ARCHITECTURAL PRIORS FOR ORDINAL REGRESSION VIA LOCAL HEADS AND ORDERED FEATURES

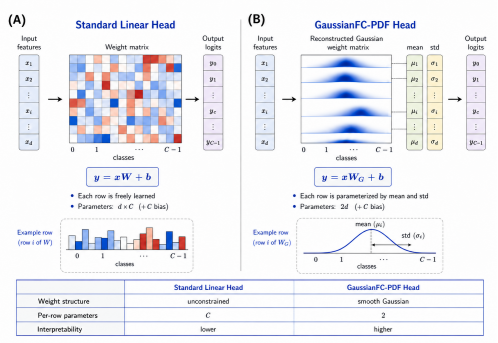
Anonymous authors
Paper under double-blind review

ABSTRACT

Ordinal prediction problems—age estimation, disease grading, historical image dating—involve labels with intrinsic order, yet deep ordinal methods act at the loss or the output-side posterior, leaving the head's mapping from features to logits agnostic to label order. We argue that the head itself is the third and right locus of inductive bias, and develop a theoretical framework around it. We introduce a family of *label-axis structured heads* in which each feature dimension affects only a local neighborhood of ordered labels, and prove that such heads operate through two complementary mechanisms: an ordered classifier mapping that localizes predictions on the label axis, and an implicit bias that drives the backbone toward ordered feature geometry through label-frequency filtering of the gradient field. Across 23 dataset–backbone cells spanning age estimation, histology and diabetic-retinopathy grading, and historical image dating (K=4 to 70), and eight ordinal baselines, our heads win or tie 177–182 of 184 Holm-corrected pairwise comparisons on MAE (128–145 of 161 on the ranked probability score) and are the only methods within 3% median gap of the per-cell best in every label-axis regime, while every baseline fails in some regime by 22–233% at worst. Mechanism-level analysis confirms the two mechanisms are dissociable: the Poisson-Unimodal baseline realizes only the second—it stays within 3–10% at large K—the best of any baseline in that regime—but collapses at small K and transfers a weaker representation under head swapping.

1 INTRODUCTION

Ordinal prediction—age estimation (Fu et al., 2015), disease grading (Aresta et al., 2019), historical image dating (Palermo et al., 2012)—involves labels with intrinsic order, where confusing adjacent classes is less severe than confusing distant ones. Specialized losses and architectures for this setting improve substantially over the cross-entropy baseline inherited from unordered multi-class classification, but performance saturates on small and imbalanced benchmarks such as FG-NET, and—more fundamentally—existing methods are isolated heuristics enforcing unimodal output distributions, with no principled account of *why* this prior helps or *which* structural choices should be preferred (Cardoso et al., 2025; Wang et al., 2025).



	Standard Linear Head	GaussianFC-PDF Head
Weight structure	unconstrained	smooth Gaussian
Per-row parameters	C	2
Interpretability	lower	higher

Figure 1: Standard linear head vs. the proposed GaussianFC-PDF head. Each row of GaussianFC-PDF is parameterized by a mean and a standard deviation, yielding a smooth Gaussian-shaped weight profile over class positions and reducing per-row degrees of freedom from C to 2.

This paper develops a theoretical framework around a single thesis: ordinal learning is governed not by the loss function alone, but by the *implicit bias that the classifier head imposes on representation learning* (Soudry et al., 2018). Replacing the standard linear head with one whose feature-to-logit mapping is structured along the ordered label axis yields large gains even when the backbone and loss are held fixed, and stronger structural priors—locality, smoothness—amplify them. We attribute this to a tighter alignment between the head's optimization bias and the geometry that ordinal tasks

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.

demand: localized basis functions over the label axis induce localized gradients on the backbone, which in turn drive the learned representation into an *ordered feature geometry*.

We make four **contributions**. **(i)** An interpretable feature-to-label parameterization. We introduce a Gaussian ordinal head that parameterizes each feature's class-weight profile by an explicit label center and bandwidth. This embeds ordinal structure directly into the feature-to-label weight mapping, rather than encoding it through the loss or an output-side posterior parameterization—two common strategies in ordinal learning. The head aggregates feature-wise contributions across label regions, making the mapping from backbone activations to ordinal logits directly inspectable. On ResNet-18 with FG-NET, this design reduces classifier parameters by over 95% relative to a linear head.

(ii) We establish two complementary mechanisms by which such heads improve ordinal prediction: an *ordered classifier mapping* that localizes predictions and gradients on the label axis (Prop. 1), and an *implicit bias on representation learning* that drives the backbone toward ordered feature geometry via label-frequency filtering of the gradient field (Props. 2–3).

(iii) Empirical evidence of label-related representation structure and transferability. On FG-NET with ViT-Tiny, Gaussian heads exhibit class-center separation that increases with label gap on average, lower normalized class-center roughness than ordinal baselines, and improved ordinal readout when multiple leading principal components are used instead of the first alone. These diagnostics characterize complementary aspects of label-related geometry. Frozen-backbone and head-swap experiments on CNN and ViT backbones further indicate that predictive gains depend on backbone adaptation and remain accessible to a retrained linear readout.

(iv) Across 23 dataset–backbone–resolution cells on FG-NET, BACH, HCI and APTOS (K=4–70) with ResNet and ViT backbones, Gaussian heads combining the two proposed mechanisms achieve strong empirical performance against eight ordinal baselines. Under the reported Holm-corrected tests, they outperform or show no statistically significant difference from the compared baselines in 177–182 of 184 MAE comparisons and 128–145 of 161 ranked-probability-score comparisons. Both heads achieve median relative MAE gaps of at most 3% from the per-cell best in both the small-K and large-K regimes. The variants exhibit different empirical strengths: GFC-PDF achieves lower median relative RPS gaps at large K, whereas GFC is less sensitive to hyperparameter choice in the tested settings.

2 PRELIMINARIES

Setup. An ordinal task has ordered label space $\mathcal{Y} = \{1, \dots, K\}$ with $1 < \cdots < K$. A backbone f_ψ produces features $z \in \mathbb{R}^d$, and a head maps z to logits $o(z) \in \mathbb{R}^K$. Standard practice uses a linear head, which encodes no order.

Label-axis structured heads. Associating each label j with a position t_j on an ordered axis ($t_1 < \cdots < t_K$), we study heads of the form

$$o_j(z) = \sum_{i=1}^{d} z_i \, \Phi_i(t_j; \theta_i) + b_j, \tag{1}$$

where $\Phi_i(\cdot; \theta_i)$ is a feature-specific basis function on the label axis. We collect the head's evaluations into the *effective weight matrix* $W \in \mathbb{R}^{d \times K}$ with $W_{ij} = \Phi_i(t_j; \theta_i)$, so that $o(z) = z^\top W + b$. When Φ_i is localized, feature i influences only nearby logits; when additionally smoothed, adjacent logits vary continuously. Concrete instances appear in Section 4.

Ordered feature geometry. Let $c_k = \mathbb{E}[z \mid y = k]$ be the class mean. We say the representation exhibits *ordered feature geometry* if (i) $\|c_k - c_{k'}\|$ increases with $|k - k'|$ on average, (ii) the class-mean trajectory $\{c_k\}_{k=1}^K$ aligns with a low-dimensional ordered manifold, and (iii) the feature-curvature functional $\mathcal{K}_{\text{feat}}(C) := \sum_{k=2}^{K-1} \|c_{k+1} - 2c_k + c_{k-1}\|^2$ is small, where $C = [c_1, \dots, c_K]$. This is stronger than separability—it is *ordered separability*.

Definition 1 (Gradient locality bias) *For row i of W, let $\mu_i \in \arg\max_j |W_{ij}|$ and $A_i = |W_{i\mu_i}|$, and define the* envelope width

$$\sigma_i(W) \;:=\; \min\Big\{\sigma > 0 \;:\; |W_{ij}| \le A_i\, e^{-(j-\mu_i)^2/2\sigma^2},\; j = 1, \dots, K\Big\}$$
$$= \max_{j\,:\,0<|W_{ij}|<A_i} \frac{|j-\mu_i|}{\sqrt{2\ln\big(A_i/|W_{ij}|\big)}}, \tag{2}$$

with $\sigma_i(W) := \infty$ if the row attains its maximum at more than one position. The head induces gradient locality bias *of width $\sigma_i(W)$, and the bias is stronger the smaller σ_i/K is: an envelope with $\sigma_i \le K/4$ keeps over 95% of its mass within half the label axis, against 68% at $\sigma_i = K/2$.*

Definition 1 is architectural: $\sigma_i(W)$ is a function of the trained weight matrix alone. Since $\partial\ell/\partial z_i = \sum_j (\partial\ell/\partial o_j)\, W_{ij}$, the envelope directly bounds how labels at distance $> r$ from the target can contribute to backbone updates. Section 5 (Prop. 1) shows our heads admit an envelope with A_i, σ_i tied to the head's learned amplitude and width; values are tabulated in Appendix A.4. A standard linear head has no architectural mechanism holding σ_i down, and measurably does not: on FG-NET/ResNet-18 the median σ_i/K is 0.27–0.36 for the GaussianFC family versus 1.12 for Linear+CE (App. C.6).

Note on sufficiency. Ordered feature geometry is a *necessary* but not *sufficient* property for strong ordinal prediction: a model that induces ordered features but lacks a classifier mapping that exploits this order locally on the label axis will still produce non-local errors. Section 5 establishes that label-axis structured heads address both axes simultaneously—inducing ordered geometry through their implicit-bias on optimization (Prop. 3) while implementing an ordered classifier mapping through their architecture (Prop. 1).

3 Related work

Ordinal regression. Surveys (Cardoso et al., 2025) organize ordinal methods around losses (UR (Liu et al., 2020), CO2 (Albuquerque et al., 2021)) and models (Binomial Unimodal (da Costa & Cardoso, 2005), Poisson Unimodal (Beckham & Pal, 2017), VS-SL (Yamasaki, 2022)). A second line builds order into the output layer: CORAL (Cao et al., 2020) imposes rank-consistent thresholds on a shared projection, CORN (Shi et al., 2023) restores rank consistency through conditional probabilities without CORAL's shared-weight constraint, and DLDL (Gao et al., 2017) replaces the one-hot target with a Gaussian-smoothed label distribution. All act at the loss or the output-side posterior, leaving the head's mapping from features to logits agnostic to label order. We instead embed ordinal structure into the *head architecture itself*, parameterizing the per-feature weight row as a localized basis function on the label axis: the third locus, and the one at which a *local* prior along the label axis can be imposed at all. Our heads remain compatible with the cross-entropy and unimodal-output losses above. Rank-N-Contrast (RNC) (Zha et al., 2023) is orthogonal to all three: its prior acts on the *representation* (embedding distances ordered by label distance) and leaves the readout free.

Relation to Radial basis function networks. While our use of Gaussian-shaped functions resembles radial basis function networks, the similarity is purely formal. Classical RBF networks (Broomhead & Lowe, 1988) apply basis functions to inputs to construct nonlinear feature representations, whereas we apply basis functions over the ordered label axis in the final classifier head, with the explicit goal of inducing an ordered classifier mapping and ordered feature geometry.

Feature geometry and Neural Collapse. Neural Collapse (Kothapalli, 2022) characterizes the symmetric simplex geometry of standard multi-class training; Ordinal Neural Collapse (ONC; (Ma et al., 2025)) shows that under cumulative-link ordinal regression, class means asymptotically collapse onto a one-dimensional ordered subspace. ONC analyzes the asymptotic endpoint of a specific (threshold-based) model class; we study the *optimization-side* mechanism by which an arbitrary backbone, paired with a structured head, is driven toward such geometry—the class-center update lies in the head's low-frequency label modes up to an explicitly bounded leakage term, and of those modes only the Fiedler mode is monotone in the label index (Prop. 3; a detailed comparison with ONC's asymptotic characterization is given in Appendix B).

Implicit bias and architectural priors. Gradient descent has well-documented implicit biases on linear and homogeneous classifiers (Soudry et al., 2018; Wang et al., 2024), and recent work points

to feature–classifier mismatch as a source of suboptimal training (Wu et al., 2024). Our analysis is in this spirit but architectural: changing the head from a linear map to a label-axis structured one alters the gradient distribution along the label axis, and thereby the geometry the backbone learns. The *locality* side of our prior also relates to prototype-based classifiers (Zhang et al., 2025), which localize each class in feature space; we instead localize each *feature* on the label axis.

4 METHOD

Comparison of classification heads. We propose two structured heads—GaussianFC (GFC) and GaussianFC-PDF (GFC-PDF)—both instances of Eq. (1) (Figure 1): where a linear classifier learns each row of W freely, a Gaussian row is set by a mean and a width. Imposing only this constraint, with everything else unchanged, wins or ties the overwhelming majority of Holm-corrected comparisons against eight ordinal baselines (Section 6, Table 1).

Structured head family. For backbone features $z \in \mathbb{R}^d$, our heads have the form $o_j(z) = \sum_i z_i \, \Phi_i(t_j; \theta_i) + b_j$ with $\Phi_i(\cdot; \theta_i)$ a feature-specific basis function on the label axis (Eq. 1). The simplest instance uses a Gaussian basis $\Phi_i(t_j; \theta_i) = a_i \exp(-(t_j - \mu_i)^2 / 2\sigma_i^2)$ with (a_i, μ_i, σ_i) controlling amplitude, center and spread. Both variants below are realized as a particular choice of Φ_i.

Heads as graph operators. Treat the ordered labels $\{1, \ldots, K\}$ as a path graph with Laplacian L. Each head becomes an operator on this graph and is summarized by two scalars: *locality* $|\mathcal{S}_i(\tau)| = |\{j : |W_{ij}| \geq \tau\}|$ (small=local) and *Dirichlet energy* $\mathcal{E}_{\text{head}}(W) = \text{Tr}(WLW^\top) = \sum_{i,j}(W_{i,j+1} - W_{ij})^2$ (small=smooth). Each variant occupies a different point of the (locality, smoothness) plane, and Section 5 shows this position determines both the gradient signal sent to the backbone and the learned feature geometry.

Two concrete heads. **(a) GaussianFC** is a Gaussian-shaped row normalized across labels and rescaled by a learnable amplitude a_i, $W_{ij} = a_i \exp(-(t_j - \mu_i)^2 / 2\sigma_i^2) / \sum_k \exp(-(t_k - \mu_i)^2 / 2\sigma_i^2)$. The Gaussian basis itself is graph-smooth (its Dirichlet energy decays as $O(\sigma_i^{-3})$); locality and smoothness are jointly controlled by a single knob σ_i. **(b) GaussianFC-PDF** replaces the amplitude with the Gaussian PDF normalizer, $W_{ij} = \exp(-(t_j - \mu_i)^2 / 2\sigma_i^2) / (\sigma_i \sqrt{2\pi})$—same coupled locality/smoothness as (a), one fewer degree of freedom, the strongest implicit regularizer in the family.

When to use which. The two variants divide by metric: GFC is ahead under MAE (90/2 Holm-corrected wins/losses against 86/7), while GFC-PDF separates clearly under RPS (80/16 against 54/33; median gap to the per-cell best 2–3% against 7–9%). Point accuracy and predictive-distribution quality are therefore not optimized by the same variant. Architecturally, GFC's learnable amplitudes matter when feature scales are heterogeneous—deeper backbones with high-dimensional embeddings—whereas GFC-PDF's parameter parsimony is the stronger implicit regularizer on small datasets.

5 THEORETICAL FRAMEWORK

We organize our analysis around two coupled mechanisms. **Mechanism A: label-structured feature aggregation.** Each feature contributes through a Gaussian-shaped weight profile over ordered labels, with an explicit center and bandwidth. Proposition 1 quantifies the resulting envelope on far-label contributions to its feature gradient. **Mechanism B: head-induced structure in representation updates.** Propositions 2–3 connect head roughness and spectral energy to a conditional bound on one-step class-center updates. Both mechanisms arise from the same parameterization; complementary experiments examine their gradient signatures and the geometry and transferability of the learned representations.

Labels lie on the integer grid $t_j = j$. Let $L = U \Lambda U^\top$ be the path-graph Laplacian, with orthonormal eigenvectors u_n and eigenvalues $0 = \lambda_1 < \lambda_2 \leq \cdots \leq \lambda_K < 4$. For the effective head $W \in \mathbb{R}^{d \times K}$, the cross-entropy residual is $q = p - e_y$, where $p = \text{softmax}(o(z))$, and the feature gradient is

$\partial \ell / \partial z = Wq$. This residual expression assumes one-hot softmax cross-entropy; soft targets and additional loss terms require the corresponding derivatives. We write $C = [c_1, \ldots, c_K]$ for the class-center matrix. Head roughness admits the spectral expansion

$$\mathcal{E}_{\text{head}}(W) = \operatorname{Tr}(WLW^\top) = \sum_{n=1}^{K} \lambda_n \|Wu_n\|_2^2.$$

Proposition 1 (Tail decay of gradient contributions) *Suppose row i admits a Gaussian envelope $|W_{ij}| \le A_i \exp(-(j-\mu_i)^2/(2\sigma_i^2))$, with $A_i \ge 0$, $\mu_i \in \mathbb{R}$, and $\sigma_i > 0$. For any label y and radius $r \ge 0$, define $P_{\text{off}}(r; y) = \sum_{|j-y|>r} p_j$. Then*

$$T_i(r; y) := \left| \sum_{|j-y|>r} W_{ij} q_j \right| \le P_{\text{off}}(r; y) A_i \exp\left(-\frac{(r - |\mu_i - y|)_+^2}{2\sigma_i^2} \right). \tag{3}$$

The bound separates the weight envelope from the off-target residual mass. Its attenuation depends on the width and on the offset between the feature's weight center and the target label. The parameterization specifies the functional form, while its learned parameters and the prediction residual determine the bound at a given training step.

Proposition 2 (Head roughness and spectral energy) *For exact sampled-Gaussian rows $W_{ij} = \alpha_i \exp(-(j-\mu_i)^2/(2\sigma_i^2))$, where $\alpha_i \in \mathbb{R}$, $\mu_i \in \mathbb{R}$, and $\sigma_i > 0$, set $A_i = |\alpha_i|$. Then*

$$\mathcal{E}_{\text{head}}(W) \le \frac{\sqrt{\pi}}{2} \sum_i \frac{A_i^2}{\sigma_i}, \operatorname{Tr}(WL^2W^\top) \le \lambda_K \mathcal{E}_{\text{head}}(W). \tag{4}$$

For a spectral filter $H = h(L)$ with $0 \le h(\lambda_n) \le 1$, $\mathcal{E}_{\text{head}}(W_0 H) \le \mathcal{E}_{\text{head}}(W_0)$; when h is also non-increasing, higher-frequency modes receive no larger multiplicative weights.

For GFC-PDF, $A_i = (\sigma_i \sqrt{2\pi})^{-1}$, giving the bound $(4\sqrt{\pi})^{-1} \sum_i \sigma_i^{-3}$. For GFC, $A_i = |a_i|/Z_i$, where Z_i is the discrete Gaussian normalizer; the corresponding $O(\sum_i a_i^2 \sigma_i^{-3})$ bound requires the width and center-containment conditions in Appendix A.2. These bounds apply to the effective head at its current parameter values. The direct sampled-Gaussian proof does not require identifying the head with an exact finite-graph heat kernel.

Proposition 3 (Low-frequency bias of class-center dynamics) *Consider one population gradient-descent step on the backbone with rate η, using the current effective head W. Assume the local linearization and feature-kernel conditions of Appendix A.3: diagonal class-kernel blocks are close to their mean $\bar{\Theta}_t$, cross-class blocks are weak, and the kernel–residual factorization error is bounded. With $\bar{Q} = [\mathbb{E}[q \mid y = 1], \ldots, \mathbb{E}[q \mid y = K]]$ and class-prior matrix $\Pi = \operatorname{diag}(\pi_1, \ldots, \pi_K)$,*

$$\Delta C_t := C_{t+1} - C_t = -\eta \bar{\Theta}_t W \bar{Q} \Pi + \mathcal{R}_t,$$

where $\mathcal{R}_t$ collects the stated approximation errors and the Taylor remainder. Let $\Pi_\star$ be the orthogonal projector in feature space onto $\operatorname{span}\{\bar{\Theta}_t W u_n : \lambda_n \le \lambda_\star\}$. Then

$$\|(I - \Pi_\star)\Delta C_t\|_F \le \eta \|\bar{\Theta}_t\|_{\text{op}} \|\bar{Q}\Pi\|_F \, \tau(\lambda_\star) + \|\mathcal{R}_t\|_F, \tau(\lambda_\star)^2 := \sum_{\lambda_n > \lambda_\star} \|Wu_n\|_2^2. \tag{5}$$

The projected subspace has dimension at most $|\{n : \lambda_n \le \lambda_\star\}|$.

Connecting head roughness to the update bound. For $\lambda_\star > 0$, the spectral expansion gives

$$\tau(\lambda_\star)^2 \le \frac{\mathcal{E}_{\text{head}}(W)}{\lambda_\star}.$$

Combining Propositions 2 and 3 therefore yields the direct corollary

$$\|(I - \Pi_\star)\Delta C_t\|_F \le \eta \|\bar{\Theta}_t\|_{\text{op}} \|\bar{Q}\Pi\|_F \sqrt{\frac{\mathcal{E}_{\text{head}}(W)}{\lambda_\star}} + \|\mathcal{R}_t\|_F. \tag{6}$$

Thus, the explicit Gaussian head-energy bound enters a bound on the component of the class-center update outside the specified feature subspace. Its informativeness depends on the head energy, the residual and kernel factors, and the approximation remainder.

Head	metric	K	CE	PU	UR	CORAL	CORN	DLDL	RNC+CE	RNC+L1	all	own gap med/max
GFC	MAE	K = 4–5	7/7/0	14/0/0	2/12/0	13/1/0	1/13/0	4/10/0	2/12/0	8/6/0	51/61/0	2% / 16%
GFC	MAE	K = 62–70	9/0/0	3/6/0	5/4/0	7/1/1	7/2/0	3/6/0	2/6/1	3/6/0	39/31/2	0% / 11%
GFC	RPS	K = 4–5	4/10/0	4/6/4	9/3/2	12/2/0	2/11/1	3/9/2	2/10/2	—	36/51/11	9% / 24%
GFC	RPS	K = 62–70	6/0/3	0/5/4	0/5/4	6/0/3	4/2/3	1/5/3	1/6/2	—	18/23/22	7% / 19%
GFC_PDF	MAE	K = 4–5	7/7/0	14/0/0	3/10/1	13/1/0	2/12/0	4/10/0	2/12/0	7/7/0	52/59/1	1% / 17%
GFC_PDF	MAE	K = 62–70	9/0/0	2/5/2	6/3/0	6/1/2	6/3/0	2/6/1	1/8/0	2/6/1	34/32/6	3% / 6%
GFC_PDF	RPS	K = 4–5	7/7/0	6/6/2	10/2/2	12/2/0	5/8/1	4/9/1	5/7/2	—	49/41/8	2% / 23%
GFC_PDF	RPS	K = 62–70	9/0/0	0/5/4	1/6/2	8/0/1	7/2/0	3/6/0	3/5/1	—	31/24/8	3% / 15%
gap of the opponent	MAE	K = 4–5	11% / 26%	32% / 98%	3% / 22%	36% / 132%	6% / 24%	1% / 233%	6% / 35%	14% / 53%	—	—
gap of the opponent	MAE	K = 62–70	49% / 78%	3% / 10%	10% / 15%	30% / 128%	16% / 38%	6% / 11%	3% / 26%	3% / 22%	—	—
gap of the opponent	RPS	K = 4–5	14% / 28%	9% / 43%	36% / 92%	41% / 98%	10% / 28%	4% / 186%	8% / 36%	—	—	—
gap of the opponent	RPS	K = 62–70	39% / 56%	0% / 6%	2% / 3%	46% / 152%	18% / 38%	9% / 16%	4% / 19%	—	—	—

Table 1: Win/tie/loss of each Gaussian head against eight fixed baselines, per K regime, under MAE (point prediction) and RPS (ordinal proper scoring rule). Wilcoxon signed-rank, Holm–Bonferroni corrected within each metric over the full set of tests in the table (368 for MAE, and 322 for RPS). Bottom rows give each baseline's own median/worst relative gap to the per-cell best.

6 EXPERIMENTS

We test predictive performance (E1, E3), the source of the gain (E2), the two mechanisms of §5—A (E4), B (E5), both in weight space (E6)—and robustness (E7). **Setup.** Five datasets span two label-axis regimes. Large K (face age): **FG-NET** (Fu et al., 2015) (K=70, N=1,002, imbalance ratio 3.7; our primary diagnostic) and **MORPH** (Ricanek & Tessema, 2023) (K=62, N=50,015). Small K: **BACH** (Aresta et al., 2019) (K=4), **HCI** (Palermo et al., 2012) (K=5), **APTOS** (K=5), and **FG-NET-5**, FG-NET coarsened to K=5 on the same splits. Only the final head of ResNet-18/50 or ViT-Tiny (Steiner et al., 2021) (128/224 px) is replaced, giving 23 (dataset, backbone, resolution) cells (Table 1). Each method is tuned separately on a nested validation split; results are means over 5 subject-exclusive folds $\times$ 5 seeds, compared on matched runs by Holm-corrected Wilcoxon signed-rank tests (App. C.1). We report MAE and the ranked probability score (RPS), the ordinal proper scoring rule; NLL, accuracy and QWK are in App. Table C.1. **Baselines.** Our GFC and GFC-PDF place the prior in the head's feature-to-logit mapping; the baselines cover the other two loci (§3): the loss (LINEAR+UR, DLDL) and the output-side posterior (PU, CORAL, CORN). We add prior-free Linear+CE and representation-side RNC.

6.1 PERFORMANCE AND ORDINAL CONSISTENCY (E1–E3)

E1: main results. Table 1 summarizes 184 Holm-corrected pairwise comparisons per head per metric (23 cells $\times$ eight baselines; per-cell values in App. Table C.1). Under MAE the two Gaussian heads win or tie $182/184$ (GFC, 90 wins) and $177/184$ (GFC-PDF, 86 wins); under RPS, $128/161$ and $145/161$. Their median gap to the per-cell best is at most 3% in every MAE–regime pair, and at K=62 $-$ 70 their worst-case gap is 6–11$\%$. No baseline shares this profile—each fails in some regime, by 22–233$\%$ at worst: PU, RNC and DLDL collapse at small K (max gaps 98%, 53% and 233%), Linear+CE and CORN degrade at large K (worst 78% and 38% under MAE), and CORAL trails in both regimes; the per-cell anatomy is in App. C.2.

The strongest large-K baseline is PU (own gap 3–10$\%$); Holm-corrected differences against it split 5–2 in favor of the Gaussian heads—despite PU's strong ordered feature geometry there (E5). This is consistent with Mechanism A contributing on top of the representation-side prior PU shares; the controlled evidence for the dissociation remains E2's transfer asymmetry and E4–E6.

E2: classifier vs. representation. The mechanism diagnostics (E2–E6) are computed on FG-NET under a subject-exclusive evaluation protocol: configurations are selected on validation, and every number pools 25 (fold, seed) runs per method on ViT-Tiny at 128 px with test MAE as the metric (cross-backbone replications in the appendix). We disentangle the source of the structured-head gain through frozen-backbone retraining and asymmetric head-swap transfer (Figure 2).

(i) The structured head induces a representation better suited to ordinal prediction. A Linear+CE probe retrained on a frozen GFC-trained backbone reaches 4.87 MAE, far better than Linear+CE trained end-to-end on its own backbone (7.17; -2.30 MAE). The head is identical in both cases—only the representation differs—so the induced representation generalizes beyond the head that produced it.

(ii) Head and representation are complementary. The same probe still trails the structured-head end-to-end result (4.87 vs. 4.53): the structured head is what reads the ordered representation out

into a locality-aware label-axis mapping. This matches §5—Mechanism B shapes the representation, Mechanism A reads it out.

(iii) The transfer is asymmetric. In the reverse direction, Gaussian probes on frozen linear-trained backbones degrade (5.92–20.91 MAE, worst on the PU backbone), and on a backbone frozen at its generic ImageNet initialization structured heads underperform linear ones (10.64–23.07 vs. 6.93–7.72): the locality prior matches a co-adapted ordered representation but mismatches a generic or Linear-trained one. The asymmetry rules out the trivial reading that probes are simply easier to train. Linear+CE serves as the forward probe to avoid conflating representation quality with the probe's own ordinal prior; the pattern replicates on ResNet-18 and ResNet-50 (App. C.3).

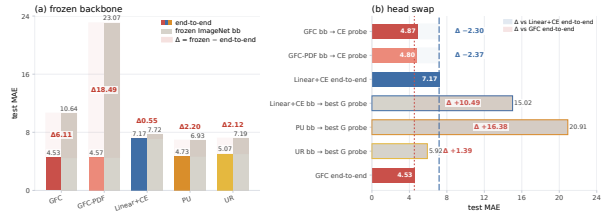
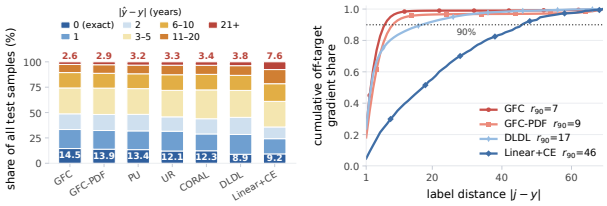


Figure 2: Source of the gain: classifier or representation? (E2, FG-NET, ViT-Tiny.) (a) Frozen backbone: test MAE trained end-to-end vs. with the backbone frozen at its ImageNet initialization (head-only training). The bracket on each pair is the adaptation gap Δ = frozen − end-to-end. (b) Head swap: a Linear+CE probe retrained on frozen structured-trained backbones, and Gaussian probes on frozen linear-trained backbones. Both end-to-end references are drawn as bars and as guide lines.

E3: ordinal consistency. Figure 3(a) decomposes predictions by ordinal error distance. Ordinal-aware heads reduce aggregate large-error mass ($d\geq3$) relative to Linear+CE (51.3–56.1% vs. 64.2%), but this aggregate barely separates the ordinal heads from one another—they span 4.8 pp among themselves while sitting 8 pp or more below Linear+CE—so the informative differences sit at the two ends of the distribution. Exact hits (d=0) range from 8.9% (DLDL) to 14.5% (GFC), and catastrophic errors ($d\geq21$, an age error of more than two decades) from 2.6% (GFC) to 7.6% (Linear+CE), a 3.0× spread. Linear+CE is worst at both ends—second-fewest exact hits (9.2%) and, by a wide margin, the most catastrophic errors—whereas GFC is best at both, with the most exact hits and the fewest catastrophic errors. At the catastrophic end the ordinal heads are tightly bunched (GFC 2.6%, GFC-PDF 2.9%, PU 3.2%, UR 3.3%, CORAL 3.4%, DLDL 3.8%).



(a) Error distance (E3). (b) Gradient locality (E4).

Figure 3: Ordinal consistency (E3) and gradient locality (E4) on FG-NET, ViT-Tiny. (a) E3. Predictions decomposed by ordinal distance $d = |\hat{y}-y|$ as shares of all test samples; in-bar labels give the exact-hit share (d=0), labels above the bars the catastrophic share ($d \geq 21$). (b) E4. Cumulative share of off-target gradient mass as a function of label distance $|j - y|$; curves that saturate earlier indicate stronger gradient locality, summarized by the 90% radius r_{90}. Heads whose output dimension differs from K (PU, UR, CORAL, CORN) expose no per-label gradient field and cannot be measured by this diagnostic; neither can RNC, whose contrastive stage trains no head at all and whose ℓ_1 probe sees a frozen backbone.

6.2 Verifying the two mechanisms (E4–E6)

E4: gradient locality studies (Mechanism A). Prop. 1 predicts that off-target update mass concentrates near the target label. Figure 3(b) plots the cumulative share of off-target gradient mass against label distance: the radius enclosing 90% of the mass collapses from r_{90}=46—most of the 69-step label axis—for Linear+CE to $r_{90}\leq9$ for both GaussianFC-family heads. DLDL, whose ordinal prior lives in soft label targets rather than in the head, sits in between (r_{90}=17) and is a useful control: it matches GaussianFC in the near field (59% vs. 61% of the mass within $d\leq3$) but leaks a longer tail (83% within $d\leq10$ vs. 93–97%), so label-side priors concentrate gradients partially, architecture-side structure concentrates them most. The ordering is consistent across ResNet-18 and ResNet-50 (Linear+CE r_{90}=38–49 vs. 4–7 for the family, App. C.4).

Two clarifications keep this result honest. *Mechanism:* the measured locality is carried by the concentration of the residual field rather than by the Gaussian envelope of W acting as a direct gate—the residual-only view and the W-weighted view of Prop. 1 nearly coincide, and the column norms

of W vary only mildly along the label axis (coefficient of variation 0.11–0.17 on ViT, 0.18–0.21 on the ResNets). The causal chain supported by the data is therefore *head structure → concentrated predictive distributions → concentrated gradients*, i.e., the envelope operates through the outputs it induces. *Scope:* heads whose output parameterization is not a length-K logit vector (PU, UR, CORAL) cannot be measured by this diagnostic, so E4 compares architecture-side and label-side priors only; where the ordinal prior of PU acts is instead visible in E5 and E6.

E5: ordered feature geometry studies (Mechanism B) . We probe class centers $c_k = \mathbb{E}[z \mid y{=}k]$ on FG-NET with ViT-Tiny (Figure 4) along three axes.

(a) Distance vs. gap. Linear+CE is flattest—a 2.0× ratio between $\Delta{=}69$ and $\Delta{=}1$—meaning its class-center separation is nearly independent of label gap, the signature of an unordered scatter. Fluctuations beyond $\Delta{\approx}50$ reflect the few class pairs at extreme gaps, so the reliable comparison is over $\Delta < 50$, where the Gaussian heads lead the other ordinal baselines clearly.

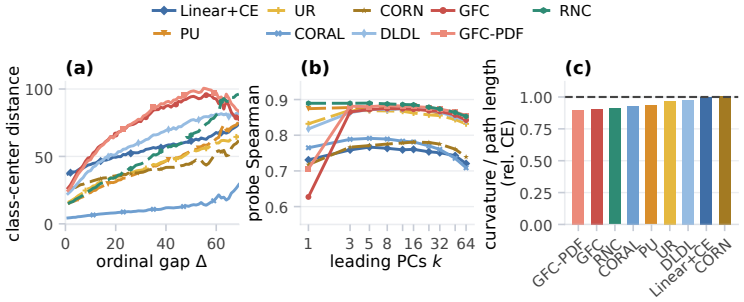


Figure 4: Ordered feature geometry on FG-NET, ViT-Tiny (Mechanism B). Solid: heads trained with cross-entropy. Dashed: dedicated ordinal losses. (a) Class-center distance vs. ordinal gap; fluctuations beyond $\Delta{\approx}50$ reflect the few class pairs at extreme gaps. (b) Out-of-fold Spearman correlation of a ridge probe on the leading k PCs: PU and RNC are highest at $k{=}1$, while GaussianFC-family heads start lower and climb, GFC-PDF overtaking by $k{=}3$—ordinal information distributed across multiple low-frequency modes. (c) Ratio of squared second to squared first differences of the class-center sequence, relative to Linear+CE (the unnormalized $\mathcal{K}_{\text{feat}}$ is confounded by feature norm; §6.2). Per-backbone versions in App. C.5.

(b) Top-k PC probe. With class-stratified 5-fold cross-validation we fit the PCA basis on training folds only, project onto the leading k components, ridge-regress to the label, and report the Spearman correlation of the out-of-fold predictions. At $k{=}1$ RNC and PU are highest (0.890 / 0.874) and the Gaussian heads lowest (0.627 / 0.706), but the family climbs steeply: GFC-PDF overtakes PU at $k{=}3$ (0.881 vs. 0.878) and is level with RNC by $k{=}64$ (0.855 vs. 0.853). This is consistent with Prop. 3: ordinal information spread over several low-frequency modes rather than compressed into one leading direction. The diagnostic identifies *where* ordinal information lives.

(c) Scale-free curvature. We report the ratio of squared second to squared first differences of the class-center sequence, normalized by Linear+CE. On ViT-Tiny the two Gaussian heads are smoothest (GFC-PDF 0.896, GFC 0.903).

Cross-backbone, the geometry is capacity-dependent: the loss- and posterior-side priors drift monotonically toward Linear+CE as the backbone grows, while the two architecture-side heads improve monotonically—architecture-side structure acts on the gradient field directly and does not weaken with capacity, and neither does RNC's representation-side prior ($0.984 \to 0.901 \to 0.909$): what capacity erodes is not every prior outside the head, but every prior that reaches the representation through the output layer. We hypothesize that Gaussian heads offer greater relative representational benefits as backbone capacity increases (App. C.5). However, the current comparisons do not isolate capacity from architecture and optimization, so this hypothesis requires controlled evaluation.

E6: weight-space verification. Direct measurements on the trained effective weight matrix confirm both mechanisms in weight space (Figure 8). For locality we report, per row, the rank correlation ρ between $|W_{ij}|$ and the distance to the row's peak: the GaussianFC family attains $\rho = 1.000$ (every trained row decays monotonically from its peak), versus 0.149 for UR and 0.058 for Linear+CE. We use ρ rather than the fixed-threshold support width of Definition 1 because that statistic saturates at K once the trained envelope is wider than the label axis, hiding the profile shape (at the winning working point it yields a spurious near-tie, 57.5 vs. 61.0; App. C.6). For smoothness, the label-axis Dirichlet energy of Gaussian-family heads is ≈2,300–3,400× lower than Linear+CE—the first direct weight-space verification of the smoothness mechanism against a linear control. Two caveats: ρ

is a *structural descriptor* that localizes where the ordinal prior lives (in W), not a performance predictor—DLDL and UR carry their prior in soft labels or the loss, score low ρ, and remain strong; and heads with output dimension $\neq K$ (PU, UR, CORAL) have no label-axis weight profile, so ρ is undefined for them. Per-method values are in Appendix C.6.

E7: Robustness studies. We report robustness studies (training-set subsampling and neighbourhood label noise, $4{,}025$ runs on FG-NET-R18). Appendix C.7 shows robustness is opponent-specific: under label noise the structured heads degrade significantly more slowly than Linear+CE, UR and the RNC probe, but are indistinguishable from PU, CORAL and DLDL—and CORN degrades more slowly still—while under subsampling no method degrades significantly faster than any other. The RNC probe is the fastest to degrade on both axes.

7 Discussion, limitations, and future work

What did we learn. The two-mechanism framework explains a consistent picture across all experiments: structured heads improve ordinal prediction not by acting as a stronger terminal classifier, but by acting simultaneously as (Mechanism A) an ordered classifier mapping—each feature dimension affects only a local label neighborhood, with a Gaussian envelope on its off-target gradient mass (E4: $r_{90}{=}7$ vs. 49 for Linear+CE)—and (Mechanism B) an implicit optimization bias that drives the backbone toward ordered feature geometry (E5). The two mechanisms are dissociable: PU realizes Mechanism B alone, Linear+CE realizes neither, and only structured heads realize both.

Three loci of ordinal prior. The framework clarifies the design space of ordinal priors by distinguishing three loci at which the prior can be injected: the loss function, the output-side posterior parameterization, and the head's architectural mapping from features to logits. *Loss-based methods* (e.g., UR, CO2) shape the output distribution but leave the head's mapping from features to logits unconstrained, so the gradient field driving backbone learning remains order-agnostic. *Posterior-side methods* (e.g., PU, VS-SL) shape both the output and the representation, but at the cost of collapsing class centers onto a near-one-dimensional manifold (E5(b): probe correlation saturates at $k{=}1$). *Our architecture-side approach* shapes the gradient field and the resulting geometry while preserving *multi-dimensional* ordered structure—ordinal information is distributed across multiple low-frequency modes of the path-graph Laplacian (E5(b): correlation peaks at moderate k). The framework also gives an optimization-side account of Ordinal Neural Collapse: the Fiedler mode is the only non-constant path-graph mode monotone in the label index, so it is the only route by which a monotone component survives the head's filter (App. B). A fourth route is orthogonal rather than alternative: representation-side pretraining (RNC) changes what the encoder is trained to be, and composes with any of the three. E5 uses it as the informative contrast: its ordinal structure sits in the representation and, unlike the loss- and posterior-side priors, does not wash out as backbone capacity grows.

Limitations. *(i)* Prop. 3 holds on a local-linearity window in which the cross-class feature-NTK blocks are weak and the within-class Jacobian and residual are near-uncorrelated (App. A.3); as in NTK, mean-field and unconstrained-features analyses, these constants are assumed small rather than measured, and what we verify empirically are the qualitative predictions (E4–E6). *(ii)* It bounds one gradient step, and from above: it constrains the dimension and curvature of what *can* reach the backbone, not which directions training uses, and it does not characterize the endpoint—that is where ONC-type asymptotic analyses apply (App. B).

Future work. Two extensions follow from the framework. First, the Gaussian envelope of Definition 1 is symmetric, but real-world ordinal label distributions are often asymmetric (e.g., age-estimation annotation skew). Second, our diagnostics are geometric rather than semantic: E4–E6 locate the ordinal information but say nothing about what an individual feature encodes. Because every feature i carries an explicit label center μ_i, the head supports a per-feature claim a linear head cannot even formulate—that a feature centered at the high-age end is driven by high-age evidence such as wrinkle density. Testing it needs attribution or μ_i-banded ablation, which we leave to future work.

AI USE STATEMENT

Large language models were used to polish the writing, including correcting grammar and spelling. Experimental design, implementation, training, evaluation, analysis and all scientific claims are the authors'.

REPRODUCIBILITY STATEMENT

Appendix C.1 specifies the complete experimental protocol: the label definitions (including the fixed clinical order for BACH and the fixed K for APTOS), the subject-exclusive split construction and its verified zero overlap across partitions, the per-method learning-rate and hyperparameter grids, the seed protocol (seeds 0–4 for reporting, a disjoint set 100–102 for selection), the early-stopping and checkpoint rule, and the paired Wilcoxon/Holm–Bonferroni procedure behind every significance claim. Upon acceptance, we will release the full training and evaluation code together with the exact train/validation/test file lists for every fold of every dataset, so that every reported number can be regenerated from fixed artefacts rather than from a re-run of a random generator. All datasets are publicly available, MORPH under its academic licence.

ETHICS STATEMENT

This work studies output heads for ordinal regression and collects no new data. All benchmarks are established public datasets used under their respective terms of access; the medical datasets (BACH, APTOS-2019) are de-identified challenge releases, and no human subjects were recruited. For the facial-age benchmarks, our subject-exclusive protocol ensures that models cannot exploit identity, and we neither train nor evaluate any identification or biometric matching capability. Our medical-imaging results are methodological benchmarks, not validated clinical tools.

REFERENCES

Tomé Albuquerque, Ricardo Cruz, and Jaime S Cardoso. Ordinal losses for classification of cervical cancer risk. *PeerJ Computer Science*, 7:e457, 2021.

Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, et al. Bach: Grand challenge on breast cancer histology images. *Medical image analysis*, 56:122–139, 2019.

Christopher Beckham and Christopher Pal. Unimodal probability distributions for deep ordinal classification. In *International Conference on Machine Learning*, pp. 411–419. PMLR, 2017.

David S Broomhead and David Lowe. Radial basis functions, multi-variable functional interpolation and adaptive networks. Technical report, 1988.

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020.

Jaime S Cardoso, Ricardo PM Cruz, and Tomé Albuquerque. Unimodal distributions for ordinal regression. *IEEE Transactions on Artificial Intelligence*, 2025.

Joaquim Pinto da Costa and Jaime S Cardoso. Classification of ordinal data using neural networks. In *European conference on machine learning*, pp. 690–697. Springer, 2005.

Yanwei Fu, Timothy M Hospedales, Tao Xiang, Jiechao Xiong, Shaogang Gong, Yizhou Wang, and Yuan Yao. Robust subjective visual property prediction from crowdsourced pairwise labels. *IEEE transactions on pattern analysis and machine intelligence*, 38(3):563–577, 2015.

Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 26(6):2825–2838, 2017.

Vignesh Kothapalli. Neural collapse: A review on modelling principles and generalization. *arXiv preprint arXiv:2206.04041*, 2022.

Xiaofeng Liu, Fangfang Fan, Lingsheng Kong, Zhihui Diao, Wanqing Xie, Jun Lu, and Jane You. Unimodal regularized neuron stick-breaking for ordinal classification. *Neurocomputing*, 388:34–44, 2020.

Chuang Ma, Tomoyuki Obuchi, and Toshiyuki Tanaka. Neural collapse in cumulative link models for ordinal regression: An analysis with unconstrained feature model. *arXiv preprint arXiv:2506.05801*, 2025.

Frank Palermo, James Hays, and Alexei A Efros. Dating historical color images. In *European Conference on Computer Vision*, pp. 499–512. Springer, 2012.

Kurt Ricanek and Tesfaye Tessema. Morph-2, 2023. URL `https://www.kaggle.com/ds/3576052`.

Xintong Shi, Wenzhi Cao, and Sebastian Raschka. Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. *Pattern Analysis and Applications*, 26(3):941–955, 2023.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

Jinhong Wang, Jintai Chen, Jian Liu, Dongqi Tang, Danny Z Chen, and Jian Wu. A survey on ordinal regression: Applications, advances and prospects. *arXiv preprint arXiv:2503.00952*, 2025.

Zheng Wang, Geyong Min, and Wenjie Ruan. The implicit bias of gradient descent toward collaboration between layers: A dynamic analysis of multilayer perceptions. *Advances in Neural Information Processing Systems*, 37:74868–74898, 2024.

Xinghao Wu, Jianwei Niu, Xuefeng Liu, Mingjia Shi, Guogang Zhu, and Shaojie Tang. Tackling feature-classifier mismatch in federated learning via prompt-driven feature transformation. *arXiv preprint arXiv:2407.16139*, 2024.

Ryoya Yamasaki. Unimodal likelihood models for ordinal data. *Transactions on Machine Learning Research*, 2022.

Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-n-contrast: Learning continuous representations for regression. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Shihao Zhang, Yuguang Yan, and Angela Yao. Improving deep regression with tightness. *arXiv preprint arXiv:2502.09122*, 2025.

APPENDIX

A PROOFS AND ENVELOPE CONSTANTS

Throughout this appendix, labels occupy the integer grid $t_j = j$, $W \in \mathbb{R}^{d\times K}$ is the effective classifier matrix, and $L = U\Lambda U^\top$ is the unweighted path-graph Laplacian. Its orthonormal eigenvectors u_n are indexed by $0 = \lambda_1 < \lambda_2 \le \cdots \le \lambda_K < 4$. We write

$$\mathcal{E}_{\mathrm{head}}(W) = \mathrm{Tr}(WLW^\top) = \sum_{j=1}^{K-1} \|W_{:,j+1} - W_{:,j}\|_2^2 = \sum_{n=1}^{K} \lambda_n \|Wu_n\|_2^2.$$

Unless otherwise stated, the loss is one-hot softmax cross-entropy, with residual $q = p - e_y$ and feature gradient $\partial\ell/\partial z = Wq$. All head parameters are evaluated at the current training step. The continuous-center Gaussian envelope used in the proofs is distinguished from the discrete-peak diagnostic in Definition 1; see Appendix A.4.

A.1 PROOF OF PROPOSITION 1

Statement. Suppose row i admits an envelope

$$|W_{ij}| \le A_i \exp\left(-\frac{(j-\mu_i)^2}{2\sigma_i^2}\right), \qquad j = 1, \ldots, K, \tag{7}$$

where $A_i \ge 0$, $\mu_i \in \mathbb{R}$, and $\sigma_i > 0$. For a target y and radius $r \ge 0$, let $S = \{j : |j - y| > r\}$, $P_{\mathrm{off}}(r; y) = \sum_{j\in S} p_j$, and $T_i(r; y) = |\sum_{j\in S} W_{ij} q_j|$. Then

$$T_i(r; y) \le P_{\mathrm{off}}(r; y) A_i \exp\left(-\frac{(r - |\mu_i - y|)_+^2}{2\sigma_i^2}\right). \tag{8}$$

Proof. If S is empty, both sides are zero. Otherwise, $r \ge 0$ implies $y \notin S$, so $q_j = p_j \ge 0$ for every $j \in S$ and $P_{\mathrm{off}}(r; y) \le 1 - p_y \le 1$. Consequently,

$$T_i(r; y) \le \sum_{j\in S} |W_{ij}| p_j \le P_{\mathrm{off}}(r; y) \max_{j\in S} |W_{ij}|.$$

For $j \in S$, the reverse triangle inequality gives $|j - \mu_i| \ge (r - |\mu_i - y|)_+$. Since the Gaussian envelope decreases with distance from its center, substitution proves Eq. (8). □

Interpretation and tightness. The bound separates a weight-envelope factor from off-target residual mass. The envelope is centered on the feature's weight profile, whereas the far-label set is centered on the target; their offset therefore affects attenuation. The intermediate maximum-coefficient bound can be approached by concentrating off-target mass at a maximizing index. The analytic envelope bound may retain additional slack from the envelope and the grid. Its functional form is specified by the parameterization, while its parameter values may change during training. It does not imply hard support or unimodality of the final prediction.

Other residuals. For a general logit derivative $q_j = \partial\ell/\partial o_j$, the same argument with absolute values gives

$$\left|\sum_{j\in S} W_{ij} q_j\right| \le A_i \exp\left(-\frac{(r - |\mu_i - y|)_+^2}{2\sigma_i^2}\right) \sum_{j\in S} |q_j|. \tag{9}$$

For soft-target cross-entropy, $q = p - s$ rather than $p - e_y$; the residual-mass factor is therefore not generally P_{off}. Additional loss terms require their corresponding derivatives when interpreting gradient diagnostics.

A.2 PROOF OF PROPOSITION 2

Spectral contraction. Let $H = h(L) = Uh(\Lambda)U^\top$, with $0 \le h(\lambda_n) \le 1$, and let $\widetilde{W} = W_0 H$. Then

$$\mathcal{E}_{\mathrm{head}}(\widetilde{W}) = \sum_{n=1}^{K} \lambda_n h(\lambda_n)^2 \|W_0 u_n\|_2^2 \le \mathcal{E}_{\mathrm{head}}(W_0). \tag{10}$$

The equality follows by diagonalizing $HLH^\top$; the inequality holds term by term. A term is strictly reduced when $\lambda_n > 0$, $W_0 u_n \ne 0$, and $h(\lambda_n) < 1$. If h is additionally non-increasing, higher-frequency modes receive no larger multiplicative weights. Non-increasing behavior without the amplitude bound does not imply contraction.

Exact Gaussian rows. Suppose $W_{ij} = \alpha_i \exp(-(j-\mu_i)^2/(2\sigma_i^2))$, where $\alpha_i \in \mathbb{R}$, $\mu_i \in \mathbb{R}$, and $\sigma_i > 0$. Set $A_i = |\alpha_i|$ and $g_i(t) = \alpha_i \exp(-(t-\mu_i)^2/(2\sigma_i^2))$. Cauchy–Schwarz on each unit interval yields

$$(W_{i,j+1} - W_{ij})^2 = \left(\int_j^{j+1} g_i'(t)\, dt \right)^2 \le \int_j^{j+1} |g_i'(t)|^2\, dt.$$

Summing over j and extending the integral to $\mathbb{R}$ gives

$$\begin{aligned} \sum_{j=1}^{K-1} (W_{i,j+1} - W_{ij})^2 &\le \frac{A_i^2}{\sigma_i^4} \int_{\mathbb{R}} s^2 e^{-s^2/\sigma_i^2}\, ds \\ &= \frac{\sqrt{\pi}}{2} \frac{A_i^2}{\sigma_i}. \end{aligned}$$

Summing over rows proves

$$\mathcal{E}_{\text{head}}(W) \le \frac{\sqrt{\pi}}{2} \sum_i \frac{A_i^2}{\sigma_i}. \tag{11}$$

Furthermore,

$$\begin{aligned} \|WL\|_F^2 = \operatorname{Tr}(WL^2W^\top) &= \sum_n \lambda_n^2 \|W u_n\|_2^2 \\ &\le \lambda_K \mathcal{E}_{\text{head}}(W). \end{aligned} \tag{12}$$

The first bound requires an exact Gaussian row, possibly with a negative row amplitude. A pointwise Gaussian envelope alone does not exclude oscillating signs and is insufficient for that bound. □

Constants for the two heads. For GaussianFC-PDF, $A_i = (\sigma_i \sqrt{2\pi})^{-1}$, hence

$$\mathcal{E}_{\text{head}}(W) \le \frac{1}{4\sqrt{\pi}} \sum_i \sigma_i^{-3}. \tag{13}$$

For GaussianFC, define $Z_i = \sum_{j=1}^K \exp(-(j-\mu_i)^2/(2\sigma_i^2))$. Since $A_i = |a_i|/Z_i$,

$$\mathcal{E}_{\text{head}}(W) \le \frac{\sqrt{\pi}}{2} \sum_i \frac{a_i^2}{Z_i^2 \sigma_i}. \tag{14}$$

If $\sigma_i \ge 1$ and $[\mu_i - \sigma_i, \mu_i + \sigma_i] \subseteq [1, K]$, at least $2\sigma_i - 1 \ge \sigma_i$ grid points lie within one standard deviation. Thus $Z_i \ge e^{-1/2}\sigma_i$. If these conditions hold for every row,

$$\mathcal{E}_{\text{head}}(W) \le \frac{e\sqrt{\pi}}{2} \sum_i a_i^2 \sigma_i^{-3}. \tag{15}$$

Without the containment and width conditions, Eq. (14) remains valid with the exact Z_i.

Scope of the spectral interpretation. The contraction identity and the direct sampled-Gaussian bound are separate results. A sampled Gaussian on a finite path is not generally the exact action of its heat operator; the direct proof does not require that identification. Also, $\|WL\|_F^2$ measures head roughness, including path-boundary terms. It is distinct from the normalized class-center second-difference diagnostic used in E5.

A.3 Proof of Proposition 3

Setting and assumptions. All quantities below are evaluated at step t, with the time index omitted where unambiguous. Consider one population gradient-descent step on backbone parameters ψ, using the current head W. If head parameters are updated simultaneously, W denotes their pre-update effective matrix. Let $\pi_k > 0$ be fixed class priors summing to one, and define

$$\begin{aligned} J(x) &= \frac{\partial f_\psi(x)}{\partial \psi}, \\ \Theta_t(k,k') &= \mathbb{E}[J(x)J(x')^\top \mid y = k, y' = k'], \\ \bar{\Theta}_t &= \frac{1}{K} \sum_k \Theta_t(k,k), \\ \bar{q}_k &= \mathbb{E}[q \mid y = k], \qquad \bar{Q} = [\bar{q}_1, \dots, \bar{q}_K], \qquad \Pi = \operatorname{diag}(\pi_1, \dots, \pi_K). \end{aligned}$$

The samples x, x' are independently drawn conditional on their classes. We assume the expectations and local derivatives exist, permit interchange of differentiation and expectation, and satisfy:

1. The diagonal blocks are close to their mean: $\|\Theta_t(k,k) - \bar{\Theta}_t\|_{\text{op}} \le \varepsilon$.
2. The cross-class blocks are weak: $\|\Theta_t(k,k')\|_{\text{op}} \le \varepsilon_\times$ for $k \ne k'$.

3. Residual fluctuations obey $\mathbb{E}[\|q - \bar{q}_k\|_2^2 \mid y = k] \le \delta^2$, and the kernel–residual factorization error satisfies
$$\Big\|\mathbb{E}\Big[J(x)J(x')^\top W(q(x',y') - \bar{q}_{k'}) \mid y = k, y' = k'\Big]\Big\|_2 \le \kappa\delta$$
for every pair k, k'.

The factorization-error bound is an additional assumption, not a consequence of the variance bound alone. We also assume that the local Taylor expansion has a bounded second-order remainder over this step. This holds, for example, under suitable local Jacobian-Lipschitz bounds and a bounded update direction.

Class-center update. The class-conditional feature gradients satisfy
$$g_k = \mathbb{E}[\partial\ell/\partial z \mid y = k] = W\bar{q}_k, \qquad G_t = W\bar{Q}.$$
The backbone step is
$$\Delta\psi = -\eta \sum_{k'} \pi_{k'} \mathbb{E}[J(x')^\top W q(x',y') \mid y' = k'].$$
Expanding $f_{\psi+\Delta\psi}$ at ψ gives
$$\begin{aligned}\Delta c_k = & -\eta \sum_{k'} \pi_{k'} \mathbb{E}[J(x)J(x')^\top W q(x',y') \mid y = k, y' = k'] \\ & + r_{k,t}^{\mathrm{lin}}.\end{aligned} \tag{16}$$
Let $\mathcal{R}_t^{\mathrm{lin}}$ collect the Taylor remainders. Splitting $q(x',y') = \bar{q}_{k'} + (q(x',y') - \bar{q}_{k'})$, and applying the three assumptions, yields
$$\Delta C_t = -\eta\bar{\Theta}_t W\bar{Q}\Pi + \mathcal{R}_t, \tag{17}$$
where
$$\begin{aligned}\|\mathcal{R}_t\|_F \le & \ \eta\varepsilon\|W\bar{Q}\Pi\|_F \\ & + \eta\sqrt{K}\left(\varepsilon_\times \max_k \|W\bar{q}_k\|_2 + \kappa\delta\right) + \|\mathcal{R}_t^{\mathrm{lin}}\|_F.\end{aligned} \tag{18}$$
Indeed, the diagonal-block error in column k is bounded by $\eta\pi_k\varepsilon\|g_k\|_2$; the cross-class error by $\eta\varepsilon_\times \sum_{k'\neq k} \pi_{k'}\|g_{k'}\|_2$; and the factorization error by $\eta\kappa\delta$. Stacking columns gives Eq. (18). Under the stated local regularity conditions, $\|\mathcal{R}_t^{\mathrm{lin}}\|_F = O(\eta^2)$.

Spectral projection. For a cutoff $0 \le \lambda_\star < \lambda_K$, define
$$\mathcal{S}_\star = \mathrm{span}\{\bar{\Theta}_t W u_n : \lambda_n \le \lambda_\star\},$$
and let $\Pi_\star \in \mathbb{R}^{d\times d}$ be the orthogonal projector onto $\mathcal{S}_\star$. It is distinct from the class-prior matrix $\Pi \in \mathbb{R}^{K\times K}$. Let $U_>$ collect eigenvectors with $\lambda_n > \lambda_\star$. The low-frequency part is annihilated by $I - \Pi_\star$, and hence
$$\begin{aligned}\|(I - \Pi_\star)\Delta C_t\|_F &\le \eta\|\bar{\Theta}_t\|_{\mathrm{op}}\|WU_>U_>^\top\bar{Q}\Pi\|_F + \|\mathcal{R}_t\|_F \\ &\le \eta\|\bar{\Theta}_t\|_{\mathrm{op}}\|WU_>\|_F\|\bar{Q}\Pi\|_F + \|\mathcal{R}_t\|_F.\end{aligned}$$
Therefore,
$$\begin{aligned}\|(I - \Pi_\star)\Delta C_t\|_F &\le \eta\|\bar{\Theta}_t\|_{\mathrm{op}}\|\bar{Q}\Pi\|_F\,\tau(\lambda_\star) + \|\mathcal{R}_t\|_F, \\ \tau(\lambda_\star)^2 &= \sum_{\lambda_n > \lambda_\star} \|Wu_n\|_2^2.\end{aligned} \tag{19}$$
The subspace dimension is at most $|\{n : \lambda_n \le \lambda_\star\}|$. This proves the claimed one-step bound. □

Direct corollary: head energy and update leakage. For $\lambda_\star > 0$,
$$\begin{aligned}\tau(\lambda_\star)^2 &\le \frac{1}{\lambda_\star}\sum_{\lambda_n > \lambda_\star} \lambda_n\|Wu_n\|_2^2 \\ &\le \frac{\mathcal{E}_{\mathrm{head}}(W)}{\lambda_\star}.\end{aligned}$$
Substitution into Eq. (19) gives
$$\|(I - \Pi_\star)\Delta C_t\|_F \le \eta\|\bar{\Theta}_t\|_{\mathrm{op}}\|\bar{Q}\Pi\|_F\sqrt{\frac{\mathcal{E}_{\mathrm{head}}(W)}{\lambda_\star}} + \|\mathcal{R}_t\|_F. \tag{20}$$
For the exact Gaussian rows of Proposition 2, the square-root factor is further bounded by
$$\left(\frac{\sqrt{\pi}}{2\lambda_\star}\sum_i \frac{A_i^2}{\sigma_i}\right)^{1/2}.$$
This explicitly connects Gaussian head-energy bounds to the conditional class-center update bound.

Target term and complete gradient-field roughness. Let $\bar{P}$ collect class-conditional mean predictions. Then $\bar{Q} = \bar{P} - I_K$ and $G_t = W\bar{P} - W$. The target term satisfies $\|(-W)L\|_F^2 \le \lambda_K \mathcal{E}_{\mathrm{head}}(W)$. For the complete field,

$$\begin{aligned} \|G_t L\|_F &\le \|W\bar{P}L\|_F + \|WL\|_F \\ &\le \|W\|_{\mathrm{op}}\|\bar{P}L\|_F + \sqrt{\lambda_K \mathcal{E}_{\mathrm{head}}(W)}. \end{aligned} \tag{21}$$

Thus, the complete field also depends on the label-axis variation of $\bar{P}$. E5 measures feature geometry, not $\|\bar{P}L\|_F$. The class-center update additionally involves class priors, the feature kernel, and the remainder.

Measured head-tail energy. At the reported FG-NET/ResNet-18 checkpoint ($K = 70$, validation-selected configuration, fold 0, seed 0), the normalized unweighted tail $\widehat{\tau}^2 = \sum_{n>m} \|Wu_n\|_2^2 / \|W\|_F^2$ is:

Head	$n > 3$	$n > 5$	$n > 10$	$n > 20$
GFC	3.1×10^{-2}	2.1×10^{-3}	1.3×10^{-4}	1.2×10^{-5}
GFC-PDF	3.5×10^{-2}	1.7×10^{-3}	1.2×10^{-4}	1.2×10^{-5}
Linear+CE	9.0×10^{-1}	8.6×10^{-1}	7.9×10^{-1}	6.6×10^{-1}

The corresponding ratios $\|Wu_K\|_2/\|Wu_1\|_2$ are 8.3×10^{-6}, 8.2×10^{-6}, and 0.58, respectively. These are descriptors of the measured head spectrum.

The square root $\widehat{\tau}$ compares the tail coefficient with $\|W\|_F$ in otherwise identical norm estimates of the leading update term. It is neither the observed fraction of the update outside $\mathcal{S}_\star$ nor the ratio of complete bounds after adding $\|\mathcal{R}_t\|_F$. Also, $\mathbf{1}^\top q = 0$, so the constant label mode u_1 does not contribute to Wq. Normalization by total head energy or by its constant-mode coefficient therefore does not directly quantify effective gradient attenuation. Dirichlet-energy shares additionally weight each mode by λ_n and are distinct from the unweighted tail used here.

Scope of the update bound. The bound is informative when the head-tail term and the remainder are small on the scale of interest. It constrains one update relative to the current feature subspace; it does not control the motion of that subspace over training. E5 provides complementary diagnostics, without directly verifying the kernel assumptions or measuring the remainder. The derivation is for population gradient descent; a quantitative treatment of Adam or stochastic updates would additionally need to account for preconditioning, momentum, or sampling error.

A.4 Envelope Constants for the Head Family

GaussianFC. For center μ_i^{head}, width $\sigma_i^{\mathrm{head}} > 0$, and amplitude a_i,

$$W_{ij} = \frac{a_i}{Z_i} \exp\left(-\frac{(j-\mu_i^{\mathrm{head}})^2}{2(\sigma_i^{\mathrm{head}})^2}\right),$$

$$Z_i = \sum_{j=1}^{K} \exp\left(-\frac{(j-\mu_i^{\mathrm{head}})^2}{2(\sigma_i^{\mathrm{head}})^2}\right).$$

An admissible envelope triple is

$$A_i = \frac{|a_i|}{Z_i}, \qquad \mu_i = \mu_i^{\mathrm{head}}, \qquad \sigma_i = \sigma_i^{\mathrm{head}}.$$

GaussianFC-PDF. The row is

$$W_{ij} = \frac{1}{\sigma_i^{\mathrm{head}}\sqrt{2\pi}} \exp\left(-\frac{(j-\mu_i^{\mathrm{head}})^2}{2(\sigma_i^{\mathrm{head}})^2}\right),$$

so the same center and width apply, with $A_i = (\sigma_i^{\mathrm{head}}\sqrt{2\pi})^{-1}$. This amplitude is determined by the width, removing one parameter per row relative to GaussianFC. The formulas apply to current parameter values, whether a given parameter is fixed or optimized in the experimental configuration.

Relation to the discrete-peak envelope. Definition 1 fixes the envelope center at an integer maximizer ν_i and its amplitude at $|W_{i\nu_i}|$. Denote the resulting minimum width by $\sigma_i^{\mathrm{env}}(W)$. For a nonzero Gaussian row with a unique interior maximizer $1 < \nu_i < K$, let $\phi_i = \mu_i^{\mathrm{head}} - \nu_i$, where $|\phi_i| < 1/2$. For a nonzero integer displacement m within the grid,

$$\frac{|W_{i,\nu_i+m}|}{|W_{i,\nu_i}|} = \exp\left(-\frac{m^2 - 2m\phi_i}{2(\sigma_i^{\mathrm{head}})^2}\right).$$

An envelope centered at ν_i therefore requires

$$s^2 \geq \frac{(\sigma_i^{\mathrm{head}})^2}{1 - 2\phi_i/m}$$

for each such m. The tightest constraint comes from the adjacent position toward the continuous center, giving

$$\sigma_i^{\mathrm{env}}(W) = \frac{\sigma_i^{\mathrm{head}}}{\sqrt{1 - 2|\phi_i|}}. \tag{22}$$

At a half-grid center with two maxima, the discrete-peak convention gives an infinite envelope width although σ_i^{head} remains finite. Boundary cases depend on the available grid positions.

Proposition 1 permits any admissible envelope triple, including the continuous-center constants above. Proposition 2 uses the exact Gaussian row parameters. Neither proposition requires equating σ_i^{head} with the discrete-peak diagnostic.

A.5 Reading the Three Bounds

Gradient contributions. Proposition 1 separates the weight envelope from the residual mass produced by a trained predictor. Both factors matter: for a center aligned with the target, $K = 70$, $\sigma_i = 0.30K$, and $r = 7$, the envelope multiplier is $\exp(-1/18) \approx 0.946$. Such a multiplier alone cannot explain a small observed gradient-mass radius. The bound neither excludes far-label errors nor establishes the necessity of a Gaussian head.

Head structure. Proposition 2 gives explicit amplitude- and width-dependent bounds. At fixed envelope amplitude, the bound scales as σ_i^{-1}; the two normalization schemes lead to the variants derived above, with the stated conditions. Width thus couples concentration and the smoothness bound, without implying that both improve monotonically together. Raw energy also depends on feature scale: replacing z by cz and W by W/c preserves logits but changes head energy by c^{-2}. Row-wise profile shape is defined in the learned feature coordinates. These quantities describe head structure rather than independent evidence of predictive benefit.

Conditional representation updates. Proposition 3 and Eq. (20) connect head energy to a one-step feature-subspace bound. The directions are obtained by transforming label modes through W and the feature kernel; their coefficients depend on the residual field and class priors. The result does not imply that class centers inherit the order of an individual label eigenvector.

Connection to empirical geometry. E5 examines average distance–gap relationships, normalized class-center roughness, and ordinal readout from leading PCA subspaces. These diagnostics characterize complementary aspects of the learned representation. PCA directions are not label-Laplacian modes, and improved readout with additional principal components does not rule out another one-dimensional readout direction.

B Implications and connection to Ordinal Neural Collapse

What the three propositions say jointly. Prop. 1 bounds the spatial extent of the supervision signal sent to the backbone, Prop. 2 bounds the head's own roughness and curvature in closed form, and Prop. 3 turns both into class-center dynamics through the two head-side quantities read off in §A.5: the dimension of the subspace the update can occupy and the curvature of the field along the label axis. Together this is a head $\rightarrow$ gradient $\rightarrow$ feature-geometry chain that is causal in the direction of optimization, not a static endpoint description.

The curvature statement. The corollary in §A.3 is the part of the chain that needs no convergence argument. Since $\bar{q}_k = \bar{p}_k - e_k$ we have $\bar{Q} = \bar{P} - I_K$ and hence $G_t = W\bar{P} - W$: the target-driven half of the class-conditional gradient field *is* the head matrix, so its label-axis curvature obeys $\mathrm{Tr}(WL^2W^\top) \leq \lambda_K \, \mathcal{E}_{\mathrm{head}}(W)$ with $\lambda_K < 4$, and Prop. 2 places $\mathcal{E}_{\mathrm{head}}$ on the $O(\sum_i a_i^2 \sigma_i^{-3})$ scale. The width σ_i that makes the head smooth therefore makes the field it sends to the backbone smooth in the same sense—at every step, for every residual field, and without any assumption about where training ends. The bound is a statement about the field; E5(c) measures its effect on the centers. The two architecture-side heads are the smoothest on ViT-Tiny (GFC-PDF 0.896, GFC 0.903, within 0.007 of each other, as a shared scale with different constants predicts), and, unlike the loss- and posterior-side priors, they improve monotonically as backbone capacity grows (GFC $0.956 \rightarrow 0.904 \rightarrow 0.903$, GFC-PDF $0.934 \rightarrow 0.925 \rightarrow 0.896$ across ResNet-18 / ResNet-50 / ViT-Tiny).

Connection to ONC. The Ordinal Neural Collapse (ONC) literature characterizes the asymptotic geometry of class means in cumulative-link ordinal regression: at convergence, class means concentrate on a one-dimensional ordered subspace whose direction is monotone in the label index. The propositions above give an optimization-side account with a *different* prediction. On dimension, Eq. (5) bounds the part of the class-center

update lying outside $\mathcal{S}_\star = \mathrm{span}\{\bar{\Theta}_t W u_n : \lambda_n \le \lambda_\star\}$; that subspace has dimension at most $|\{n : \lambda_n \le \lambda_\star\}|$—low-dimensional, but not one-dimensional. On monotonicity, u_2 is the only non-constant path-graph mode that is monotone in the label index, so a monotone component can enter only through it, while the remaining low-frequency modes carry ordered but non-monotone structure. E5(b) is the measurement that separates the two pictures: the posterior-side prior PU saturates at k=1, as an endpoint collapse onto one direction implies, while the GaussianFC family starts lower and climbs, GFC-PDF overtaking PU at k=3 (0.881 vs. 0.878). ONC describes where cumulative-link models end up; the analysis here describes what the head-induced filter admits along the way, and the two do not coincide.

Design implications. In this family locality and smoothness are not independent knobs: the same σ_i sets the envelope of Prop. 1 and the energy constant of Prop. 2, so a head cannot be made local without also being made smooth. The remaining design freedom is how tightly the amplitude is coupled to the width—$A_i = |a_i|/Z_i$ with a_i learned for GFC, amplitude determined by the width for GFC-PDF—and the propositions place both on one scale differing only in the constant. The dissociation of the two mechanisms is therefore carried by the baselines rather than by a third Gaussian variant: PU realizes Mechanism B alone and Linear+CE neither, and each has a failure regime (Table 1). A head whose smoothing strength is decoupled from σ_i would make the two mechanisms independently tunable, and the bounds above say exactly which two quantities to read off it: the envelope width of Eq. (3) and the energy constant of Eq. (4).

C Experiments

C.1 Experimental Protocol

Datasets. We evaluate on five ordinal-regression datasets spanning two label-cardinality regimes. **FG-NET** contains 1,002 facial photographs of 82 subjects annotated with age 0–69 ($K = 70$). **MORPH** (Album 2) contains 50,015 facial photographs of 19,033 subjects annotated with age 16–77 ($K = 62$, label = age $- 16$), at 2.6 images per subject on average (median 2, maximum 53). It is the large-N, large-K counterpart to FG-NET: the same task and the same label-axis structure, but $50\times$ the images and $232\times$ the subjects. **FG-NET-5** is the same 1,002 images re-labelled into 5 equal-frequency age bands with cut-points at ages 5/10/16/25 ($K = 5$, class sizes 233/178/201/199/191); the quantile edges are computed once over the full set and shared by all folds, so the coarsening is a controlled manipulation of K alone—the images, subjects, and splits are identical to FG-NET. **BACH** (ICIAR-2018) contains 400 breast-histology images in four clinically ordered grades, Normal $<$ Benign $<$ InSitu $<$ Invasive ($K = 4$, 100 per class); we fix this order explicitly rather than sorting class-directory names alphabetically, which would place *Normal* last and destroy the ordinal axis. **HCI** (Historical Colour, ECCV 2012) contains 1,325 photographs across five decades, 1930s–1970s ($K = 5$, 265 per decade). **APTOS-2019** contains 3,662 retinal fundus photographs graded on the 5-point ICDR scale ($K = 5$); the grades are heavily imbalanced (1805/370/999/193/295) and we report per-fold class counts rather than assuming the imbalance away. For APTOS we fix $K = 5$ from the ICDR scale rather than inferring it from `max(label)+1`, so that a fold missing the rarest grade cannot silently shrink the label axis the head is built on.

Splits. All datasets use the same protocol: an outer 5-fold partition with a **nested validation set** held out from within each outer training fold. For FG-NET (and hence FG-NET-5), the outer partition is **subject-exclusive**: folds are formed by `GroupKFold` on subject identity (the first three characters of the filename), so no subject contributes images to more than one of train/validation/test. This matters because FG-NET averages 12 photographs per subject; an image-level split would let the network recognise the individual rather than estimate age. The inner validation set is drawn by `GroupShuffleSplit` on the same subject groups. The resulting per-fold sizes are 675/127/200 images over 56/10/16 subjects. MORPH uses the same subject-exclusive protocol, with the subject id taken from the filename prefix. Here it is not merely preferable but necessary: the folders the dataset ships with cannot be used as a split at all, since they share 3,530 (Train–Test), 3,565 (Train–Validation) and 1,014 (Validation–Test) subjects pairwise. We therefore treat those folders purely as storage locations and re-partition from scratch, giving per-fold sizes of 33,905–34,080 / 5,932–6,107 / 10,003 images over 12,942–12,943 / 2,284–2,285 / 3,805–3,807 subjects. BACH, HCI and APTOS have no subject structure—one image per patient or per source photograph—so every image forms its own group and `GroupKFold` degenerates to ordinary K-fold; per-fold sizes are 272/48/80, 901/159/265 and 2490/440/732 respectively. Splits are generated once, written to JSON as explicit file lists, and will be released with the code upon acceptance, so every number in this paper is reproducible from fixed artefacts rather than from a re-run of a random generator.

Runs and seeds. Each configuration is trained on all 5 folds $\times$ 5 seeds (0–4) = **25 runs**. Seeds control weight initialisation, data ordering and augmentation; the splits themselves are fixed and shared across seeds, so a (fold, seed) pair identifies exactly one run and the same pair is comparable across methods. Hyperparameter search uses a **disjoint seed set** (100–102) so that the seeds used to select a configuration never appear among the seeds used to report it. Every model is trained for 300 or 100 (only on MORPH Dataset) epochs with **early stopping** (`patience` $= 20$), and the reported checkpoint is the one minimising validation MAE.

Learning rate and hyperparameter grids. The learning rate is **selected per method on validation**, never shared globally: a fixed rate systematically under-tunes baselines whose optimum lies elsewhere, and in our audit this was the single largest source of inflated head effects. The grid is five values, $\{1, 5\} \times 10^{-5}$, $\{1, 5\} \times 10^{-4}$, and 10^{-3}, applied identically to every method in every cell.

Beyond the learning rate, the Gaussian heads sweep the relative width $\sigma/K \in \{0.1, 0.2, \ldots, 1.5\}$; **DLDL** sweeps its soft-label width $\sigma \in \{0.1, 0.2, \ldots, 2.0\}$; **UR** uses $\eta \in \{0.1, 0.15, 0.2, 0.25\}$, $\tau \in \{0.1, 0.5, 1.0, 2.0\}$; **PU** uses $T \in \{0.1, 0.15, 0.2, 0.25\}$; and RNC uses $\tau \in \{0.5, 0.7, 0.8, 1.0, 1.5, 2.0\}$. The linear baselines (**Linear+CE**, **CORAL**, **CORN**) have no method-specific hyperparameter, so the learning rate is their entire search space. Budgets are therefore deliberately unequal—a method with more genuine degrees of freedom is allowed to search more—which is sound because selection happens on the nested validation set and never on test.

Selection discipline. The test fold is used **exactly once per configuration**, to produce the reported number. All model selection—hyperparameters and learning rate—uses the nested validation set only. Where two variants of the same function differ only in parameterisation (the two σ-initialisations of the PDF head), the choice between them is likewise made on validation, never on test.

Statistical analysis. Methods are compared pairwise on matched (fold, seed) pairs: for a given cell, the 25 runs of method A and the 25 runs of method B use the same five subject-exclusive test folds and, within each fold, the same five seeds for the inner train/validation re-draw, so the difference series isolates the method effect from fold difficulty. Fold difficulty varies substantially — on FG-NET the mean MAE spans 1.25–1.38 across the five folds, three to seven times the median effect under test (0.17–0.42) — so an unpaired comparison would be dominated by which folds a method happened to draw. We report the Wilcoxon signed-rank test on the paired differences rather than a paired t-test, because the difference distribution is heavy-tailed — a single hard fold can dominate a mean — and the rank statistic is insensitive to that. Because five seeds share a test fold, the 25 differences are not fully independent, but pairing removes the shared fold component: the median intra-fold ICC of the difference series is 0.016 (design effect 1.07, effective n 23.5). p-values are corrected with Holm–Bonferroni, the family being the full set of tests displayed in the corresponding table.

C.2 Main Results (Experiment 1)

Table C.1: All metrics for every (cell, method). RPS and NLL are recomputed from each run's stored test probabilities; Acc and QWK are secondary metrics under val-MAE selection.

Cell	K	Method	test MAE	± se	RPS	NLL	test Acc%	test QWK	val MAE
BACH / R18@128	4	CORN	0.3055	0.0134	0.081908	0.6606	74.20	0.7590	0.2283
BACH / R18@128	4	DLDL	0.3085	0.0099	0.079769	0.6761	71.17	0.7747	0.2400
BACH / R18@128	4	PDF_r	0.3110	0.0129	0.088969	0.7797	73.60	0.8259	0.2217
BACH / R18@128	4	UR	0.3122	0.0140	0.122270	0.9571	73.91	0.7705	0.2214
BACH / R18@128	4	CE	0.3160	0.0126	0.082098	0.6792	74.62	0.7451	0.2275
BACH / R18@128	4	GFC	0.3275	0.0110	0.091368	0.9004	71.15	0.8230	0.2575
BACH / R18@128	4	RNC+CE	0.4130	0.0142	0.108340	0.9116	63.07	0.7154	0.3392
BACH / R18@128	4	RNC+L1	0.4688	0.0127	—	—	56.24	0.7138	0.4165
BACH / R18@128	4	PU	0.5116	0.0124	0.102783	0.9361	60.40	0.7445	0.4350
BACH / R18@128	4	CORAL	0.6745	0.0219	0.157685	1.7797	48.11	0.6263	0.5333
BACH / R50@128	4	PDF_r	0.2770	0.0153	0.073761	0.9171	78.40	0.8395	0.1833
BACH / R50@128	4	UR	0.2797	0.0140	0.119816	0.9542	76.09	0.7850	0.1794
BACH / R50@128	4	CORN	0.2910	0.0125	0.079814	0.8012	74.43	0.7732	0.1767
BACH / R50@128	4	GFC	0.3040	0.0160	0.084602	1.0704	75.05	0.8263	0.1983
BACH / R50@128	4	CE	0.3295	0.0168	0.089638	0.8827	73.58	0.7383	0.1867
BACH / R50@128	4	RNC+CE	0.3530	0.0225	0.094641	0.8191	70.08	0.7363	0.3233
BACH / R50@128	4	RNC+L1	0.3702	0.0246	—	—	66.36	0.7500	0.3649
BACH / R50@128	4	PU	0.5474	0.0133	0.105319	0.9326	58.05	0.7258	0.4183
BACH / R50@128	4	CORAL	0.5730	0.0144	0.137833	1.6209	51.71	0.6973	0.4550
BACH / R50@128	4	DLDL	0.6585	0.0093	0.169631	1.2641	33.86	0.5115	0.6133
BACH / ViT@128	4	CORN	0.3310	0.0118	0.093610	0.8780	72.92	0.7375	0.2175
BACH / ViT@128	4	CE	0.3620	0.0122	0.097538	0.9252	71.45	0.7154	0.2175
BACH / ViT@128	4	UR	0.3727	0.0171	0.126688	0.9746	69.59	0.7279	0.2546
BACH / ViT@128	4	GFC	0.3855	0.0150	0.110162	1.3182	71.90	0.7443	0.2475
BACH / ViT@128	4	PDF_r	0.3856	0.0165	0.113951	1.3146	71.15	0.7068	0.2408
BACH / ViT@128	4	RNC+CE	0.4400	0.0161	0.109637	0.9122	62.31	0.7066	0.3892
BACH / ViT@128	4	RNC+L1	0.4791	0.0178	—	—	56.33	0.7053	0.4191

Cell	K	Method	test MAE	± se	RPS	NLL	test Acc%	test QWK	val MAE
BACH / ViT@128	4	PU	0.5207	0.0102	0.103973	0.9432	59.95	0.7385	0.4025
BACH / ViT@128	4	DLDL	0.7435	0.0104	0.179561	1.2981	27.27	0.4382	0.6683
BACH / ViT@128	4	CORAL	0.7695	0.0134	0.178671	2.1428	45.88	0.5739	0.6742
APTOS / R18@128	5	DLDL	0.2478	0.0018	0.051120	0.6636	75.86	0.8402	0.2240
APTOS / R18@128	5	GFC	0.2498	0.0034	0.059265	1.0927	79.79	0.8874	0.2279
APTOS / R18@128	5	PDF_r	0.2501	0.0035	0.062501	0.6958	80.38	0.8847	0.2257
APTOS / R18@128	5	UR	0.2504	0.0028	0.095058	0.9449	77.60	0.8347	0.2367
APTOS / R18@128	5	RNC+CE	0.2560	0.0034	0.055111	0.8637	75.52	0.8359	0.2365
APTOS / R18@128	5	CORN	0.2635	0.0033	0.054837	0.7503	76.27	0.8237	0.2309
APTOS / R18@128	5	CE	0.2728	0.0038	0.056562	0.8486	76.04	0.8148	0.2389
APTOS / R18@128	5	RNC+L1	0.2757	0.0037	—	—	73.98	0.8385	0.2685
APTOS / R18@128	5	CORAL	0.2847	0.0054	0.065564	0.8957	74.53	0.8185	0.2575
APTOS / R18@128	5	PU	0.3555	0.0074	0.060498	0.7645	73.13	0.8469	0.3240
APTOS / R50@128	5	RNC+CE	0.2428	0.0032	0.052076	0.7364	76.71	0.8400	0.2245
APTOS / R50@128	5	DLDL	0.2444	0.0031	0.051697	0.6732	76.51	0.8383	0.2150
APTOS / R50@128	5	PDF_r	0.2447	0.0045	0.050995	1.0390	80.53	0.8876	0.2112
APTOS / R50@128	5	GFC	0.2453	0.0038	0.053588	1.0899	80.35	0.8879	0.2115
APTOS / R50@128	5	CORN	0.2476	0.0030	0.051439	0.7271	76.84	0.8340	0.2127
APTOS / R50@128	5	UR	0.2489	0.0041	0.092397	0.9307	77.64	0.8360	0.2281
APTOS / R50@128	5	RNC+L1	0.2535	0.0029	—	—	75.89	0.8454	0.2413
APTOS / R50@128	5	CE	0.2585	0.0034	0.055298	0.8708	76.86	0.8249	0.2253
APTOS / R50@128	5	CORAL	0.2670	0.0054	0.058760	0.8318	75.48	0.8288	0.2385
APTOS / R50@128	5	PU	0.3381	0.0056	0.060310	0.7650	73.87	0.8527	0.3032
APTOS / ViT@128	5	DLDL	0.2590	0.0030	0.052706	0.6552	74.90	0.8348	0.2365
APTOS / ViT@128	5	PDF_r	0.2627	0.0035	0.051829	0.7289	79.13	0.8785	0.2371
APTOS / ViT@128	5	GFC	0.2644	0.0030	0.054478	0.8290	79.03	0.8771	0.2365
APTOS / ViT@128	5	RNC+CE	0.2754	0.0026	0.055601	0.7481	74.48	0.8240	0.2592
APTOS / ViT@128	5	CE	0.2755	0.0034	0.058002	0.8824	75.49	0.8160	0.2467
APTOS / ViT@128	5	CORN	0.2761	0.0041	0.057623	0.8124	75.21	0.8186	0.2391
APTOS / ViT@128	5	UR	0.2788	0.0040	0.095838	0.9520	75.78	0.8200	0.2529
APTOS / ViT@128	5	RNC+L1	0.3092	0.0022	—	—	70.96	0.8284	0.2954
APTOS / ViT@128	5	CORAL	0.3384	0.0055	0.085498	1.3162	71.98	0.7931	0.3203
APTOS / ViT@128	5	PU	0.3509	0.0038	0.062332	0.8022	73.16	0.8464	0.3292
APTOS / ViT@224	5	UR	0.2261	0.0022	0.091139	0.9241	78.50	0.8517	0.2038
APTOS / ViT@224	5	GFC	0.2309	0.0026	0.050286	0.8544	81.75	0.8929	0.2045
APTOS / ViT@224	5	CORN	0.2317	0.0031	0.049307	0.7354	77.56	0.8455	0.2050
APTOS / ViT@224	5	PDF_r	0.2351	0.0034	0.047456	0.7613	81.60	0.8899	0.2002
APTOS / ViT@224	5	CE	0.2362	0.0028	0.049491	0.7379	77.73	0.8394	0.2050
APTOS / ViT@224	5	RNC+CE	0.2366	0.0030	0.047927	0.6407	76.83	0.8457	0.2232
APTOS / ViT@224	5	RNC+L1	0.2688	0.0036	—	—	73.48	0.8495	0.2623
APTOS / ViT@224	5	CORAL	0.2969	0.0061	0.077722	1.2237	73.86	0.8204	0.2785
APTOS / ViT@224	5	PU	0.3123	0.0054	0.056404	0.7292	75.10	0.8733	0.2836
APTOS / ViT@224	5	DLDL	0.7537	0.0048	0.135726	1.2877	27.67	0.5877	0.7505
FG-NET-5 / R18@128	5	PDF_r	0.4639	0.0094	0.099964	1.0135	59.82	0.8535	0.3795
FG-NET-5 / R18@128	5	RNC+CE	0.4946	0.0104	0.091231	1.0863	52.58	0.7935	0.4203
FG-NET-5 / R18@128	5	DLDL	0.4955	0.0123	0.095779	1.0776	54.17	0.7852	0.3980
FG-NET-5 / R18@128	5	RNC+L1	0.5069	0.0101	—	—	52.28	0.7956	0.4473
FG-NET-5 / R18@128	5	GFC	0.5074	0.0114	0.100186	1.2384	56.86	0.8263	0.3945
FG-NET-5 / R18@128	5	UR	0.5648	0.0113	0.124445	1.2281	51.68	0.7602	0.4754
FG-NET-5 / R18@128	5	CORN	0.5734	0.0108	0.116487	1.7334	51.45	0.7402	0.4508
FG-NET-5 / R18@128	5	CE	0.5852	0.0127	0.116051	1.7754	51.09	0.7364	0.4790
FG-NET-5 / R18@128	5	PU	0.6007	0.0123	0.096269	1.0880	53.19	0.8075	0.5148
FG-NET-5 / R18@128	5	CORAL	0.6378	0.0116	0.121608	1.7314	47.54	0.7574	0.5330
FG-NET-5 / R50@128	5	PDF_r	0.4868	0.0135	0.095183	1.5140	57.85	0.8361	0.3735
FG-NET-5 / R50@128	5	GFC	0.4898	0.0118	0.101401	1.6238	57.63	0.8329	0.3907
FG-NET-5 / R50@128	5	RNC+CE	0.4989	0.0129	0.100129	1.2862	53.66	0.7884	0.4388
FG-NET-5 / R50@128	5	UR	0.5014	0.0153	0.115107	1.1728	55.74	0.7867	0.4125
FG-NET-5 / R50@128	5	DLDL	0.5074	0.0127	0.096279	1.0711	53.46	0.7814	0.3999
FG-NET-5 / R50@128	5	RNC+L1	0.5124	0.0129	—	—	53.29	0.7908	0.4387
FG-NET-5 / R50@128	5	CORN	0.5366	0.0123	0.107043	1.3884	53.26	0.7613	0.4375
FG-NET-5 / R50@128	5	CE	0.5705	0.0131	0.108883	1.4227	52.88	0.7399	0.4595
FG-NET-5 / R50@128	5	PU	0.6149	0.0127	0.093950	1.0670	52.39	0.8067	0.5085
FG-NET-5 / R50@128	5	CORAL	0.6570	0.0208	0.126903	1.9568	47.44	0.7505	0.5412

Cell	K	Method	test MAE	± se	RPS	NLL	test Acc%	test QWK	val MAE
FG-NET-5 / ViT@128	5	PDF_r	0.4762	0.0098	0.085809	1.0775	57.97	0.8464	0.3896
FG-NET-5 / ViT@128	5	GFC	0.4780	0.0106	0.093328	1.2702	57.58	0.8449	0.3894
FG-NET-5 / ViT@128	5	DLDL	0.4782	0.0108	0.089243	1.0371	54.18	0.7980	0.3927
FG-NET-5 / ViT@128	5	RNC+CE	0.4913	0.0103	0.093188	1.1855	53.01	0.7945	0.4329
FG-NET-5 / ViT@128	5	CORN	0.5019	0.0123	0.103544	1.5876	54.33	0.7819	0.4162
FG-NET-5 / ViT@128	5	UR	0.5336	0.0146	0.115604	1.1793	52.95	0.7747	0.4525
FG-NET-5 / ViT@128	5	RNC+L1	0.5405	0.0130	—	—	50.38	0.7842	0.4690
FG-NET-5 / ViT@128	5	CE	0.5469	0.0092	0.107803	1.5831	52.57	0.7617	0.4536
FG-NET-5 / ViT@128	5	PU	0.6055	0.0123	0.092110	1.0592	52.56	0.8148	0.5084
FG-NET-5 / ViT@128	5	CORAL	0.7785	0.0212	0.146123	1.8610	43.21	0.6787	0.6656
HCI / R18@128	5	PDF_r	0.7879	0.0118	0.149036	1.6037	45.24	0.6087	0.7157
HCI / R18@128	5	DLDL	0.7882	0.0088	0.149565	1.4230	41.37	0.5637	0.7343
HCI / R18@128	5	RNC+CE	0.7922	0.0068	0.155369	1.8719	40.72	0.5753	0.7577
HCI / R18@128	5	GFC	0.7961	0.0146	0.155102	1.4983	44.75	0.6020	0.7248
HCI / R18@128	5	CORN	0.8181	0.0107	0.153831	1.5904	44.88	0.5511	0.7258
HCI / R18@128	5	RNC+L1	0.8334	0.0072	—	—	37.27	0.5675	0.8268
HCI / R18@128	5	UR	0.8523	0.0116	0.153338	1.3662	43.62	0.5331	0.7832
HCI / R18@128	5	CE	0.8720	0.0122	0.159966	1.7406	45.68	0.5114	0.7816
HCI / R18@128	5	PU	0.9687	0.0141	0.145330	1.3981	40.88	0.5722	0.8684
HCI / R18@128	5	CORAL	1.0113	0.0248	0.178484	2.1403	37.27	0.5239	0.9077
HCI / R50@128	5	GFC	0.7810	0.0136	0.156604	1.7186	45.58	0.6059	0.7230
HCI / R50@128	5	DLDL	0.7879	0.0098	0.152986	1.4568	43.33	0.5614	0.7026
HCI / R50@128	5	UR	0.8029	0.0100	0.145397	1.3155	47.41	0.5573	0.7333
HCI / R50@128	5	PDF_r	0.8057	0.0126	0.160311	2.4709	44.60	0.5854	0.7187
HCI / R50@128	5	RNC+CE	0.8402	0.0107	0.173847	1.9236	44.17	0.5327	0.8186
HCI / R50@128	5	RNC+L1	0.8419	0.0113	—	—	40.88	0.5602	0.8355
HCI / R50@128	5	CORN	0.8519	0.0151	0.168912	1.8850	45.31	0.5097	0.7353
HCI / R50@128	5	CE	0.8740	0.0114	0.173276	2.1302	45.93	0.5068	0.7376
HCI / R50@128	5	CORAL	0.9396	0.0218	0.171488	2.0679	38.47	0.5491	0.8294
HCI / R50@128	5	PU	0.9862	0.0151	0.146617	1.3863	39.88	0.5538	0.8795
HCI / ViT@128	5	DLDL	0.7211	0.0084	0.138919	1.4029	45.06	0.6135	0.6870
HCI / ViT@128	5	GFC	0.7263	0.0099	0.149762	2.0808	46.58	0.6547	0.6684
HCI / ViT@128	5	PDF_r	0.7374	0.0102	0.130628	1.3315	47.52	0.6601	0.6810
HCI / ViT@128	5	UR	0.7751	0.0107	0.142666	1.3036	47.08	0.5876	0.7198
HCI / ViT@128	5	CORN	0.7834	0.0105	0.160161	2.0887	47.21	0.5707	0.6923
HCI / ViT@128	5	RNC+CE	0.8106	0.0081	0.155527	1.8179	39.40	0.5737	0.7879
HCI / ViT@128	5	CE	0.8294	0.0129	0.167495	2.2733	46.03	0.5300	0.7341
HCI / ViT@128	5	RNC+L1	0.8552	0.0071	—	—	36.21	0.5657	0.8574
HCI / ViT@128	5	PU	0.8813	0.0138	0.135529	1.3432	42.75	0.6205	0.8108
HCI / ViT@128	5	CORAL	0.9982	0.0205	0.179541	1.9095	37.25	0.5108	0.8843
HCI / ViT@224	5	UR	0.6668	0.0079	0.134595	1.2571	51.48	0.6523	0.6189
HCI / ViT@224	5	GFC	0.7043	0.0108	0.154087	2.8738	50.91	0.6578	0.6352
HCI / ViT@224	5	CORN	0.7251	0.0098	0.144447	1.8157	49.84	0.6103	0.6460
HCI / ViT@224	5	RNC+CE	0.7346	0.0106	0.139612	1.6424	44.34	0.6274	0.6979
HCI / ViT@224	5	PDF_r	0.7380	0.0100	0.153049	2.6685	51.65	0.6193	0.6616
HCI / ViT@224	5	CE	0.7423	0.0113	0.149751	1.9491	50.12	0.5953	0.6644
HCI / ViT@224	5	RNC+L1	0.7636	0.0076	—	—	40.54	0.6335	0.7359
HCI / ViT@224	5	PU	0.8470	0.0170	0.124211	1.2366	44.33	0.6646	0.7776
HCI / ViT@224	5	DLDL	0.9236	0.0062	0.155368	1.4476	24.27	0.4305	0.9200
HCI / ViT@224	5	CORAL	0.9894	0.0228	0.179630	2.0315	38.44	0.5279	0.8901
MORPH / R18@128	62	CORAL	2.7639	0.0086	0.033878	2.8048	12.73	0.8877	2.7419
MORPH / R18@128	62	PU	2.8100	0.0102	0.033013	2.6811	13.24	0.9382	2.7517
MORPH / R18@128	62	GFC	2.8380	0.0121	0.037875	4.1325	13.04	0.9354	2.7994
MORPH / R18@128	62	RNC+CE	2.8444	0.0090	0.039304	4.8298	12.11	0.8863	2.8281
MORPH / R18@128	62	RNC+L1	2.8473	0.0102	—	—	12.49	0.8917	2.9014
MORPH / R18@128	62	PDF_r	2.8566	0.0119	0.034694	3.0369	12.84	0.9349	2.8184
MORPH / R18@128	62	DLDL	2.9324	0.0096	0.035855	2.8194	11.53	0.8827	2.9238
MORPH / R18@128	62	CORN	2.9894	0.0119	0.036411	2.9656	11.98	0.8797	2.9820
MORPH / R18@128	62	UR	3.0307	0.0123	0.033720	2.6794	13.34	0.8801	3.0759
MORPH / R18@128	62	CE	3.1314	0.0126	0.035859	2.8581	11.46	0.8736	3.1146
MORPH / R50@128	62	GFC	2.6938	0.0097	0.037378	4.8802	13.79	0.9413	2.6721
MORPH / R50@128	62	PDF_r	2.7137	0.0100	0.032870	2.9959	13.50	0.9405	2.6973
MORPH / R50@128	62	CORAL	2.7334	0.0111	0.033786	2.8269	12.99	0.8886	2.7076

Cell	K	Method	test MAE	± se	RPS	NLL	test Acc%	test QWK	val MAE
MORPH / R50@128	62	PU	2.7497	0.0105	0.032337	2.6661	13.50	0.9406	2.7005
MORPH / R50@128	62	RNC+L1	2.8108	0.0131	—	—	12.44	0.8932	2.8794
MORPH / R50@128	62	DLDL	2.9095	0.0089	0.035966	2.8638	11.57	0.8826	2.8746
MORPH / R50@128	62	UR	2.9138	0.0090	0.033438	2.6994	12.60	0.8864	2.9817
MORPH / R50@128	62	CORN	2.9162	0.0093	0.035431	2.9168	11.95	0.8837	2.9129
MORPH / R50@128	62	CE	3.1674	0.0100	0.035598	2.8985	11.47	0.8729	3.1523
MORPH / R50@128	62	RNC+CE	3.3866	0.0178	0.037583	4.4601	13.13	0.9364	3.3884
MORPH / ViT@128	62	GFC	2.9975	0.0100	0.040063	4.0291	12.27	0.9284	2.9660
MORPH / ViT@128	62	CORAL	3.0222	0.0105	0.036928	2.8941	11.58	0.8773	2.9962
MORPH / ViT@128	62	DLDL	3.0262	0.0093	0.037214	2.8767	11.03	0.8780	3.0110
MORPH / ViT@128	62	PU	3.0704	0.0111	0.035990	2.7678	12.06	0.9270	3.0077
MORPH / ViT@128	62	RNC+L1	3.0867	0.0099	—	—	11.33	0.8833	3.1567
MORPH / ViT@128	62	CORN	3.1056	0.0112	0.037718	2.9916	11.31	0.8747	3.0988
MORPH / ViT@128	62	PDF_r	3.1268	0.0109	0.035761	2.7683	11.88	0.9221	3.1068
MORPH / ViT@128	62	RNC+CE	3.1385	0.0100	0.038351	3.2779	11.04	0.8759	3.1264
MORPH / ViT@128	62	UR	3.2023	0.0096	0.035695	2.7398	11.34	0.8754	3.2656
MORPH / ViT@128	62	CE	3.3710	0.0100	0.037877	2.9891	10.96	0.8627	3.3679
FG-NET / R18@128	70	RNC+CE	4.4824	0.0842	0.050598	3.6868	12.50	0.7789	3.8129
FG-NET / R18@128	70	RNC+L1	4.5652	0.0700	—	—	11.36	0.7862	4.1186
FG-NET / R18@128	70	PU	4.6302	0.0875	0.049436	3.1896	12.60	0.8197	3.6928
FG-NET / R18@128	70	GFC	4.7260	0.0924	0.058654	3.2641	10.17	0.8286	3.9264
FG-NET / R18@128	70	PDF_r	4.7440	0.1041	0.053860	3.2307	10.87	0.8120	3.7622
FG-NET / R18@128	70	DLDL	4.9433	0.1226	0.057114	3.3926	9.64	0.7354	3.9524
FG-NET / R18@128	70	UR	5.1744	0.1274	0.050317	3.0971	11.74	0.7377	4.1084
FG-NET / R18@128	70	CORN	5.1862	0.1125	0.054295	3.4461	11.57	0.7659	4.3602
FG-NET / R18@128	70	CORAL	5.7049	0.1451	0.072255	4.4624	8.64	0.7177	5.1787
FG-NET / R18@128	70	CE	6.7496	0.1755	0.071835	5.2422	9.65	0.5720	5.3839
FG-NET / R18@224	70	RNC+CE	4.0920	0.0820	0.044582	3.3965	13.05	0.8080	3.3911
FG-NET / R18@224	70	RNC+L1	4.2513	0.0834	—	—	10.40	0.8098	3.7526
FG-NET / R18@224	70	PDF_r	4.2580	0.0977	0.051337	3.1828	10.32	0.8558	3.5434
FG-NET / R18@224	70	PU	4.3062	0.0818	0.045561	3.0452	13.63	0.8514	3.4468
FG-NET / R18@224	70	DLDL	4.3367	0.1201	0.047652	3.1086	9.56	0.7895	3.7126
FG-NET / R18@224	70	CORN	4.5310	0.1111	0.049137	3.3837	11.29	0.7795	3.8828
FG-NET / R18@224	70	GFC	4.5321	0.0967	0.047293	2.6183	11.26	0.8485	3.6443
FG-NET / R18@224	70	UR	4.6799	0.1484	0.045556	3.0165	12.88	0.7767	3.7449
FG-NET / R18@224	70	CORAL	5.4010	0.1582	0.067207	4.3365	9.04	0.7424	4.8188
FG-NET / R18@224	70	CE	6.1318	0.1819	0.059487	4.2751	11.33	0.6099	4.9893
FG-NET / R50@128	70	GFC	4.2955	0.1146	0.049379	3.3587	13.65	0.8445	3.5241
FG-NET / R50@128	70	PDF_r	4.4606	0.1227	0.049912	3.3424	11.26	0.8221	3.7376
FG-NET / R50@128	70	RNC+CE	4.5065	0.1177	0.051034	3.9271	12.82	0.7722	3.9302
FG-NET / R50@128	70	PU	4.5439	0.1057	0.048458	3.1553	12.60	0.8302	3.5833
FG-NET / R50@128	70	DLDL	4.7786	0.1379	0.054687	3.3328	9.17	0.7541	3.8036
FG-NET / R50@128	70	UR	4.8709	0.1241	0.049619	3.0935	11.72	0.7586	4.0650
FG-NET / R50@128	70	RNC+L1	5.2376	0.1288	—	—	11.88	0.8383	4.7146
FG-NET / R50@128	70	CORN	5.5206	0.1152	0.061205	4.3006	10.48	0.7032	4.3931
FG-NET / R50@128	70	CORAL	5.5829	0.1606	0.069192	4.3454	9.40	0.7312	5.0306
FG-NET / R50@128	70	CE	6.3927	0.1353	0.067998	5.3068	9.99	0.6153	5.2290
FG-NET / R50@224	70	RNC+L1	4.2570	0.0815	—	—	11.82	0.8182	3.7619
FG-NET / R50@224	70	GFC	4.2772	0.1200	0.048440	3.2348	12.83	0.8448	3.3545
FG-NET / R50@224	70	PU	4.2969	0.1136	0.045426	3.0489	13.26	0.8545	3.4873
FG-NET / R50@224	70	RNC+CE	4.3176	0.1175	0.047354	3.6131	12.57	0.7909	3.6313
FG-NET / R50@224	70	DLDL	4.3347	0.1131	0.049499	3.1967	10.74	0.7863	3.5997
FG-NET / R50@224	70	PDF_r	4.3691	0.1353	0.047693	3.2789	13.08	0.8293	3.4467
FG-NET / R50@224	70	UR	4.5612	0.1188	0.045353	2.9861	12.45	0.7823	3.7871
FG-NET / R50@224	70	CORN	5.1467	0.1311	0.057328	4.1245	10.96	0.7263	4.0082
FG-NET / R50@224	70	CORAL	5.6271	0.1368	0.069713	4.4371	8.63	0.7238	5.1081
FG-NET / R50@224	70	CE	6.1782	0.1671	0.062909	4.3070	10.99	0.6068	5.0636
FG-NET / ViT@128	70	GFC	4.5266	0.0855	0.051127	3.2526	12.88	0.8324	3.8748
FG-NET / ViT@128	70	DLDL	4.5474	0.0946	0.052658	3.3980	10.32	0.7725	3.9268
FG-NET / ViT@128	70	PDF_r	4.5722	0.0928	0.049914	3.2115	12.62	0.8269	3.9733
FG-NET / ViT@128	70	RNC+L1	4.7038	0.1062	—	—	11.52	0.7766	4.3120
FG-NET / ViT@128	70	RNC+CE	4.7101	0.0778	0.050953	3.5397	12.73	0.7562	4.0682
FG-NET / ViT@128	70	PU	4.7246	0.0940	0.050797	3.2387	12.49	0.7964	4.0825

Cell	K	Method	test MAE	± se	RPS	NLL	test Acc%	test QWK	val MAE
FG-NET / ViT@128	70	UR	5.0725	0.1324	0.051454	3.2575	10.91	0.7435	4.1900
FG-NET / ViT@128	70	CORN	5.8873	0.1315	0.063665	4.1993	10.17	0.6399	4.9545
FG-NET / ViT@128	70	CE	7.1701	0.1330	0.069878	4.9877	9.18	0.5370	5.9472
FG-NET / ViT@128	70	CORAL	9.1862	0.5038	0.109559	5.6862	6.18	0.5428	8.4185
FG-NET / ViT@224	70	DLDL	3.9241	0.0892	0.044561	3.1856	10.89	0.8243	3.3926
FG-NET / ViT@224	70	PDF_r	3.9695	0.0760	0.043075	3.0423	13.13	0.8783	3.4608
FG-NET / ViT@224	70	GFC	3.9932	0.0726	0.044483	3.0750	13.18	0.8776	3.4478
FG-NET / ViT@224	70	RNC+L1	4.0379	0.0757	—	—	12.29	0.8315	3.6766
FG-NET / ViT@224	70	RNC+CE	4.0493	0.0916	0.043372	3.3469	13.26	0.8134	3.5601
FG-NET / ViT@224	70	UR	4.2827	0.1045	0.043620	3.0502	12.78	0.8031	3.7872
FG-NET / ViT@224	70	PU	4.2972	0.0896	0.045614	3.0471	12.94	0.8392	3.6834
FG-NET / ViT@224	70	CORN	5.4039	0.1206	0.059247	4.1148	11.06	0.6688	4.5530
FG-NET / ViT@224	70	CE	6.9943	0.1601	0.067152	4.7036	10.11	0.5563	5.7384
FG-NET / ViT@224	70	CORAL	8.9444	0.6660	0.108679	5.9304	6.54	0.5464	8.7004

Table C.1 reports the per-cell test metrics behind Table 1: 23 (dataset, backbone, resolution) cells—BACH (K=4); HCI, APTOS and FG-NET-5 (K=5); MORPH(K=62); FG-NET (K=70)—with ten heads per cell, configurations selected on validation only.

Per-cell anatomy of the failure regimes summarized in the main text. *(i)* PU is the strongest large-K baseline, but its collapse at small K is uniform: its MAE is 31–97% higher than the per-cell best across all small-K cells. DLDL exhibits a bimodal pattern: among the small-K cells, its gap has a median of only 1% but reaches as high as 233%, whereas it remains competitive at large K (within 11% of the per-cell best in every cell). *(ii)* UR is the strongest small-K baseline and remains competitive at large K. *(iii)* Linear+CORAL exhibits its sharpest large-K degradation on ViT at 224 px, attaining an MAE of 8.94, compared with 3.92 for the best-performing head.

C.3 Frozen-backbone retraining and Asymmetric head-swap transfer (Experiment 2)

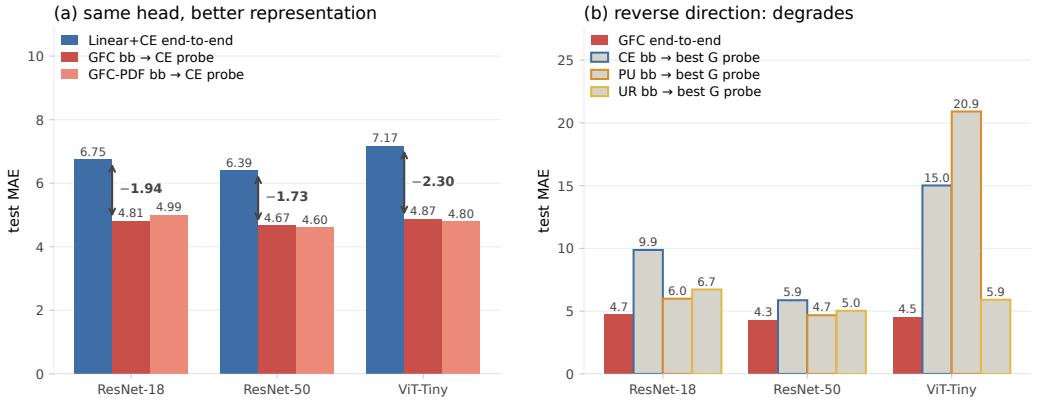


Figure 5: Backbone-resolved head-swap results for Experiment 2. All entries report test MAE on FG-NET (lower is better), 25 (fold, seed) runs per bar. (a) Forward direction: a Linear+CE probe retrained on frozen backbones trained with GFC or GFC-PDF beats Linear+CE trained end-to-end on all three backbones; arrows mark the improvement. (b) Reverse direction: Gaussian probes on frozen backbones trained with Linear (CE/PU/UR) degrade relative to GFC end-to-end.

Backbone-level breakdown. Appendix Figure 5 shows that the transfer asymmetry is consistent across architectures. In the forward direction, replacing nothing but the training head—the evaluated head is Linear+CE in both arms—improves test MAE on every backbone: 6.75 → 4.81 (−1.94) on ResNet-18, 6.39 → 4.67 (−1.72) on ResNet-50, and 7.17 → 4.87 (−2.30) on ViT-Tiny. In the reverse direction, Gaussian probes on frozen Linear-trained backbones degrade on every backbone (mildly on ResNet-50, catastrophically on ViT-Tiny), so the effect cannot be explained by probes being easier to optimize than end-to-end training. These backbone-resolved results reinforce the representation-shaping interpretation of E2 with a consistent effect size across all three architectures.

C.4 Gradient locality studies (Experiment 4)

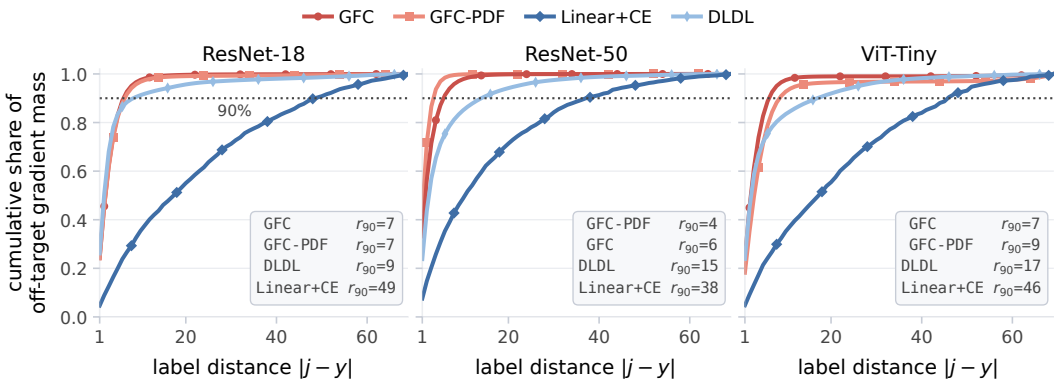


Figure 6: Cross-backbone gradient locality on FG-NET. Cumulative share of off-target gradient mass vs. label distance (on-target term removed), with the 90% radius r_{90} tabulated per panel. The ordering is consistent on all three backbones: Linear+CE $r_{90} = 38$–49, GaussianFC family 4–9, DLDL in between (9–17). Heads with output dimension $\neq K$ (PU, UR, CORAL) expose no per-label gradient field and are structurally absent from this diagnostic.

Figure 6 shows that the main-text locality result is not an artifact of one backbone. Linear+CE needs $r_{90} = 49/38/46$ (ResNet-18 / ResNet-50 / ViT-Tiny) to cover 90% of its off-target mass, while the GaussianFC family needs only 4–9 on every backbone. DLDL provides an informative middle point ($r_{90} = 9$–17): its ordinal prior enters through soft label targets, which concentrates gradients relative to CE but less than head-side structure. The same mechanism note as in the main text applies here: the residual-only and W-weighted views nearly coincide, so the locality is routed through the concentration of the predictive distribution rather than through the W envelope acting as a direct multiplicative gate.

C.5 Ordered feature geometry studies (Experiment 5)

Figure 7 provides the full backbone-wise version of the ordered feature geometry analysis. Three observations hold across backbones. First, class-center distance grows with ordinal gap on every backbone, and more regularly for ordinal-aware methods than for Linear+CE; the growth is monotone except at the extreme gaps, where few class pairs exist (on ViT-Tiny the Gaussian-family curves peak near $\Delta \approx 50$–55 and dip beyond). Second, the probe decomposition of the main text replicates: PU and RNC are the highest methods already at $k=1$ on all three backbones and stays nearly flat thereafter (its own maximum is at $k=3$ and its curve varies by less than 0.04 over the whole range of k), while the GaussianFC family starts lower and climbs; on ViT-Tiny the family's $k=1$ values are notably lower (0.62–0.71) and only catch up by $k \approx 3$, i.e., on ViT the ordinal information of Gaussian heads is the least concentrated in the first principal component. Third, the scale-free curvature of every ordinal-aware head most ratios fall in 0.85–0.98 (a consistent but modest 2–15% effect), with CORAL as low as 0.63 on ResNet-18 and DLDL essentially at parity (0.995), and we deliberately do not use the unnormalized $\mathcal{K}_{\text{feat}}$, which is confounded by the larger feature norms of Gaussian heads (§6.2).

Backbone capacity separates the three loci. The curvature ordering reverses as the backbone grows (Figure 7, bottom row). Across ResNet-18 / ResNet-50 / ViT-Tiny, the loss-side and posterior-side priors degrade monotonically toward Linear+CE—PU $0.850 \to 0.906 \to 0.934$, UR $0.952 \to 0.957 \to 0.971$, CORAL $0.634 \to 0.916 \to 0.933$—while the two architecture-side heads improve monotonically: GFC $0.956 \to 0.904 \to 0.903$, GFC-PDF $0.934 \to 0.925 \to 0.896$. On ResNet-18 CORAL induces the smoothest geometry and the Gaussian family the roughest of the ordinal-aware heads; on ViT-Tiny that order is inverted. This is what the three-loci view predicts. A prior injected in the loss or in the output parameterization constrains the representation only indirectly, through whatever the backbone happens to propagate back; as capacity grows there are more ways to satisfy the same output constraint, so that indirect grip loosens. Head-side structure instead reshapes the gradient field along the label axis directly, and does not weaken with capacity. Two qualifications: the trend is carried by the curvature row and by the rank changes in the probe row—the distance-growth ratios of the top row decline for every method, ours included—and CORN is a partial exception.

C.6 Weight-space verification (Experiment 6)

This section provides the weight-space verification of M-A2 (head locality, Mechanism A) and M-B1 (smoothness, Mechanism B) on FG-NET at 128 px; every value pools 25 runs per method at its validation-selected configuration. Locality is reported on ResNet-18; smoothness is reported on all three backbones.

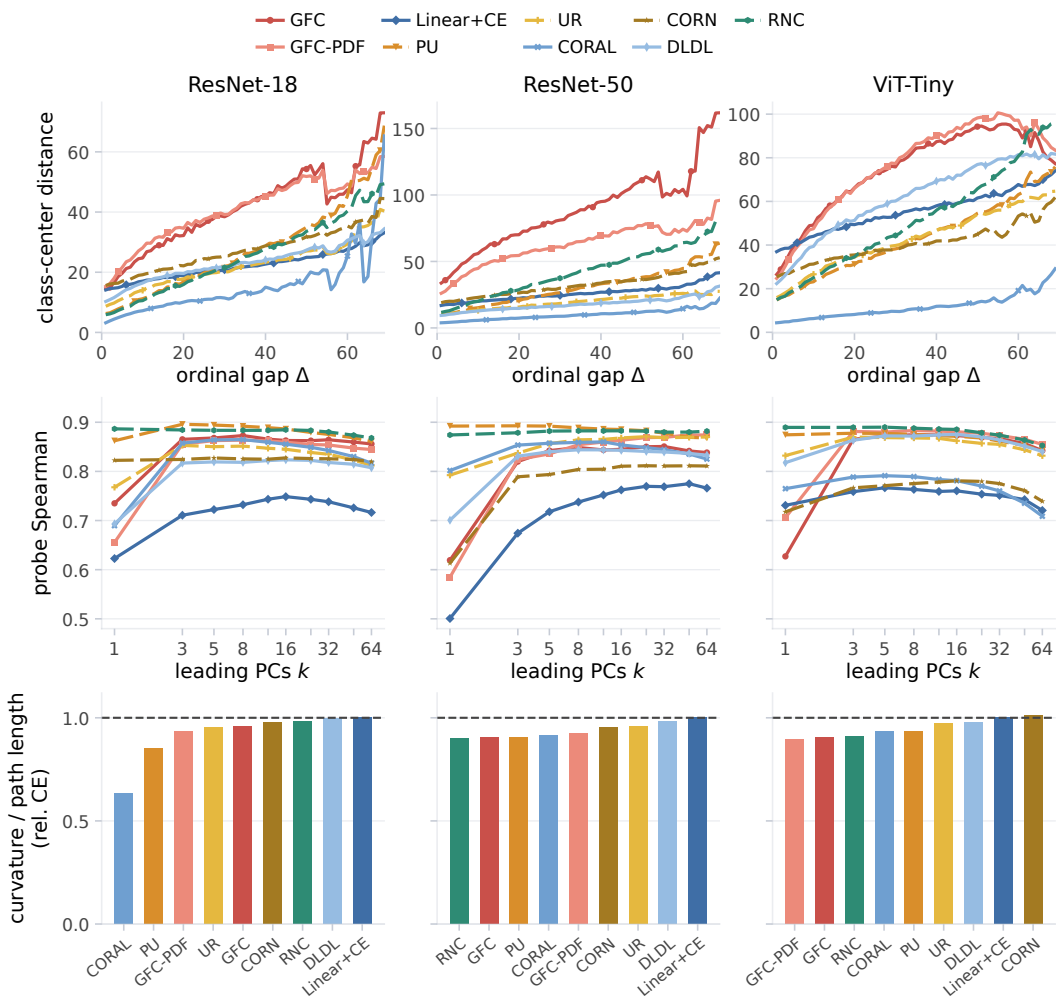


Figure 7: Backbone-wise ordered-geometry diagnostics on FG-NET. Rows: (1) mean class-center distance vs. ordinal gap; (2) ordinal probe Spearman vs. number of leading principal components; (3) scale-free curvature (curvature normalized by path length, relative to Linear+CE). Columns: ResNet-18 / ResNet-50 / ViT-Tiny. PU is the highest method already at k=1 on all three backbones and its curve then varies by less than 0.04 over the whole range of k; the GaussianFC family starts much lower and climbs. All ordinal-aware heads have scale-free curvature below Linear+CE on all backbones.

Locality in weight space (M-A2). For each row of the trained effective weight matrix we compute the Spearman correlation ρ between $|W_{ij}|$ and the negative distance $-|j - \mathrm{argmax}_k |W_{ik}||$ to the row's peak; $\rho = 1$ means the row decays monotonically on both sides of its peak. Figure 8(a): the GaussianFC family attains $\rho = 1.000$, UR 0.149, DLDL 0.106, CORN 0.059, and Linear+CE 0.058. The family's ρ=1.000 is enforced by the parameterization; the informative number is the baselines', which stay at $\rho \leq 0.149$ after full training—monotone decay away from a per-row center in exactly the form Definition 1 intends is not something an unconstrained head acquires on its own. We report ρ rather than the fixed-threshold support width $|\mathcal{S}_i(\alpha)|$: once the trained envelope is wider than the label axis, every position exceeds $\alpha \cdot \max_k |W_{ik}|$ and that statistic saturates at K, yielding a spurious near-tie at the winning working point (57.5 for GFC vs. 61.0 for Linear+CE at α=0.1). ρ measures the profile *shape* and does not saturate. Three honesty notes. First, ρ is scale-blind: the diamonds in panel (a) show that the per-row dynamic range spans more than an order of magnitude *within* the family (50 for GFC vs. 1,391 for GFC-PDF on ResNet-18) and does not separate family from baselines (137 for Linear+CE, 153 for UR, 156 for DLDL); monotone decay and concentration of the envelope are logically independent properties, and ρ certifies only the first. Second, ρ is a structural descriptor, not a performance predictor: DLDL and UR score low ρ yet are strong, because their prior lives in soft labels and in the loss, where ρ cannot see it. Third, ρ is undefined for heads whose output dimension differs from K (PU, CORAL).

Envelope width. The scale that ρ deliberately ignores is exactly what Definition 1 measures. Computing $\sigma_i(W)$ per row and reporting its median over rows (then averaging over the 25 runs of each method's validation-selected configuration) gives, on FG-NET/ResNet-18 with K=70: GFC-PDF 0.274 ± 0.008, GFC $0.357 \pm$

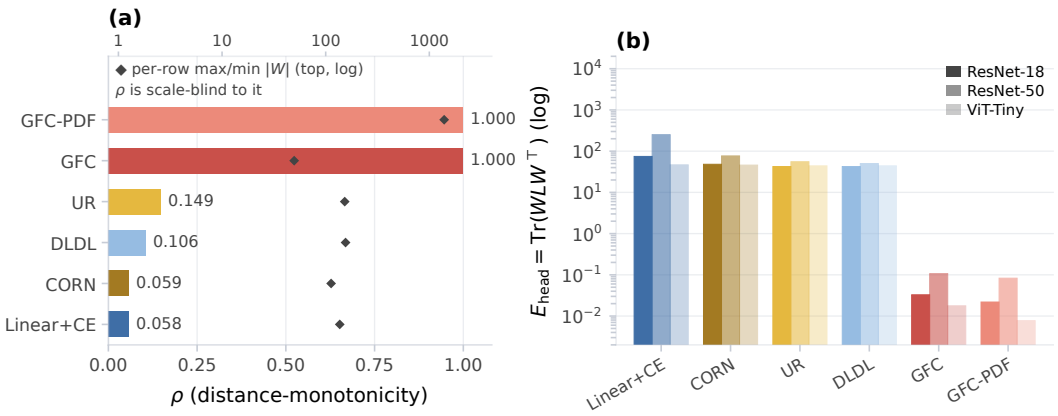


Figure 8: Weight-space evidence for locality and smoothness. (a) Distance-monotonicity ρ on ResNet-18: per-row rank correlation between $|W_{ij}|$ and distance to the row's peak, averaged over rows and runs. GaussianFC-family heads attain exactly 1.000; linear-head baselines stay below 0.15. Diamonds on the upper log axis give the per-row dynamic range $\max_j |W_{ij}| / \min_j |W_{ij}|$, to which ρ is blind by construction. (b) Label-axis Dirichlet energy $\mathcal{E}_{\text{head}}(W) = \text{Tr}(WLW^\top)$ on a log scale, for all three backbones. PU (a single output) and CORAL (one shared weight vector with per-threshold biases) have no label-axis weight matrix and appear in neither panel; RNC is absent for the same reason—its encoder has no head, and its ℓ_1 probe is one-dimensional. CORN's $K-1$ columns are the cumulative thresholds $y > k$ rather than the classes themselves; like UR's, they still lie on the label axis, so it appears in both panels.

0.025, UR 0.885 ± 0.137 and Linear+CE 1.119 ± 0.200, all as fractions σ_i/K. The family therefore sits near $K/3$ and the linear heads at or beyond K—a clean separation, but one worth stating on the scale itself rather than as a threshold: at $\sigma_i/K = 0.27$–0.36 the envelope still leaves 7–16% of its mass outside half the label axis, so the family is *local*, not *concentrated*, at K=70. This is the quantitative content behind the architectural factor of Prop. 1, and it is measured on the trained W rather than assumed.

Label-axis Dirichlet energy (M-B1). Figure 8(b) reports $\mathcal{E}_{\text{head}}(W)$ per method on a log scale. On ResNet-18 the Gaussian-family heads sit at 0.022–0.033, versus 74.6 for Linear+CE and $\approx 42 - 48$ for UR, DLDL and CORN—a $\approx$2,300–3,400$\times$ gap to Linear+CE. The separation replicates on both other backbones: across ResNet-18, ResNet-50 and ViT-Tiny the family's energy is $(1.7\text{–}4.4) \times 10^{-4}$ of the same-backbone Linear+CE, i.e. a factor of $\approx$2,300–6,000. This is the first verification of the smoothness mechanism against a linear control group: normalizing within the Gaussian family alone hides the three orders of magnitude that the Gaussian basis itself contributes (the low-pass filter of Prop. 2). Within the family at validation-selected configurations, GFC-PDF has the lower energy on every backbone (0.67$\times$, 0.78$\times$ and 0.44$\times$ GFC on ResNet-18, ResNet-50 and ViT-Tiny); because each variant selects its own width σ on validation, these within-family numbers conflate σ with the head parameterization and are not a controlled test of Prop. 2's ceteris-paribus claim—the controlled contrast here is family vs. linear. Beyond the scalar energy, the same weights show where that energy sits: expanding W in the eigenbasis of the path-graph Laplacian, 90% of the label-axis Dirichlet spectrum lies in the 3 lowest of 70 modes for both GFC and GFC-PDF, against 63 of 70 for Linear+CE (a flat spectrum would need 63). The lowest 10% of modes carry 100.0% of the family's spectral energy and 16.5% of Linear+CE's. This is the mode-resolved form of Prop. 2 and the same quantity that bounds the class-center dynamics in Prop. 3. The two panels together also quantify an asymmetry between the two mechanisms as realized in weight space: the smoothness separation is three orders of magnitude, while the locality separation is qualitative and bimodal ($\rho \leq 0.15$ vs. 1.000, with no method in between).

Joint verification. Together, the two views establish that both mechanisms are realized at the level of trained weights, not merely architecturally specified: the monotone row profiles of (a) show that the envelope of Definition 1 is preserved through training, and the Dirichlet energies of (b) show that the Gaussian basis operationally suppresses high-frequency label-axis variation as predicted by Prop. 2. Combined with the gradient-side verification of E4 and the geometry-side verification of E5, this closes the head $\rightarrow$ gradient $\rightarrow$ geometry chain at the weight level.

C.7 Robustness studies (Experiment 7)

We evaluate two practical stressors at the working point selected for the main comparison; no hyperparameter is re-tuned per condition. *Subsampling*: we keep a fraction $\rho \in \{0.1, 0.25, 0.5, 0.75\}$ of the training *subjects* (6 of 56 subjects, i.e. 59–85 images, at ρ=0.1), leaving validation and test untouched. Because patience is counted in epochs, a smaller training set would otherwise be cut off after proportionally fewer optimisation

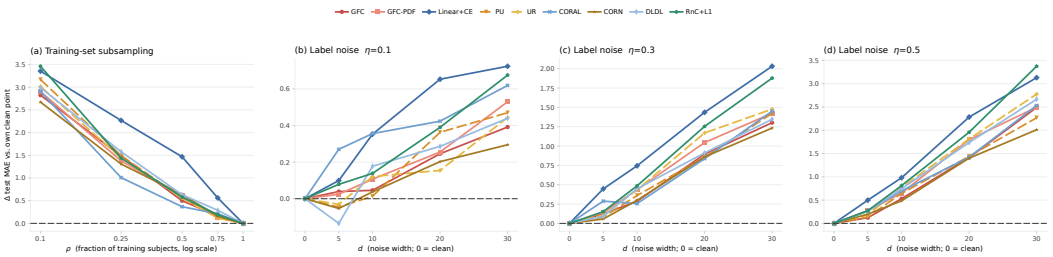


Figure 9: Degradation under training-set subsampling and neighbourhood label noise (FG-NET, ResNet-18 at 128 px; 25 (fold, seed) runs per point; patience 20). Each curve is Δ test MAE relative to that method's *own* clean full-data point, so a flatter curve is a more robust method. (a) subsampling ratio ρ (fraction of training *subjects*; epochs and patience are both scaled by $1/\rho$ so that the early-stopping window is a constant number of optimisation steps). (b–d) neighbourhood label noise: a fraction η of training labels is moved to a uniformly chosen class within d positions of the truth (d=0 is the clean point); validation and test labels stay clean. At the widest settings the window is truncated by the label range and, because FG-NET is skewed young, 70% of the perturbed labels move upward, so d=30 carries a systematic upward shift in addition to noise.

steps—a bias that grows as the training set shrinks and that falls hardest on late-peaking methods—so we scale both the epoch budget and the patience by $1/\rho$. *Neighbourhood label noise*: at strength η and width d, exactly $\mathrm{round}(\eta N)$ training labels are moved to a class drawn uniformly from those within d positions of the true label, with the window truncated at the ends of the label range; d=1 recovers the adjacent-class model used in earlier work. We sweep the full grid $\eta \in \{0.1, 0.3, 0.5\} \times d \in \{5, 10, 20, 30\}$ and summarise each cell by its *dose*, the mean absolute label error over the training set $\eta \cdot \mathbb{E}|\Delta y|$, which ranges from 0.30 to 6.99 years against a clean test MAE of about 4.5 years. Restricting the perturbation to a neighbourhood keeps the noise ordinally plausible—the dominant annotation error mode in real ordinal tasks (age estimation, severity grading, decade dating)—while d controls its magnitude. Width matters: with d=1 the dose is at most 0.7 years even at η=0.7, and no method moves measurably.