025	0.1790±.0037	0.0817±.0040	0.1943±.0011	0.0501±.0003	0.1639±.0058
cola	0.6975±.0000	0.6895±.0069	0.6123±.0105	0.6886±.0074	0.6972±.0005	0.6978±.0005
covertype	0.6207±.0474	0.7130±.0076	0.6689±.0512	0.7046±.0073	0.6125±.0093	0.6760±.0094
cremad	0.3753±.0269	0.3438±.0160	0.3162±.0163	0.3631±.0390	0.3881±.0027	0.3589±.0138
dbpedia14	0.1313±.0121	0.2694±.0099	0.5311±.0141	0.2625±.0108	0.4329±.0748	0.2193±.0097
dd	0.6061±.0147	0.5906±.0032	0.5788±.0000	0.5858±.0087	0.6033±.0272	0.5561±.0306
digits	0.7970±.0364	0.9637±.0066	0.8479±.0128	0.9753±.0115	0.8357±.0327	0.9651±.0063
dna_enhancers	0.5563±.0249	0.5811±.0134	0.6501±.0109	0.5671±.0037	0.6393±.0141	0.5732±.0129
dtd	0.0906±.0074	0.1609±.0030	0.0950±.0031	0.1666±.0049	0.0919±.0033	0.1527±.0061
ecg200	0.8315±.0437	0.8970±.0131	0.7830±.0413	0.8924±.0102	0.8646±.0344	0.8845±.0222
ecg5000	0.9085±.0052	0.9301±.0017	0.8819±.0152	0.9267±.0043	0.9152±.0034	0.9320±.0021
electricdevices	0.5586±.0092	0.6410±.0125	0.5655±.0153	0.6342±.0118	0.5662±.0128	0.6388±.0140
emnist_letters	0.3080±.0277	0.7306±.0047	0.2112±.0489	0.7233±.0138	0.3422±.0470	0.6625±.0125
enzymes	0.2786±.0507	0.4475±.0257	0.3932±.0291	0.4334±.0355	0.2675±.0537	0.4684±.0192
esc50	0.1097±.0093	0.2150±.0093	0.0699±.0186	0.2299±.0124	0.1112±.0124	0.2247±.0101
eurosat	0.6172±.0169	0.7917±.0162	0.5580±.0718	0.8095±.0052	0.6379±.0090	0.7539±.0191
flowers102	0.1477±.0028	0.2920±.0072	0.1198±.0183	0.3011±.0108	0.1382±.0105	0.2495±.0075
fmnist	0.6437±.0137	0.7929±.0022	0.6243±.0466	0.7973±.0074	0.6347±.0337	0.7679±.0023
food101	0.0247±.0006	0.0779±.0036	0.0411±.0055	0.0924±.0017	0.0278±.0033	0.0715±.0027
forda	0.6420±.0706	0.9148±.0019	0.5023±.0208	0.9113±.0136	0.7026±.0325	0.8768±.0265
fordb	0.6475±.0224	0.7418±.0120	0.5241±.0160	0.7280±.0400	0.6700±.0454	0.7248±.0291
gtsrb	0.1814±.0092	0.4081±.0022	0.1256±.0076	0.4380±.0170	0.1872±.0100	0.3538±.0092
gtzan	0.3841±.0551	0.4128±.0421	0.3060±.0679	0.4635±.0418	0.3385±.0515	0.3919±.0200
imdb	0.5063±.0058	0.5099±.0032	0.5702±.0222	0.5079±.0003	0.5740±.0049	0.5085±.0052
iris	0.7778±.0509	0.6556±.0839	0.9111±.0192	0.7000±.0882	0.8444±.0192	0.6778±.0385
kmnist	0.3734±.0122	0.7265±.0089	0.3688±.0661	0.7283±.0168	0.3881±.0248	0.6644±.0199
letter	0.3360±.0337	0.8557±.0053	0.7761±.0143	0.8534±.0108	0.3791±.0645	0.8032±.0191
magic04	0.7732±.0180	0.8439±.0064	0.8212±.0190	0.8428±.0036	0.7842±.0314	0.8262±.0129
mnist	0.7336±.0214	0.9121±.0014	0.4321±.0985	0.9177±.0070	0.7877±.0298	0.8705±.0057
mutag	0.8289±.0132	0.8333±.0152	0.8333±.0152	0.8333±.0152	0.8509±.0304	0.8509±.0152
news20	0.0651±.0023	0.0845±.0029	0.1756±.0079	0.0826±.0066	0.1497±.0069	0.0805±.0002
nsynth	0.3258±.0108	0.5885±.0103	0.3500±.0104	0.5948±.0088	0.3569±.0269	0.6082±.0026
oxford_pets	0.0657±.0067	0.1141±.0114	0.0749±.0046	0.1247±.0041	0.0683±.0014	0.1012±.0035
proteins	0.7596±.0074	0.7653±.0123	0.7471±.0116	0.7563±.0120	0.7694±.0091	0.7517±.0170
ravdess	0.3403±.1007	0.3030±.0290	0.1871±.0106	0.2869±.0209	0.3739±.0156	0.2817±.0184
shuttle	0.9392±.0203	0.9831±.0006	0.9952±.0012	0.9851±.0018	0.9416±.0214	0.9763±.0092
speechcmd	0.0777±.0073	0.3988±.0098	0.0759±.0042	0.4503±.0109	0.0897±.0042	0.4211±.0112
sst2	0.4966±.0134	0.5099±.0262	0.5382±.0118	0.5207±.0123	0.5297±.0159	0.5129±.0108
starlightcurves	0.8756±.0562	0.9640±.0108	0.8428±.0013	0.9683±.0076	0.9043±.0187	0.9600±.0087
stl10	0.3649±.0263	0.5150±.0106	0.3202±.0010	0.5220±.0063	0.4012±.0079	0.4914±.0053
svhn	0.2031±.0089	0.3557±.0033	0.1952±.0022	0.3824±.0107	0.2041±.0091	0.2985±.0204
trec6	0.4371±.0727	0.5787±.0130	0.6976±.0191	0.5901±.0139	0.6949±.0254	0.5667±.0131
urbansound8k	0.3455±.0357	0.5450±.0190	0.3327±.0300	0.5638±.0266	0.3683±.0241	0.5566±.0096
uwgesture	0.8659±.0188	0.9473±.0020	0.6008±.0203	0.9498±.0040	0.8735±.0050	0.9519±.0021
wafer	0.9679±.0093	0.9937±.0010	0.9372±.0166	0.9948±.0010	0.9811±.0061	0.9940±.0016
wine	0.8981±.0424	0.9444±.0556	0.9444±.0278	0.9167±.0278	0.8889±.0000	0.8981±.0160
yelp_full	0.2308±.0040	0.2186±.0044	0.2834±.0092	0.2211±.0019	0.2884±.0062	0.2242±.0048
modelnet40	0.0621±.0141	0.3687±.0147	0.3944±.0259	0.3752±.0262	0.1046±.0327	0.3740±.0132
scanobjectnn	0.1321±.0014	0.1963±.0059	0.3287±.0158	0.1915±.0112	0.1606±.0172	0.1960±.0032
modelnet10	0.2691±.0053	0.6749±.0209	0.5464±.0370	0.6652±.0172	0.3477±.0412	0.6760±.0168
shapenet16	0.4021±.0244	0.7963±.0128	0.8258±.0034	0.7910±.0089	0.5270±.0812	0.7968±.0094

Table 10: Per-dataset test accuracy (mean ± sample std, three seeds, full-test protocol), hilbert world line. Means and standard deviations are computed from full-precision values and rounded at display; recomputation from rounded per-seed values may differ by one unit in the last digit.

Dataset	etf-plain	mmoe	perc	ple	t01	xst
agnews	0.2560±.0135	0.3521±.0086	0.5230±.0164	0.3456±.0266	0.3555±.0354	0.3357±.0095
breast_cancer	0.7951±.0140	0.8768±.0089	0.9350±.0105	0.8849±.0114	0.7681±.0189	0.8768±.0089
cifar10	0.2318±.0061	0.4734±.0054	0.3090±.0153	0.4912±.0049	0.2512±.0331	0.4555±.0066
cifar100	0.0146±.0058	0.1777±.0067	0.0840±.0111	0.1875±.0115	0.0199±.0037	0.1659±.0036
cola	0.6975±.0000	0.6972±.0005	0.6134±.0311	0.6941±.0032	0.6969±.0011	0.6954±.0030
covertype	0.4968±.0093	0.6848±.0243	0.6955±.0094	0.7123±.0052	0.5349±.0840	0.6840±.0097
cremad	0.2642±.0193	0.3621±.0198	0.3278±.0252	0.3725±.0212	0.2419±.0170	0.3607±.0016
dbpedia14	0.0868±.0017	0.2352±.0190	0.5349±.0091	0.2578±.0142	0.2133±.0253	0.2196±.0129
dd	0.5820±.0055	0.5787±.0105	0.5788±.0000	0.5724±.0514	0.5782±.0010	0.5426±.0472
digits	0.4786±.0313	0.9625±.0121	0.8635±.0178	0.9762±.0013	0.4398±.0904	0.9656±.0087
dna_enhancers	0.5025±.0087	0.5865±.0054	0.6508±.0041	0.5684±.0161	0.6025±.0346	0.5816±.0082
dtd	0.0430±.0064	0.1533±.0043	0.0907±.0098	0.1631±.0058	0.0478±.0087	0.1389±.0017
ecg200	0.7785±.0202	0.8840±.0170	0.7645±.0322	0.8759±.0078	0.7915±.0266	0.8866±.0112
ecg5000	0.8917±.0135	0.9335±.0012	0.8884±.0097	0.9264±.0028	0.8563±.0373	0.9312±.0026
electricdevices	0.4162±.0284	0.6304±.0143	0.5835±.0198	0.6407±.0162	0.3712±.0274	0.6345±.0063
emnist_letters	0.0908±.0092	0.6596±.0390	0.1997±.0379	0.6804±.0313	0.1173±.0538	0.6756±.0087
enzymes	0.2414±.0209	0.4219±.0236	0.3594±.0241	0.4271±.0081	0.1923±.0139	0.4475±.0084
esc50	0.0402±.0238	0.2135±.0161	0.0826±.0102	0.2307±.0239	0.0409±.0087	0.2113±.0217
eurosat	0.3917±.0393	0.7551±.0288	0.5774±.0318	0.8012±.0033	0.4538±.0742	0.7730±.0066
flowers102	0.0250±.0097	0.2714±.0076	0.1218±.0217	0.2927±.0039	0.0223±.0095	0.2397±.0075
fmnist	0.4124±.0196	0.7649±.0178	0.6114±.0370	0.7846±.0182	0.4374±.0917	0.7676±.0075
food101	0.0110±.0004	0.0818±.0045	0.0387±.0041	0.0973±.0010	0.0108±.0010	0.0718±.0008
forda	0.4969±.0106	0.9025±.0185	0.5082±.0235	0.9028±.0166	0.5088±.0174	0.8862±.0227
fordb	0.5159±.0094	0.7645±.0068	0.5271±.0053	0.7346±.0085	0.5346±.0234	0.7416±.0097
gtsrb	0.0914±.0077	0.3890±.0143	0.1241±.0019	0.4168±.0033	0.1008±.0181	0.3599±.0051
gtzan	0.1309±.0512	0.4141±.0583	0.2969±.0396	0.4023±.0396	0.1667±.0251	0.4141±.0423
imdb	0.5045±.0009	0.5097±.0020	0.5678±.0132	0.5076±.0010	0.5154±.0078	0.5077±.0020
iris	0.7389±.0585	0.7222±.0192	0.9222±.0192	0.7667±.0667	0.6389±.0585	0.6889±.1018
kmnist	0.2071±.0215	0.6462±.0289	0.3570±.0513	0.6724±.0466	0.2252±.0366	0.6608±.0132
letter	0.0972±.0205	0.7954±.0375	0.7805±.0194	0.8217±.0343	0.0826±.0431	0.8149±.0128
magic04	0.7089±.0146	0.8259±.0209	0.8317±.0084	0.8375±.0149	0.7094±.0158	0.8358±.0075
mnist	0.2766±.0155	0.8789±.0182	0.3819±.0437	0.8878±.0220	0.4077±.1471	0.8719±.0084
mutag	0.8158±.0263	0.8421±.0263	0.8596±.0152	0.8070±.0152	0.7368±.0696	0.8509±.0152
news20	0.0549±.0069	0.0868±.0024	0.1785±.0091	0.0816±.0082	0.0838±.0062	0.0751±.0052
nsynth	0.2344±.0223	0.5823±.0328	0.3565±.0241	0.5982±.0185	0.2209±.0503	0.5938±.0076
oxford_pets	0.0343±.0041	0.1145±.0080	0.0812±.0024	0.1215±.0038	0.0379±.0072	0.1039±.0054
proteins	0.7548±.0178	0.7611±.0137	0.7469±.0101	0.7547±.0090	0.7628±.0081	0.7519±.0105
ravdess	0.2094±.0455	0.2756±.0421	0.1753±.0287	0.2999±.0588	0.2331±.0326	0.2865±.0138
shuttle	0.8272±.0605	0.9795±.0025	0.9953±.0012	0.9810±.0045	0.8173±.0551	0.9809±.0028
speechcmd	0.0503±.0035	0.4053±.0197	0.0805±.0110	0.4548±.0058	0.0544±.0039	0.4305±.0119
sst2	0.5060±.0042	0.5084±.0003	0.5408±.0121	0.5013±.0042	0.5060±.0068	0.5048±.0087
starlightcurves	0.8558±.0032	0.9621±.0024	0.8396±.0117	0.9661±.0037	0.8421±.0068	0.9625±.0017
stl10	0.2515±.0095	0.5090±.0041	0.3305±.0096	0.5178±.0039	0.2749±.0223	0.4855±.0042
svhn	0.1849±.0109	0.3283±.0224	0.1911±.0031	0.3712±.0172	0.1920±.0064	0.3024±.0104
trec6	0.3139±.1075	0.5552±.0359	0.6707±.0269	0.5865±.0062	0.5531±.0312	0.5304±.0088
urbansound8k	0.2189±.0123	0.5187±.0511	0.3643±.0181	0.5735±.0238	0.2545±.0107	0.5604±.0206
uwgesture	0.4970±.0129	0.9468±.0057	0.5702±.0291	0.9524±.0069	0.5049±.0549	0.9479±.0028
wafer	0.9064±.0247	0.9940±.0014	0.9338±.0240	0.9940±.0018	0.9116±.0337	0.9946±.0015
wine	0.8657±.0765	0.9167±.0278	0.9074±.0321	0.9352±.0160	0.8241±.0642	0.9167±.0000
yelp_full	0.2263±.0059	0.2281±.0034	0.2837±.0049	0.2228±.0050	0.2468±.0042	0.2260±.0019
modelnet40	0.1203±.0195	0.7164±.0255	0.5702±.0045	0.7276±.0075	0.1072±.0371	0.7041±.0141
scanobjectnn	0.1593±.0133	0.3311±.0090	0.3503±.0106	0.3316±.0117	0.1548±.0165	0.3074±.0154
modelnet10	0.3479±.0282	0.8514±.0097	0.5852±.0184	0.8532±.0053	0.3153±.0500	0.8387±.0182
shapenet16	0.5363±.1132	0.9565±.0049	0.8910±.0147	0.9605±.0024	0.5785±.0457	0.9557±.0015

Table 11: Per-dataset test accuracy (mean ± sample std, three seeds, full-test protocol), views world line. Means and standard deviations are computed from full-precision values and rounded at display; recomputation from rounded per-seed values may differ by one unit in the last digit.

Dataset	etf-plain	mmoe	perc	ple	t01	xst
agnews	0.3495±.0022	0.3461±.0091	0.5381±.0079	0.3328±.0267	0.7301±.0106	0.3368±.0075
breast_cancer	0.9053±.0135	0.8886±.0329	0.9131±.0515	0.8823±.0306	0.8927±.0159	0.8960±.0115
cifar10	0.5068±.0011	0.4705±.0107	0.2879±.0180	0.4921±.0109	0.4880±.0072	0.4598±.0074
cifar100	0.1250±.0034	0.1724±.0040	0.0711±.0091	0.1851±.0072	0.1230±.0056	0.1607±.0009
cola	0.6980±.0005	0.6941±.0037	0.6441±.0425	0.6932±.0066	0.6200±.0381	0.6960±.0019
covertype	0.7263±.0060	0.6876±.0320	0.6801±.0232	0.6874±.0311	0.7287±.0015	0.6821±.0116
cremad	0.4177±.0191	0.3565±.0110	0.2845±.0608	0.3630±.0150	0.4128±.0209	0.3615±.0161
dbpedia14	0.2778±.0089	0.2479±.0172	0.5371±.0118	0.2478±.0175	0.8380±.0048	0.2141±.0111
dd	0.5689±.0327	0.5842±.0120	0.5788±.0000	0.5830±.0263	0.5458±.0410	0.5690±.0099
digits	0.9589±.0055	0.9573±.0108	0.8349±.0432	0.9729±.0041	0.9699±.0035	0.9719±.0028
dna_enhancers	0.5668±.0103	0.5792±.0091	0.6588±.0018	0.5764±.0048	0.6334±.0034	0.5873±.0047
dtd	0.1624±.0030	0.1439±.0013	0.0850±.0052	0.1555±.0026	0.1473±.0090	0.1425±.0004
ecg200	0.8793±.0168	0.8958±.0078	0.7509±.0031	0.8892±.0079	0.8834±.0229	0.8759±.0139
ecg5000	0.9283±.0040	0.9335±.0007	0.8868±.0121	0.9274±.0031	0.9350±.0027	0.9298±.0041
electricdevices	0.6231±.0092	0.6343±.0131	0.5610±.0077	0.6299±.0134	0.6271±.0039	0.6405±.0086
emnist_letters	0.6622±.0109	0.6902±.0442	0.1534±.0251	0.6954±.0350	0.6491±.0139	0.6779±.0093
enzymes	0.4098±.0300	0.4308±.0179	0.3702±.0281	0.4092±.0209	0.4144±.0182	0.4386±.0293
esc50	0.2504±.0359	0.2128±.0157	0.0692±.0000	0.2121±.0232	0.2467±.0344	0.2091±.0229
eurosat	0.7774±.0006	0.7486±.0279	0.5194±.0359	0.7667±.0210	0.7642±.0017	0.7209±.0278
flowers102	0.2851±.0082	0.2557±.0212	0.1096±.0073	0.2848±.0122	0.2650±.0151	0.2365±.0092
fmnist	0.7896±.0101	0.7816±.0191	0.5763±.0192	0.7858±.0188	0.7903±.0130	0.7710±.0088
food101	0.0723±.0050	0.0749±.0033	0.0376±.0025	0.0874±.0034	0.0653±.0057	0.0669±.0042
forda	0.9137±.0011	0.9131±.0150	0.5140±.0020	0.9031±.0094	0.9103±.0055	0.8961±.0115
fordb	0.7565±.0107	0.7508±.0055	0.5274±.0101	0.7382±.0111	0.7488±.0048	0.7326±.0409
gtsrb	0.5604±.0195	0.3926±.0273	0.1240±.0111	0.4277±.0257	0.5143±.0225	0.3441±.0045
gtzan	0.4661±.0421	0.4023±.0135	0.2786±.0495	0.4089±.0418	0.4408±.0584	0.4102±.0341
imdb	0.5154±.0019	0.5092±.0046	0.5749±.0084	0.5101±.0014	0.6603±.0107	0.5114±.0013
iris	0.7389±.0255	0.6444±.0385	0.9000±.0000	0.6889±.0192	0.7667±.0577	0.6556±.0192
kmnist	0.6787±.0127	0.6862±.0645	0.2883±.0254	0.6997±.0420	0.6656±.0079	0.6771±.0083
letter	0.7727±.0026	0.8333±.0354	0.7759±.0302	0.8284±.0340	0.7715±.0137	0.8122±.0109
magic04	0.8440±.0044	0.8365±.0062	0.8216±.0164	0.8307±.0107	0.8426±.0089	0.8287±.0156
mnist	0.9046±.0013	0.8908±.0289	0.3010±.0037	0.9016±.0177	0.9041±.0101	0.8774±.0052
mutag	0.8596±.0608	0.8333±.0152	0.8421±.0263	0.8684±.0000	0.8684±.0263	0.8421±.0000
news20	0.0864±.0032	0.0851±.0084	0.1864±.0083	0.0827±.0028	0.3042±.0038	0.0773±.0033
nsynth	0.5984±.0174	0.5846±.0248	0.3355±.0298	0.5868±.0131	0.5872±.0230	0.5853±.0181
oxford_pets	0.1231±.0049	0.1040±.0023	0.0796±.0054	0.1238±.0014	0.1141±.0051	0.1005±.0034
proteins	0.7537±.0152	0.7548±.0097	0.7599±.0037	0.7652±.0141	0.7412±.0154	0.7572±.0176
ravdess	0.3783±.0127	0.2882±.0040	0.1697±.0419	0.3021±.0226	0.4206±.0036	0.2756±.0137
shuttle	0.9782±.0062	0.9828±.0021	0.9948±.0004	0.9836±.0012	0.9805±.0043	0.9818±.0014
speechcmd	0.4156±.0004	0.4119±.0242	0.0770±.0160	0.4537±.0082	0.3882±.0189	0.4364±.0162
sst2	0.5327±.0073	0.5161±.0180	0.5361±.0082	0.4978±.0193	0.6265±.0112	0.5084±.0105
starlightcurves	0.9622±.0013	0.9597±.0123	0.8436±.0016	0.9568±.0101	0.9657±.0020	0.9598±.0072
stl10	0.5220±.0088	0.4996±.0058	0.3136±.0170	0.5105±.0053	0.5076±.0100	0.4820±.0074
svhn	0.4730±.0221	0.3537±.0149	0.1937±.0023	0.3870±.0295	0.4463±.0199	0.3071±.0020
trec6	0.6481±.0202	0.5911±.0257	0.6490±.0469	0.5836±.0124	0.7976±.0141	0.5265±.0480
urbansound8k	0.5830±.0454	0.5540±.0200	0.3628±.0238	0.5503±.0208	0.5825±.0306	0.5609±.0207
uwgesture	0.9493±.0103	0.9452±.0057	0.5564±.0386	0.9498±.0034	0.9469±.0057	0.9475±.0058
wafer	0.9923±.0015	0.9939±.0007	0.9233±.0289	0.9931±.0010	0.9941±.0011	0.9933±.0021
wine	0.9259±.0160	0.9259±.0160	0.8611±.0000	0.9444±.0278	0.9167±.0278	0.9074±.0160
yelp_full	0.2269±.0028	0.2246±.0037	0.2966±.0023	0.2246±.0048	0.3738±.0094	0.2257±.0046
modelnet40	0.6564±.0080	0.7471±.0089	0.6221±.0421	0.7323±.0099	0.6481±.0064	0.7418±.0157
scanobjectnn	0.3774±.0061	0.4255±.0212	0.3952±.0114	0.4236±.0059	0.3805±.0053	0.4134±.0192
modelnet10	0.8346±.0179	0.8500±.0069	0.7816±.0110	0.8503±.0133	0.8299±.0129	0.8524±.0057
shapenet16	0.9472±.0012	0.9616±.0017	0.9292±.0150	0.9664±.0037	0.9447±.0046	0.9633±.0014

Table 12: Per-dataset specialists versus the flatten/ETF-plain joint cell (two size tiers, capped protocol; Δ against the joint cell's full-test per-dataset accuracy).

Dataset	Specialist small	Specialist large	Joint cell	Δ small	Δ large
agnews	0.3582±.0141	0.2501±.0015	0.3039	+0.0543	-0.0538
breast_cancer	0.8965±.0098	0.9046±.0111	0.8860	+0.0104	+0.0185
cifar10	0.4631±.0153	0.5425±.0095	0.3204	+0.1427	+0.2221
cifar100	0.1971±.0056	0.2065±.0170	0.0437	+0.1533	+0.1628
cola	0.6975	0.6975	0.6975	+0.0000	+0.0000
covertype	0.7246±.0072	0.6600±.0496	0.6207	+0.1039	+0.0393
cremad	0.4206±.0078	0.3948±.0162	0.3753	+0.0453	+0.0194
dbpedia14	0.3590±.0157	0.3094±.0790	0.1313	+0.2277	+0.1781
dd	0.5556±.0408	0.5495±.0318	0.6061	-0.0504	-0.0566
digits	0.9743±.0066	0.9772±.0010	0.7970	+0.1773	+0.1801
dna_enhancers	0.5039±.0135	0.4961±.0135	0.5563	-0.0524	-0.0602
dtd	0.1557±.0048	0.1746±.0028	0.0906	+0.0651	+0.0840
ecg200	0.8765±.0288	0.8689±.0060	0.8315	+0.0450	+0.0375
ecg5000	0.9301±.0049	0.9228±.0096	0.9085	+0.0216	+0.0143
electricdevices	0.6206±.0188	0.6127±.0396	0.5586	+0.0620	+0.0541
emnist_letters	0.7116±.0302	0.7883±.0197	0.3080	+0.4036	+0.4803
enzymes	0.4040±.0523	0.3799±.0260	0.2786	+0.1254	+0.1012
esc50	0.2753±.0158	0.2846±.0022	0.1097	+0.1656	+0.1749
eurosat	0.8287±.0071	0.8729±.0100	0.6172	+0.2115	+0.2557
flowers102	0.3903±.0027	0.4138±.0078	0.1477	+0.2426	+0.2661
fmnist	0.8245±.0101	0.8431±.0237	0.6437	+0.1809	+0.1994
food101	0.1461±.0146	0.1432±.0017	0.0247	+0.1215	+0.1185
forda	0.9193±.0119	0.5143±.0000	0.6420	+0.2773	-0.1278
fordb	0.7620±.0091	0.7730±.0184	0.6475	+0.1146	+0.1256
gtsrb	0.6825±.0707	0.8180±.0219	0.1814	+0.5010	+0.6366
gtzan	0.4093±.0303	0.4657±.0281	0.3841	+0.0252	+0.0815
imdb	0.4961±.0135	0.4961±.0135	0.5063	-0.0102	-0.0102
iris	0.7083±.1063	0.7292±.0836	0.7778	-0.0694	-0.0486
kmnist	0.7199±.0312	0.8328±.0037	0.3734	+0.3464	+0.4594
letter	0.8607±.0083	0.8797±.0078	0.3360	+0.5247	+0.5437
magic04	0.8483±.0024	0.7575±.0953	0.7732	+0.0751	-0.0157
mnist	0.9034±.0332	0.9369±.0145	0.7336	+0.1698	+0.2032
modelnet10	0.6875±.0032	0.6650±.0277	0.2691	+0.4184	+0.3959
modelnet40	0.4006±.0035	0.3458±.0216	0.0621	+0.3385	+0.2837
mutag	0.8257±.0285	0.8268±.0190	0.8289	-0.0033	-0.0022
news20	0.0861±.0033	0.0844±.0088	0.0651	+0.0209	+0.0192
nsynth	0.5927±.0325	0.6297±.0153	0.3258	+0.2669	+0.3039
oxford_pets	0.1326±.0117	0.1641±.0089	0.0657	+0.0669	+0.0984
proteins	0.7624±.0108	0.7663±.0168	0.7596	+0.0028	+0.0068
ravdess	0.3902±.0313	0.3652±.0196	0.3403	+0.0499	+0.0249
scanobjectnn_flatten	0.1933±.0096	0.1656±.0289	0.1321	+0.0611	+0.0335
shapenet16	0.8571±.0051	0.8172±.0638	0.4021	+0.4550	+0.4150
shuttle	0.9787±.0013	0.9817±.0014	0.9392	+0.0395	+0.0425
speechcmd	0.5890±.0379	0.6476±.0234	0.0777	+0.5112	+0.5699
sst2	0.5092±.0004	0.5089	0.4966	+0.0126	+0.0124
starlightcurves	0.9565±.0209	0.9540±.0054	0.8756	+0.0809	+0.0784
stl10	0.5131±.0163	0.5500±.0051	0.3649	+0.1482	+0.1851
svhn	0.6113±.0305	0.5232±.2869	0.2031	+0.4082	+0.3201
trec6	0.6849±.0580	0.7006±.0162	0.4371	+0.2478	+0.2636
urbansound8k	0.5360±.0292	0.5133±.0114	0.3455	+0.1905	+0.1678
uwgesture	0.9470±.0092	0.9455±.0044	0.8659	+0.0811	+0.0796
wafer	0.9916±.0034	0.9921±.0012	0.9679	+0.0237	+0.0242
wine	0.9618±.0151	0.9398±.0080	0.8981	+0.0637	+0.0417
yelp_full	0.2432±.0052	0.2384±.0023	0.2308	+0.0124	+0.0076

Bucket membership for the collapsed hilbert/T01 checkpoint (non-exclusive buckets): *representation failure* (train and test both below 0.5; 32 datasets): agnews, cifar10, cifar100, cremad, dbpedia14, digits, dtd, electricdevices, emnist_letters, enzymes, esc50, eurosat, flowers102, fmnist, food101, gtsrb, gtzan, kmnist, letter, mnist, modelnet10, modelnet40, news20, nsynth, oxford_pets, ravdess, scanobjectnn, speechcmd, stl10, svhn, urbansound8k, yelp_full. *Memorization* (gap above 0.10; 4): dbpedia14, electricdevices, modelnet10, wine. *Healthy* (gap at most 0.05, test at least 0.6; 12): breast_cancer, cola, ecg200, ecg5000, imdb, iris, magic04, mutag, proteins, shuttle, starlightcurves, wafer.

E GLASS BOX

Complete protocol of the four-modality late-concat glass box (ground-truth generative model; per-modality encoders parameter-matched to within 2.4%; seed and reporting conventions, which differ from §3.1); the pre-stated hypothesis factorial (five hypotheses, exactly one supported: training-time order augmentation immunizes at zero cost); the test-time reordering reliance measurements with reliance-calibration caveats; the time-reversal positive control; and Arm F (train-on-hostile-order) results.

E.1 FINDINGS

A controlled glass box isolates how much of the effect is specific to training: a single-task toy model fusing four modalities by late concatenation, each modality's contribution measurable against known ground truth. The design is hypothesis-driven: five pre-stated hypotheses, exactly one supported: training-time order augmentation immunizes the model against test-time reordering at zero cost, the same mitigation §4.7 tests at scale. A train-on-hostile-order variant (Arm F) supplies the complementary training-time measurement. All glass-box numbers are means over three training seeds under the Appendix E protocol.

Test-time corruption is cheap: permuting a modality's token order at evaluation (an information-preserving reordering) costs only 5–21% of that modality's measured reliance, falsifying the naive hypothesis that any order corruption is catastrophic at test time. The instrument can still inflict damage when a reordering destroys structure the trained encoder exploits: time-reversing the sequential modality inflicts super-proportional harm (reliance loss ratio 1.67). Decisive for scope, training-time pollution is not automatic either: in Arm F, trained from scratch on the hostile (hilbert) ordering of one modality, final performance is statistically unchanged against the clean-order control ($F - A = +0.0021$), probe R^2 of the *other* modalities' encoders is untouched (0.997 / 0.915 / 0.842 versus 0.997 / 0.914 / 0.836), and the shared representation's effective rank is essentially unchanged (29.5 versus 28.9): the hostile ordering simply gets learned around.

The difference between the glass box and SphereBench-54 is precisely the carrier indicted in §4.5: a shared, iteratively refined trunk trained jointly on 54 tasks. Late-concat fusion gives each modality a private pathway, so a hostile ordering cannot write its gradient into other modalities' encoders; a shared iterative trunk gives every task the same weights to write into. Training-time representation debt requires shared iterative dynamics to propagate. Test-time robustness to reordering (cheap, glass box) and training-time vulnerability to hostile serialization (ruinous, shared trunk) are different predicates, and the glass box disciplines our claim: not that order corruption is universally harmful, but that under large-scale joint training with a shared trunk, it is.

F STUDY DETAILS

Budget-scan nominal-versus-realized parameter accounting (nominal tiers 310K–3M; realized counts per arm from run manifests); the seven study arms completed with uncapped evaluation (B01, B02, B03_s123, B04_s456, B07_s123, B08_s456, SE01) and the four-arm cap-bias control; sample-efficiency fractions; dose-scan construction (point-cloud datasets enter in expansion order (modelnet40 at d1, scanobjectnn at d2, modelnet10 at d3) with non-dosed point-cloud datasets removed from the mixture, so dose arms train on 51, 52, and 53 datasets and dose cells are macro-accuracy over 50 shared datasets plus the dosed point-cloud datasets); scale-family configurations and realized

parameter counts (T01-deep 3,169,997; T01-XL768 10,753,613; T01-XXL 18,270,029; MMoE-XL 8,457,469).

One budget arm (the 5.6M-parameter Perceiver-IO cell) used sequential-backward after exceeding memory under the standard fused path; the paths are arithmetically equivalent up to floating-point reassociation, and the arm is pooled without adjustment. The paired Wilcoxon test of §4.1 excludes zero differences, uses tie-averaged ranks and the normal approximation, and is two-sided. The order-augmentation arm's three per-seed macro-accuracies are 0.4920, 0.4639, and 0.4930 (gated mean $0.4829 \pm .0165$, against the hilbert anchor $0.3916 \pm .0230$).

F.1 Scale probe: per-seed detail and auxiliary scans

The 18.27M (T01-XXL) cell rests on three hilbert seeds (0.2681, 0.2680, 0.2704), on the XL768 plateau ($0.2697 \pm .0020$); the views trio holds at $0.6397 \pm .0227$, a -37.1-point gap. Views gains only +1.4 points for a $7\times$ parameter increase ($0.6363 \to 0.6506$), while hilbert deteriorates in absolute terms ($0.3916 \to 0.3110 \to 0.2697$). Per-seed clustering is tight, making the widening a property of the configuration, not sampling:

- T01-XL768 (10.75M) per seed: hilbert 0.2682, 0.2688, 0.2720; views 0.6525, 0.6509, 0.6484.
- T01-deep (3.17M) per seed: views 0.6381, 0.6398, 0.6515; hilbert 0.3392, 0.2701, 0.3235 (wide spread; the hilbert mean sits between the anchor and the XL768 plateau).
- MMoE-XL (8.46M) per seed: views 0.6079, 0.6036, 0.6018 (mean 0.6044); hilbert 0.6069, 0.6080, 0.6039 (mean 0.6063).

Two flatten-line auxiliary studies ran under the capped evaluation protocol (cap 2,000). The budget scan (3 seeds; realized parameter counts in parentheses, nominal and realized sizes differ) gives:

- ETF-plain improves with parameters: $0.5582 \pm .0079$ (341K), $0.5830 \pm .0018$ (659K), $0.6022 \pm .0026$ (3.09M).
- MMoE saturates and slightly declines: $0.5863 \pm .0044$ (647K), $0.5879 \pm .0036$ (820K), $0.5808 \pm .0050$ (3.18M).
- Perceiver-IO saturates near 0.50 even at 5.60M realized parameters, the scan's largest arm.

The sample-efficiency scan (fractions subsample the training split only, stratified, test split untouched) finds the families dead-tied at 10% of training data (ETF-plain $0.5155 \pm .0039$, MMoE $0.5156 \pm .0038$, population std), MMoE marginally ahead at 25% (0.5589 versus 0.5551), and ETF-plain back ahead at 50% (0.5813 versus 0.5721): non-monotone.

Comparability note. Cross-study levels for the dense family are indicative only: the budget and sample-efficiency arms run the capped protocol, and the flatten line sits in the semi-collapse regime where levels shift with training length (Appendix C); their ETF-plain cells therefore sit above the main-scan flatten cell by more than the bounded cap bias. The MMoE rows, by contrast, match their main-scan cells within cap bias (0.586–0.588 versus 0.5872), which is the protocol behaving as designed on a family whose regime is stable.

F.2 Dose scan: construction and per-seed detail

All dose arms use the same T01 architecture and protocol as the main scan. Dose k trains on $50 + k$ datasets: the hostile point-cloud datasets enter in the family's growth order (modelnet40 at dose one, scanobjectnn at dose two, modelnet10 at dose three), each hilbert-serialized, the remaining 50 byte-identical to the main scan; dose four is the main-scan hilbert cell itself, all four present. The d3 cell carries the study's largest seed variance (± 0.0483), consistent with regime change, not smooth degradation, and its mean overlaps the four-dose reference within seed variance. Per-seed macro-accuracies: d1 (0.6358, 0.6288, 0.6379) and d2 (0.6320, 0.6285, 0.6291) cluster within a point of the views anchor, while d3 splits (0.4255, 0.3415, 0.3420), one seed partially escaping, two not.

The threshold retroactively explains an apparent contradiction in our earlier results: in the 47-dataset era, with one to two point-cloud datasets in the mixture, the hilbert line was T01's *best* serialization

(0.6331, final-test protocol); at 54 datasets with four, it is the worst. Cross-era macro-accuracy levels are not directly comparable (task-count dilution shifts the mean), but the ordering's sign reversal is what a dose-dependent mechanism predicts, and the scan turns that retro-diction into a controlled measurement, its first two levels replaying the earlier era's point-cloud share on the current benchmark. The pollution is thus dose-dependent, a property of the training mixture rather than of any single cursed dataset, and below the threshold the dense trunk is the best jointly trained model we have: the debt is incurred crossing the threshold, not by sharing per se.

G FOLLOW-UP ARMS

G.1 HEAD TRAINABILITY: THE LEARNED-ETF CONTROL

A verbatim clone of the main-scan T01 arm with one manipulated variable: the classifier prototypes keep their simplex-ETF initialization but receive gradients (`universal_mode: etf_learned`), replacing the frozen buffer. On hilbert, the learned head scores $0.4394 \pm .0359$ against the frozen head's $0.3916 \pm .0230$ (per-seed: 0.4512, 0.3991, 0.4679): head trainability repays 4.8 points of the 24.5-point gap, leaving the trunk's entanglement as the dominant term. On views, the learned head matches the frozen one ($0.6342 \pm .0069$ versus $0.6363 \pm .0033$; per-seed 0.6340, 0.6411, 0.6275): the rigidity tax vanishes on the healthy line, completing the decomposition.

G.2 HOSTILE TEXT: A REVERSE-DIRECTION CONTROL

A flatten-cloud clone of the main scan in which eight of the nine text datasets carry a fixed byte-position permutation (same byte multiset, destroyed adjacency; news20's loader path is not permuted). The manipulation costs 2.2 points against the flatten anchors under the identical full-test protocol ($0.4818 \pm .0020$: 0.4829, 0.4829, 0.4795 against 0.4992, 0.5049, 0.5061), a mild dip without the representation-failure signature: the 54-dataset cascade is not an automatic consequence of an arbitrary hostile modality at this share, and the threshold's dependence on modality and dose share is a related open direction.

G.3 NORM EQUALIZATION: THE TASKNORM CONTROL

A verbatim clone of the T01 hilbert arm with per-task gradient normalization: each task's gradient is rescaled to the mean per-task norm before accumulation, leaving the aggregate step size unchanged. The intervention lifts hilbert from $0.3916 \pm .0230$ to $0.6288 \pm .0038$ (per-seed 0.6302, 0.6318, 0.6244), repaying 23.7 of the 24.5-point gap and leaving 0.75 points to the views line; dead datasets (test accuracy below 0.05) drop to zero. PCGrad, the projection-side intervention in the same gradient slot, repays nothing in either era (classic era: 0.6571 against 0.6587 unmodified; 54-dataset era: $0.3939 \pm .0018$ against $0.3916 \pm .0230$, Appendix G.6): equalizing norms, not resolving direction conflicts, is the active ingredient — exactly the carrier §4.5 names. The control is a static member of the gradient-norm-balancing family (Chen et al., 2018), normalizing magnitudes without learnable per-task weights; norm-dominated capture of a shared objective by one task family echoes gradient starvation (Pezeshki et al., 2021), and the norm-matched control removes that capture channel.

G.4 GRADIENT ALIGNMENT OVER TRAINING

Per-epoch gradient probes on the T01 arms (one seed per world line; diagnostic, not a seed study). On hilbert, gradient alignment stays elevated and volatile throughout training (third-of-training means 0.273, 0.220, 0.366 with per-epoch spread ± 0.30; gradient norm ratio between the largest and median task of $1.4\text{–}3.1 \times 10^3$), while on views it decays to zero (0.019, 0.005, 0.004; norm ratio 3.0–4.5). The rank-one, norm-dominated signature is a training-long property of the hostile line, not a final-checkpoint artifact — consistent with the converged-weight probe of §4.5.

G.5 A VANILLA DENSE TRUNK DOES NOT COLLAPSE

A dense vanilla transformer trunk (UniversalEncoder followed by pre-norm transformer blocks and the learned-ETF bank; 1,522,971 parameters, -1.0% against T01's 1,538,693) trained under the identical hilbert protocol does not collapse: $0.6396 \pm .0058$ (per-seed 0.6369, 0.6357, 0.6463), zero

dead datasets, matching T01's own healthy views line ($0.6363 \pm .0033$). The within-family control is starker still: the same vanilla trunk on the views ordering reaches $0.6275 \pm .0021$ (per-seed 0.6274, 0.6254, 0.6297), so its cross-serialization amplitude is +1.2 points with the sign reversed, against T01's -24.5. The collapse is therefore not a property of dense shared trunks as a class; it is specific to the spherical-iterative core that T01 and ETF-plain share.

G.6 Direction surgery: the 54-dataset PCGrad control

A verbatim clone of the T01 hilbert arm with PCGrad gradient surgery (`grad_surgery: pcgrad`; equation 6): for every conflicting pair with negative cosine, each task's gradient is projected off the other's normal plane before accumulation. On the 54-dataset cell the arm scores 0.3939 $\pm$ .0018 (per-seed 0.3959, 0.3922, 0.3937) against the unmodified anchor's $0.3916 \pm .0230$ — a $+0.2$-point shift buried inside the anchor's own seed spread. Together with the classic-era null of §4.7 (0.6571 against 0.6587 unmodified), direction-conflict resolution repays nothing in either era, while the norm-matched control in the same accumulation slot repays 23.7 of 24.5 points (Appendix G.3): the debt's carrier is gradient-norm domination, not direction conflict.

H Formal statements

The mechanisms of §3–4 as numbered statements, each anchored to the reported measurements; constants are the deployed configuration (Appendix B). Table 13 maps each statement to its verification.

Table 13: Statement–verification map; every entry is a reported measurement.

Statement	Verified at	Measurement
S1 spherical-iterative dynamics	§3.3, App. B	deployed configuration
S2 norm-dominated rank-one capture	§4.5	$+0.8318$ versus $\leq +0.08$
S3 dose threshold, not gradient	§4.4	0.6341 / 0.6299 / 0.3697
S4 direction surgery never fires	§4.7, App. G.6	$+0.2$, inside seed spread
S5 repayment ledger	§4.7	4.8 / 23.7 / 9.1 of 24.5
S6 collapsed-representation geometry	§4.5	effective rank 3.9–4.5 vs 8.6–9.6
S7 iterative dynamics necessary	App. G.5	$+1.2$, sign-reversed amplitude

Statement S1 (spherical-iterative dynamics). T01 holds a 12-slot state $z \in \mathbb{R}^{12\times 248}$ constrained to a product of 4-dimensional sub-spheres $(\mathbb{S}^3)^{62}$. Reading sets $z^{(0)} = \mathrm{L2norm}\big(z_{\text{init}} + \mathrm{XAttn}_8(\mathrm{LN}(z_{\text{init}}), X, X)\big)$ with a learned z_{init}. Each of $K{=}3$ steps recomputes the self-referential context $c = \mathrm{mean}_{\text{slots}}(z)$, descends an explicit energy by one tangent-projected gradient step, blends toward the read state, and re-projects:

$$g = \nabla_z E(z,c), \qquad g_t = g - (g\cdot\hat{z})\,\hat{z}, \qquad z \leftarrow z - 0.05\,g_t, \qquad z \leftarrow 0.8\,z + 0.2\,z^{(0)}, \qquad z \leftarrow \Pi_{\text{poly}}(z),$$

$$E(z,c) = 0.5 \operatorname*{mean}_{i\neq j} \|\hat{z}_i - \hat{z}_j\|^{-1} \;-\; 0.5 \operatorname*{mean}_i \langle \hat{z}_i, \mu(c)\rangle \;+\; \operatorname*{mean}_i (\|z_i\| - 1)^2,$$

where $\hat{z} = z/\|z\|$, $\mu(c) = c/\|c\|$, and Π_{poly} re-normalizes each 4-dimensional segment. The unrolled loop is fully differentiable; the readout scores the pooled L2-normalized feature against per-task frozen ETF prototypes under a learned temperature (§3.3, Figure 1).

Statement S2 (norm-dominated rank-one capture). If one task's gradient norm dominates, $\|g_{d^*}\| \gg \mathrm{median}_e \|g_e\|$, and pairwise cosines are positive, the accumulated update direction is approximately $g_{d^*}/\|g_{d^*}\|$ irrespective of the remaining tasks. Anchor: under hilbert/T01 the mean pairwise cosine is $+0.8318$ (median $+0.9072$) while every other configuration measures $\leq +0.08$ (Table 5); modelnet10's gradient norm is 10136 against 4–11 for all others, a max-to-median ratio of $1.4\text{–}3.1 \times 10^3$ sustained across training (Appendix G.4).

Statement S3 (dose threshold, not gradient). $\mathrm{Acc}(k) \approx A_{\text{healthy}} - \mathcal{D}\cdot \mathbb{1}[k \geq k^*]$ with $k^* \in (2, 3]$: doses one and two sit within a point of the healthy line (0.6341, 0.6299), dose three collapses (0.3697), and dose four sits above the trough ($0.3916 > 0.3697$) — threshold crossing, not monotone descent. The dose-three seed split (0.4255 against 0.3415 and 0.3420) is the bistable signature expected at a critical point (Table 4; Appendix F).

Statement S4 (direction surgery never fires). In the captured regime the conflicting-pair set $\{(d,e) : \cos(g_d, g_e) < 0\}$ is nearly empty — the minimum pairwise cosine over all 1431 pairs is

-0.0361 (Table 5) — so the projection condition of equation 6 is met with frequency ≈ 0 and the operator degenerates to the identity. Anchor: PCGrad repays $+0.2$ points, inside the anchor's own seed spread (Appendix G.6).

Statement S5 (repayment ledger). With $R(\text{fix}) = \big(\text{Acc}_{\text{fix}} - \text{Acc}_{\text{hilbert}}\big)/\mathcal{D}$: a trainable ETF head repays 4.8 of 24.5, per-task norm equalization equation 5 repays 23.7 of 24.5 with dead datasets driven to zero, and data-side order augmentation repays 9.1 of 24.5; direction surgery repays 0.2. The ledger localizes the carrier to gradient-norm domination — not the readout, not direction conflict (Appendices G.1, G.3).

Statement S6 (geometry of the collapsed representation). Effective rank is the participation ratio $(\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$ of the pooled latent's covariance spectrum; centroid coupling is the mean cosine between per-dataset latent centroids across modality lines. Collapse halves the first ($3.9 \rightarrow 4.5$ under refinement, against $8.6 \rightarrow 9.6$ on views) and more than doubles the second (0.804 against 0.328 and 0.365), while task identity persists (0.77–0.89): the collapsed representation is entangled along modality lines and compressed into a low-rank coupled subspace, not scrambled (Table 6).

Statement S7 (shared iterative dynamics are necessary). A parameter-matched vanilla transformer trunk under the identical hilbert protocol does not collapse ($0.6396 \pm .0058$; amplitude $+1.2$ points, sign-reversed against T01's -24.5): representation debt requires the spherical-iterative loop of Statement S1 as its carrier, not dense sharing as such (Appendix G.5).