LEARNING TO SPACE AND SCORE ORDINAL LABELS ACROSS VISION AND TOUCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Ordinal labels specify class order, but not the distances between classes. Yet ordinal soft supervision must choose both how target probability is distributed across neighboring classes and how the resulting ordered distributions are scored. We make these two choices explicit. First, we formulate cumulative ordinal training as proper scoring over nested threshold events, connecting the commonly used squared cumulative objective (EMD^2/RPS-type scoring) to cumulative Brier scoring and CDF-CE to cumulative logarithmic scoring. We prove cumulative strict propriety, derive their distinct cross-boundary gradient scaling, and show that soft and hard cumulative targets induce different probability-matching geometries. Second, we replace fixed equal spacing with a lightweight learnable supervision geometry in which positive gaps define strictly ordered, fixed-span class positions z, optionally with a single global softness parameter τ. Across five benchmarks spanning vision and touch, Gaussian soft targets improve the mean task-primary metric over hard targets for every dataset under CE, EMD^2, and CDF-CE. Under Gaussian supervision, CDF-CE improves over EMD^2 on all five tasks, whereas the ordering reverses on four of five under hard targets. Learning the target geometry further improves AADB, HCI, and PHAC-2, while the physically grounded equal spacing of Whisker remains preferable. Across the controlled target-and-scoring variants, a fixed or adaptive CDF-CE model achieves the best mean primary result on every benchmark. Together, these results show that ordinal soft supervision depends jointly on target construction and scoring.

1 INTRODUCTION

Ordinal prediction lies between nominal classification and continuous regression. Labels provide an ordering, but that ordering alone does not define a metric: knowing which categories are higher or lower does not imply that adjacent categories are equally separated. Some applications inherit an external metric, while in others the relative class spacing is unknown. Yet ordinal soft-label methods often introduce additional geometric structure through distance-based target similarities or probability assigned to neighboring classes (Diaz & Marathe, 2019; Geng et al., 2013). In other words, *order is observed, while geometry is chosen*.

Geometry is only one part of ordinal supervision. For ordered categories, cumulative probabilities provide a natural representation because moving probability mass across more class boundaries changes more cumulative terms (Epstein, 1969). A categorical prediction p and target q induce cumulative probabilities P_k and Q_k over the nested events $Y \leq k$, but this representation does not determine how their discrepancies should be scored. The cumulative squared objective used in deep ordinal learning (Hou et al., 2016) penalizes $(P_k - Q_k)^2$; we instead examine the logarithmic counterpart obtained by applying Bernoulli log loss to the same events. This isolates the scoring rule while keeping the target and cumulative representation fixed. Thus, target construction, cumulative representation, and scoring are separable design choices.

Figure 1 illustrates why target geometry should not be assumed. Our benchmarks span aesthetic scores, semantic age groups, historical decades, human haptic ratings, and tactile depth bins, whose underlying scales differ substantially. Some are subjective or irregularly grouped, whereas others inherit an external numerical coordinate. In particular, Whisker discretizes physical depth uniformly

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

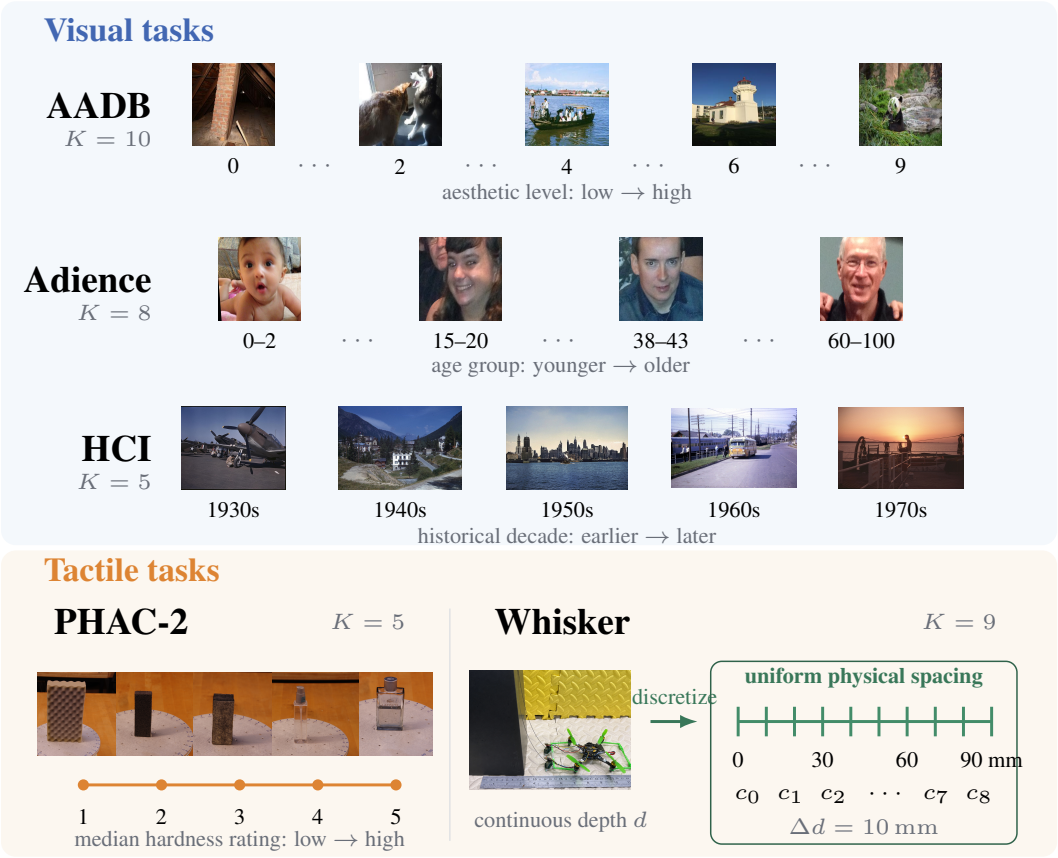

Figure 1: **Evaluation domains and supervision geometry.** Representative inputs are drawn from AADB (Kong et al., 2016), Adience (Eidinger et al., 2014), HCI (Palermo et al., 2012), PHAC-2 (Chu et al., 2015), and Whisker (Ye et al., 2026), and are arranged according to increasing ordinal labels. AADB and Adience contain more classes than shown, whereas all five classes are displayed for HCI and PHAC-2. For AADB, Adience, and PHAC-2, the observed ordering does not by itself specify equal perceptual distances between adjacent classes. HCI has an equal decade coordinate, although equal temporal spacing need not imply equal visual or perceptual change. In contrast, the Whisker benchmark is explicitly constructed from uniform 10-mm physical depth intervals, giving equal spacing a direct task-level metric interpretation.

into 10-mm intervals, giving equal spacing a direct task-level interpretation. The usefulness of geometry adaptation should therefore depend on how the ordinal labels are constructed.

We study two questions under controlled pipelines: *how should a fixed ordinal target be scored?* and *what geometry should construct that target?* For scoring, we hold the target, architecture, and training protocol fixed and compare categorical cross-entropy, cumulative squared scoring, and CDF-CE. For geometry, we hold CDF-CE fixed and replace equal spacing with ordered learnable positions z, optionally with a global softness parameter τ. A direct monotone-CDF head is analyzed separately as a change in output representation.

Our contributions are:

- **A cumulative proper-scoring view of ordinal learning.** We show that squared cumulative training and CDF-CE are cumulative quadratic and logarithmic proper scores over the same nested ordinal events. We establish cumulative strict propriety and characterize their distinct gradient scaling and the interior-versus-boundary geometry induced by soft and hard cumulative targets.
- **Learnable supervision geometry.** We learn positive ordinal gaps whose cumulative sums define strictly ordered, fixed-span class coordinates. The construction preserves order and unimodal-

ity, recovers equal spacing at initialization, and optionally learns a single global target-softness parameter.

- **Evidence across vision and touch.** Across three visual and two tactile tasks, Gaussian soft targets improve the mean primary metric under all three scoring rules. Under Gaussian supervision, CDF-CE improves over cumulative squared scoring on all five benchmarks, while adaptive geometry further improves three. Across the controlled target-and-scoring variants, the fixed/adaptive CDF-CE family contains the best mean primary result on every benchmark.

2 Related Work

Ordinal structure can enter through target construction, the scoring objective, or the prediction parameterization. We separate these components so that target construction and scoring can be varied independently under a fixed prediction architecture.

Ordinal output parameterization. Niu et al. (Niu et al., 2016) decompose a K-class ordinal problem into $K-1$ binary threshold tasks. CORAL enforces rank consistency through shared weights (Cao et al., 2020), while CORN constructs rank probabilities from conditional threshold probabilities (Shi et al., 2023). Other approaches impose unimodality directly on the categorical output, for example using Poisson or binomial distributions (Beckham & Pal, 2017). These methods modify the prediction representation. Our main comparisons instead retain the same K-class softmax head and vary only target construction and scoring; a direct monotone-CDF head is included separately as a representation analysis.

Soft ordinal supervision and adaptive label distributions. Label-distribution learning replaces single-label supervision with a distribution over labels, allowing neighboring classes to contribute to training (Geng et al., 2013). SORD constructs soft ordinal targets from a user-defined inter-class metric (Diaz & Marathe, 2019). TactONet replaces one-hot labels with unimodal Poisson or binomial target distributions to exploit ordinal relations in tactile hardness classification (Fang et al., 2022). Related methods adapt label distributions in other ways: SALDL learns age-dependent Gaussian widths (Hou et al., 2017), unimodal-concentrated learning allows sample-dependent predicted distributions under shape constraints (Li et al., 2022), and OLDL incorporates spatial, semantic, and temporal ordinal relations into distribution-level objectives (Wen et al., 2023). Our adaptation instead learns a single shared set of ordered class coordinates, $z_0 < \cdots < z_{K-1}$, used to construct supervision targets across samples. These coordinates represent supervision geometry rather than a uniquely recovered perceptual metric.

Cumulative scoring. Hou et al. derive a squared cumulative loss from the one-dimensional EMD for ordered classes (Hou et al., 2016). In probabilistic forecasting, the corresponding cumulative quadratic score is the Ranked Probability Score (RPS) (Epstein, 1969); for soft targets, squared CDF discrepancy differs from expected RPS only by a target-dependent constant. Its continuous analogue, CRPS, integrates binary Brier scores over thresholds (Matheson & Winkler, 1976; Gneiting & Raftery, 2007). More generally, proper scoring rules provide a unified framework for probabilistic prediction (Gneiting & Raftery, 2007; Waghmare & Ziegel, 2025). We use this view to separate the cumulative transform from the binary score applied at each threshold, and study scoring jointly with target geometry.

3 Method

3.1 Problem Setup and Target Construction

Let $\mathcal{Y} = \{0, \ldots, K-1\}$ be ordered. A network with logits $a \in \mathbb{R}^K$ produces

$$p_j = \frac{\exp(a_j)}{\sum_i \exp(a_i)}. \tag{1}$$

Hard CE uses $q = \delta_y$. A fixed Gaussian ordinal target places mass according to index distance,

$$q_j^{(y)}(\tau_0) = \frac{\exp[-(j-y)^2/(2\tau_0^2)]}{\sum_i \exp[-(i-y)^2/(2\tau_0^2)]}. \tag{2}$$

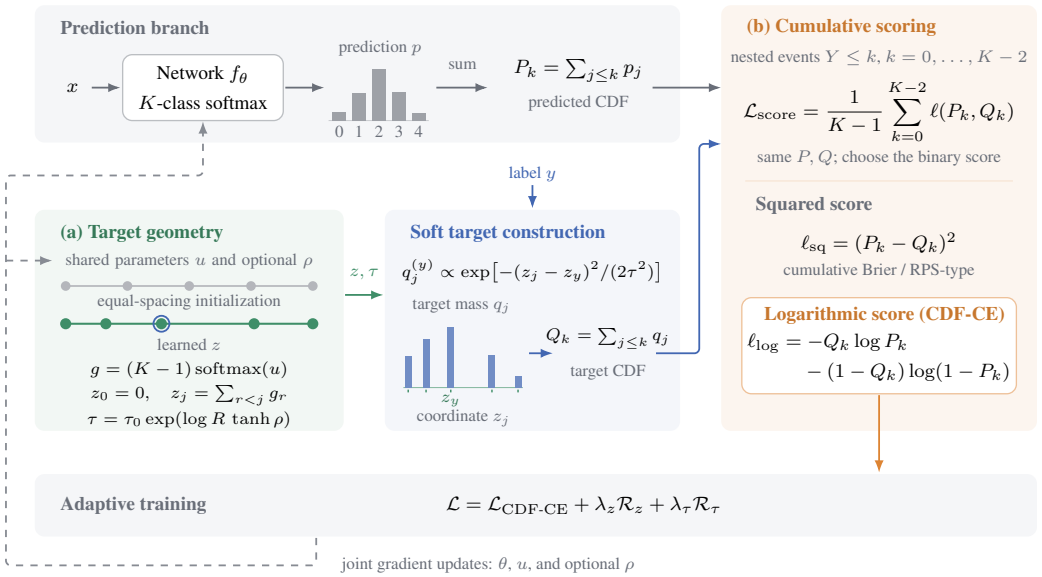

Figure 2: **Method overview.** Ordinal soft supervision separates target geometry from cumulative scoring. (a) Positive normalized gaps are accumulated into strictly ordered class coordinates z, which together with label y and global softness τ define the soft target q. Target bars represent probability masses at the learned coordinates, with the maximum at the labeled coordinate z_y. (b) A standard softmax network predicts p; prediction and target distributions are accumulated into P and Q, then compared using squared or logarithmic scores over the same ordinal thresholds. For the adaptive-geometry variants, CDF-CE is held fixed while the regularized objective jointly learns the network, class geometry, and optionally the global softness through backpropagation.

This familiar construction already contains two assumptions: classes are embedded on a line and adjacent indices are equally spaced. Our adaptive formulation keeps the first assumption—ordinal order—while relaxing the second.

3.2 Cumulative Proper Scoring

An ordinal K-class distribution can be viewed through its $K-1$ cumulative thresholds. For each threshold k,

$$P_k = \sum_{j \le k} p_j, \qquad Q_k = \sum_{j \le k} q_j, \qquad k = 0, \ldots, K-2, \tag{3}$$

where P_k is the predicted probability of the binary event $Y \le k$, and Q_k is the corresponding target probability. Thus, rather than comparing the two categorical distributions class by class, we compare them through the same ordered set of nested binary events.

We consider two cumulative scores. The cumulative squared objective is

$$\mathcal{L}_{\text{EMD}^2} = \frac{1}{K-1} \sum_{k=0}^{K-2} (P_k - Q_k)^2, \tag{4}$$

while cumulative logarithmic scoring gives

$$\mathcal{L}_{\text{CDF-CE}} = -\frac{1}{K-1} \sum_{k=0}^{K-2} \left[Q_k \log P_k + (1 - Q_k) \log(1 - P_k) \right]. \tag{5}$$

Both score exactly the same ordered events $Y \le k$; they differ only in the binary scoring rule used at each threshold.

Theorem 1 (Cumulative strict propriety). *Let $\ell(r, b)$ be a strictly proper binary loss. Then*

$$S_\ell(p, y) = \frac{1}{K-1} \sum_{k=0}^{K-2} \ell(P_k, \mathbb{I}[y \le k]) \tag{6}$$

is strictly proper over categorical distributions. In particular, if $Y \sim q$, then the expected score is uniquely minimized by $p = q$.

For each threshold, strict propriety forces $P_k = Q_k$ in expectation. Since the $K - 1$ nontrivial cumulative values uniquely recover the PMF through

$$p_0 = P_0, \qquad p_j = P_j - P_{j-1}, \qquad p_{K-1} = 1 - P_{K-2},$$

matching all thresholds uniquely implies $p = q$. The complete proof is given in Section A.

For binary Brier loss,

$$\mathbb{E}[(P - B)^2] = (P - Q)^2 + Q(1 - Q),$$

so Equation (4) is the cumulative Brier score up to a target-dependent constant. Bernoulli log loss yields Equation (5) exactly.

Why the two cumulative scores optimize differently. For a fixed threshold k, summing the gradients over all logits on the left side of the boundary gives

$$G_k^{\log} := \sum_{j \le k} \frac{\partial \ell_{\text{CDF-CE},k}}{\partial a_j} = P_k - Q_k, \tag{7}$$

whereas cumulative squared scoring gives

$$G_k^{\text{sq}} := \sum_{j \le k} \frac{\partial \ell_{\text{EMD}^2,k}}{\partial a_j} = 2(P_k - Q_k)P_k(1 - P_k). \tag{8}$$

The two corrections are therefore related by

$$G_k^{\text{sq}} = 2P_k(1 - P_k)\, G_k^{\log}, \qquad 0 < 2P_k(1 - P_k) \le \frac{1}{2}. \tag{9}$$

The attenuation factor approaches zero as $P_k \to 0$ or $P_k \to 1$, so cumulative squared scoring increasingly suppresses the same cross-boundary residual near saturated cumulative predictions.

Soft and hard targets induce different scoring geometries. For an interior cumulative target $Q \in (0, 1)$,

$$\ell_{\log}(P, Q) - \ell_{\log}(Q, Q) = D_{\text{KL}}(\text{Ber}(Q) \,\|\, \text{Ber}(P)), \tag{10}$$

whereas the excess squared score is $(P - Q)^2$. Around $P = Q$, the logarithmic score has curvature $1/[Q(1 - Q)]$, while the squared score has constant curvature 2. Thus soft cumulative targets define interior probability-matching problems with target-dependent local geometry.

Hard targets instead satisfy $Q_k \in \{0, 1\}$, placing the optimum on the probability boundary. They therefore provide threshold-side supervision without an intermediate cumulative probability to recover. This distinction, together with Equation (9), motivates the target×score comparison in Table 2; complete derivations are given in Section B.

3.3 Adaptive Ordinal Target Geometry

The fixed Gaussian target in Equation (2) assumes equally spaced class coordinates $z_j = j$. We instead parameterize positive gaps

$$g_k = (K - 1)\,\text{softmax}(u)_k, \qquad k = 0, \ldots, K - 2, \tag{11}$$

and define

$$z_0 = 0, \qquad z_j = \sum_{r=0}^{j-1} g_r, \qquad j = 1, \ldots, K - 1. \tag{12}$$

The adaptive Gaussian target is

$$q_j^{(y)}(z, \tau) = \frac{\exp\left[-(z_j - z_y)^2/(2\tau^2)\right]}{\sum_i \exp\left[-(z_i - z_y)^2/(2\tau^2)\right]}. \tag{13}$$

The geometry parameters u are optimized jointly with the network parameters by backpropagating through the target construction. To keep the learned geometry close to the equal-spacing initialization, we use

$$\mathcal{R}_z = \frac{1}{K-1} \sum_{k=0}^{K-2} (g_k - 1)^2. \tag{14}$$

When global softness is learned, we parameterize

$$\tau = \tau_0 \exp(\log R \, \tanh \rho), \tag{15}$$

which constrains $\tau/\tau_0 \in (1/R, R)$, and regularize it with

$$\mathcal{R}_\tau = [\log(\tau/\tau_0)]^2. \tag{16}$$

The adaptive variants minimize

$$\mathcal{L} = \mathcal{L}_{\text{CDF-CE}}(p, q(z, \tau)) + \lambda_z \mathcal{R}_z + \lambda_\tau \mathcal{R}_\tau, \tag{17}$$

with gradients jointly updating the network parameters, u, and optionally ρ.

The positive-gap parameterization guarantees strictly ordered class coordinates with fixed endpoints,

$$0 = z_0 < z_1 < \cdots < z_{K-1} = K - 1,$$

so the adaptive target remains unimodal with a unique maximum at the observed class. Fixing the endpoints also removes the global translation and scale ambiguities of the geometry. Formal statements and proofs are given in Section C.

Representation diagnostic. M0–M5 use the same categorical softmax head. M6 instead predicts a monotone CDF with positive logit increments: $h_0 = r_0$, $h_k = h_0 + \sum_{j=1}^{k} \text{softplus}(r_j)$, and $P_k = \text{sigmoid}(h_k)$. Because this changes the output representation, M6 is not used to claim that additional target adaptation is always better.

4 Experiments

4.1 Datasets, Metrics, and Evaluation Protocols

We evaluate the same supervision design across five benchmarks spanning two sensing modalities: three visual tasks (AADB, Adience, and HCI) and two tactile tasks (PHAC-2 and Whisker). Representative inputs and label structures are shown in Figure 1. AADB measures image aesthetics (Kong et al., 2016) and uses Spearman's ρ as the primary metric, while Adience predicts eight age groups (Eidinger et al., 2014) and uses exact-match accuracy; for both datasets, we follow the task construction and evaluation conventions of Hou et al. (Hou et al., 2016), using a common ResNet-18/AdamW training pipeline across methods. HCI dates historical color images (Palermo et al., 2012) and follows the Ord2Seq setup (Wang et al., 2023), including ten splits, a VGG-16 backbone, Adam optimization, and a 1700-step training schedule; its primary metric is MAE, computed here from the expected class value of the predicted distribution (MAE-EV). PHAC-2 provides BioTac recordings with five-level haptic adjective ratings (Chu et al., 2015; Richardson & Kuchenbecker, 2020); following Richardson and Kuchenbecker (Richardson & Kuchenbecker, 2020), we use object-disjoint evaluation with five folds per adjective, a multimodal CNN, and macroaveraged MAE (MAE^M) as the primary metric. Whisker predicts discretized tactile depth from three-channel whisker pressure measurements (Ye et al., 2026); we use the data split and tactile-depth formulation of Ye et al., a shared MLP backbone across methods, and depth MAE as the primary metric. Together, these tasks cover subjective ratings, grouped semantic categories, temporal coordinates, and uniformly discretized physical measurements, providing varied settings in which the equal-spacing prior may be more or less appropriate. Complementary ordinal or continuous metrics are reported for each task; detailed dataset-specific architectures, optimization settings, and checkpoint rules are provided in Section F.

4.2 Controlled Method Variants

M0 uses hard CE; M1–M3 share the same fixed Gaussian target and respectively use categorical CE, cumulative squared scoring, and CDF-CE. M4 learns the ordered class geometry z, M5 additionally learns global softness τ, and M6 changes the output representation to a direct monotone CDF. The complete method taxonomy and correspondence to prior work are provided in Section E.

Table 1: Main controlled results (mean ± population std). Cell format: AADB Spearman / MAE-EV; Adience Acc. / one-off Acc.; HCI MAE-EV / QWK; PHAC-2 MAE^M / Acc.; Whisker depth MAE (mm) / QWK. Arrows in the header indicate the primary metric direction. Bold marks the best M0–M5 value for each reported metric; M6 is separated because it changes representation.

Method	AADB ↑ / ↓	Adience ↑ / ↑	HCI ↓ / ↑	PHAC ↓ / ↑	Whisker ↓ / ↑
M0 Hard CE	$.648 \pm .014$ $1.311 \pm .081$	$.578 \pm .058$ $.914 \pm .017$	$.715 \pm .040$ $.640 \pm .043$	$.689 \pm .155$ $.627 \pm .188$	$7.928 \pm .572$ $.693 \pm .016$
M1 Gaussian CE	$.675 \pm .006$ $1.228 \pm .013$	$.616 \pm .040$ $.955 \pm .013$	$.685 \pm .040$ $.678 \pm .029$	$.632 \pm .150$ $.639 \pm .175$	$7.143 \pm .508$ $.733 \pm .016$
M2 Gaussian EMD^2	$.674 \pm .006$ $1.232 \pm .020$	$.616 \pm .041$ $.946 \pm .012$	$.718 \pm .065$ $.659 \pm .048$	$.629 \pm .155$ $.612 \pm .192$	$7.064 \pm .494$ $.696 \pm .074$
M3 Gaussian CDF-CE	$.676 \pm .009$ $1.219 \pm .015$	$\mathbf{.620 \pm .036}$ $.953 \pm .011$	$.693 \pm .046$ $.702 \pm .036$	$.618 \pm .148$ $.643 \pm .175$	$\mathbf{6.778 \pm .231}$ $\mathbf{.753 \pm .009}$
M4 learned z	$\mathbf{.678 \pm .007}$ $\mathbf{1.217 \pm .007}$	$.615 \pm .040$ $\mathbf{.955 \pm .014}$	$\mathbf{.666 \pm .030}$ $\mathbf{.708 \pm .027}$	$.605 \pm .160$ $.643 \pm .180$	$7.182 \pm .189$ $.743 \pm .011$
M5 learned (z, τ)	$.671 \pm .011$ $1.231 \pm .014$	$.612 \pm .039$ $.948 \pm .013$	$.689 \pm .048$ $.697 \pm .042$	$\mathbf{.593 \pm .166}$ $\mathbf{.650 \pm .174}$	$7.114 \pm .171$ $.746 \pm .011$
M6 direct CDF	$.682 \pm .009$ $1.203 \pm .016$	$.605 \pm .031$ $.952 \pm .013$	$.690 \pm .057$ $.704 \pm .051$	$.594 \pm .167$ $.622 \pm .193$	$7.293 \pm .312$ $.768 \pm .006$

4.3 Main Results

Table 1 reports two metrics per task in each cell. The first line is the primary metric and the second is a complementary ordinal/continuous metric. This compact format exposes metric trade-offs while the complete metric tables remain in Section H. Across the five benchmarks, the strongest M3–M5 variant achieves the best task-primary mean among M0–M5 on every dataset, despite the tasks spanning distinct modalities and label semantics.

Soft ordinal targets provide a strong and consistent gain. M0→M1 improves the task-primary mean on all five benchmarks. The full factorial shows that this is not specific to categorical CE: for every dataset and for each of CE, EMD^2, and CDF-CE, the Gaussian target improves the aggregate task-primary mean over its hard-target counterpart. This consistency across scoring rules indicates that target construction is itself a major component of ordinal supervision, rather than merely an auxiliary choice attached to a particular loss.

Cumulative log scoring is consistently strongest under Gaussian supervision. M2 and M3 differ only in the proper binary score applied to the same cumulative target events. Across all five datasets, replacing cumulative squared scoring with CDF-CE improves the task-primary mean: AADB Spearman increases from .674 to .676 and Adience accuracy from .616 to .620, while HCI MAE-EV decreases from .718 to .693, PHAC-2 MAE^M from .629 to .618, and Whisker depth MAE from 7.064 to 6.778 mm. Thus, under the same Gaussian ordinal target, cumulative log scoring gives the better mean primary result on every benchmark.

4.4 Target × Scoring Interaction

To separate target construction from scoring, we complete a 2×3 factorial over hard versus Gaussian targets and categorical CE, cumulative squared scoring, and CDF-CE. All comparisons use the same dataset-specific architecture, split, optimization budget, and checkpoint rule.

Table 2 reveals a clear target-dependent change in score ordering. With Gaussian targets, CDF-CE achieves a better mean primary metric than EMD^2 on all five datasets. Under hard targets, this ordering reverses on four of the five benchmarks: EMD^2 gives the better mean on AADB, Adience, PHAC-2, and Whisker, while HCI remains the exception. The advantage of cumulative log scoring therefore emerges specifically in conjunction with soft ordinal supervision, rather than as a target-independent property of the scoring rule. Matched-unit uncertainty analyses are provided in Section G.

4.5 When Should Target Geometry Be Learned?

The fixed-geometry CDF-CE model M3 already gives the best controlled mean on Adience and Whisker, while learning the ordinal geometry further improves AADB, HCI, and PHAC-2. Specifi-

Table 2: Target × scoring factorial on the task-primary metric. Entries are means over matched evaluation units. Bold marks the better cumulative score within each target family; CE is included as the categorical reference.

Target	Score	AADB ρ ↑	Adience Acc.↑	HCI MAE↓	PHAC MAE^M ↓	Whisker mm↓
Hard	CE	.648	.578	.715	.689	7.928
Hard	EMD^2	**.651**	**.597**	.734	**.660**	**7.119**
Hard	CDF-CE	.650	.595	**.727**	.676	7.249
Gaussian	CE	.675	.616	.685	.632	7.143
Gaussian	EMD^2	.674	.616	.718	.629	7.064
Gaussian	CDF-CE	**.676**	**.620**	**.693**	**.618**	**6.778**

cally, M3→M4 improves HCI from .693 to .666 MAE-EV, PHAC-2 from .618 to .605 MAE^M, and AADB from .676 to .678 Spearman. Adding global softness in M5 further improves PHAC-2 to .593 MAE^M. Consequently, among the controlled M0–M5 variants, the M3–M5 family contains the best task-primary mean on all five benchmarks.

Figure 3(a) shows that these performance changes are accompanied by systematic task-dependent changes in the learned ordinal geometry. The largest softness-normalized gap deviations occur on AADB, Adience, and Whisker, reaching approximately 0.26–0.33 in magnitude, whereas HCI and PHAC-2 remain much closer to the equal-spacing reference. The magnitude of geometric deformation is not itself predictive of the performance gain: HCI benefits from comparatively small local corrections, while the substantially larger deformation learned on Whisker moves away from its externally defined uniform physical scale and does not improve depth error. These results suggest that adaptive geometry is most useful when equal spacing is an implicit modeling assumption, whereas an externally grounded physical spacing can already provide an appropriate supervision geometry.

M5 additionally learns one global softness parameter τ. All datasets learn $\tau/\tau_0 < 1$ (approximately .91–.96), while only PHAC-2 further improves the primary metric, from .605 to .593 MAE^M.

Additional output-representation and continuous-regression controls are reported in Section I.

5 Discussion

The experiments reveal a clear interaction between target construction and cumulative scoring. With Gaussian targets, CDF-CE improves over EMD^2 on all five benchmarks, whereas under hard targets the ordering reverses on four of five. This pattern is consistent with the different scoring geometries induced by the two target regimes. Gaussian targets place each nontrivial Q_k in the interior of $(0, 1)$, where logarithmic scoring performs Bernoulli-KL probability matching with target-dependent local curvature. Hard targets instead place the optimum at the probability boundary, so the additional pressure of logarithmic scoring mainly drives predictions toward more confident threshold separation rather than matching an intermediate cumulative probability. Together with the gradient relation in Equation (9), this provides a mechanism consistent with the target-dependent ordering observed in Table 2.

Geometry adaptation provides a complementary source of flexibility. Learning z improves AADB, HCI, and PHAC-2, suggesting that equal spacing can be restrictive when class distances are not externally specified. Whisker shows the opposite pattern: its classes are constructed from uniform 10-mm physical depth intervals, so equal spacing already has a direct task-level metric interpretation. Deforming that geometry therefore moves away from known physical structure and does not improve depth error. These results suggest that adaptive geometry is most useful when class spacing is implicit or uncertain, whereas externally grounded spacing can already provide an appropriate supervision geometry.

The learned z and global softness τ are shared supervision parameters optimized jointly with the predictor. They capture dataset-level target geometry and softness rather than instance-specific uncertainty. The direct-CDF variant changes the output representation and is therefore analyzed separately from the target-geometry comparison.

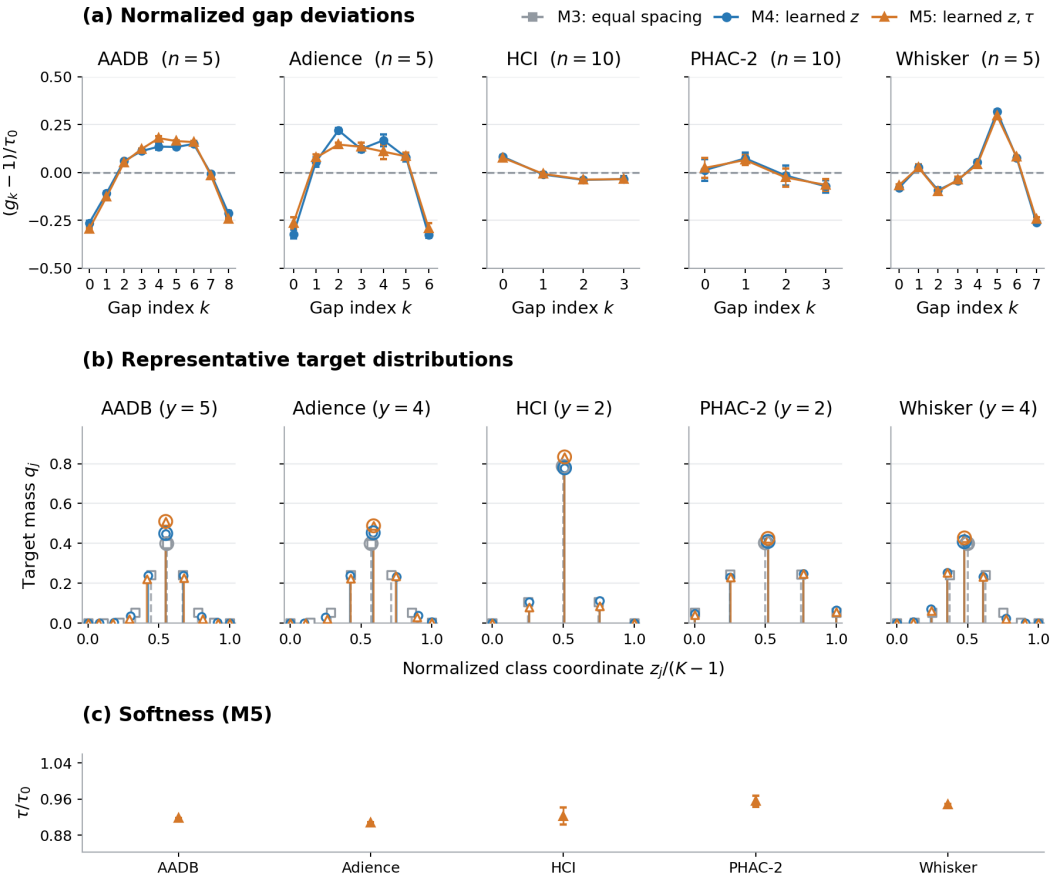

Figure 3: **Learned supervision geometry across five ordinal tasks.** (a) Softness-normalized deviations of adjacent class gaps, $(g_k - 1)/\tau_0$, where zero denotes the equal-spacing geometry used by M3. Points and error bars show mean $\pm$ population standard deviation across matched evaluation units; normalizing by τ_0 expresses each gap deviation in units of the baseline Gaussian target width. (b) Representative target distributions from matched runs for M3, M4, and M5, plotted at their normalized class coordinates $z_j/(K-1)$. (c) Learned global softness for M5, reported as τ/τ_0, with one denoting the fixed-softness reference. For PHAC-2, the five object-disjoint folds are first averaged within each adjective, yielding ten adjective-level units. Across datasets, the learned geometries exhibit systematic task-dependent departures from equal spacing, while the magnitude of the deformation need not track the downstream performance gain.

6 Conclusion

Ordinal soft supervision involves two interacting choices: how probability mass is distributed across ordered classes and how discrepancies between the resulting cumulative distributions are scored. We showed that cumulative squared loss and CDF-CE are quadratic and logarithmic proper scores over the same ordinal threshold events, but differ in both gradient scaling and the probability geometry induced by soft versus hard targets. Across five benchmarks spanning vision and touch, Gaussian targets improve the mean primary metric under all three scoring rules, and under Gaussian supervision CDF-CE improves over EMD^2 on all five tasks.

Learning the target geometry adds a complementary degree of freedom. Adaptive class positions improve tasks whose spacing is not externally prescribed, while Whisker illustrates why meaningful physical geometry can be worth preserving. Across M0–M5, fixed or adaptive CDF-CE provides the best mean primary result on every benchmark. Together, these results establish target geometry and cumulative scoring as two central, interacting components of ordinal soft supervision.

Reproducibility Statement

Complete implementation, preprocessing, hyperparameter, and evaluation details are provided in Section F. Code and machine-readable results will be released with the supplementary material.

AI Use Statement

Generative AI tools were used for literature search, manuscript editing, assistance with mathematical derivations and proofs and experimental code. The authors independently verified all AI-assisted content and take full responsibility for the submission.

References

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

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.

Vivian Chu, Ian McMahon, Lorenzo Riano, Craig G McDonald, Qin He, Jorge Martinez Perez-Tejada, Michael Arrigo, Trevor Darrell, and Katherine J Kuchenbecker. Robotic learning of haptic adjectives through physical interaction. *Robotics and Autonomous Systems*, 63:279–292, 2015.

Raul Diaz and Amit Marathe. Soft labels for ordinal regression. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 4733–4742. IEEE, 2019.

Eran Eidinger, Roee Enbar, and Tal Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on information forensics and security*, 9(12):2170–2179, 2014.

Edward S Epstein. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology (1962-1982)*, 8(6):985–987, 1969.

Senlin Fang, Zhengkun Yi, Tingting Mi, Zhenning Zhou, Chaoxiang Ye, Wanfeng Shang, Tiantian Xu, and Xinyu Wu. Tactonet: Tactile ordinal network based on unimodal probability for object hardness classification. *IEEE Transactions on Automation Science and Engineering*, 20(4):2784–2794, 2022.

Xin Geng, Chao Yin, and Zhi-Hua Zhou. Facial age estimation by learning from label distributions. *IEEE transactions on pattern analysis and machine intelligence*, 35(10):2401–2412, 2013.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Le Hou, Chen-Ping Yu, and Dimitris Samaras. Squared earth mover's distance-based loss for training deep neural networks. *arXiv preprint arXiv:1611.05916*, 2016.

Peng Hou, Xin Geng, Zeng-Wei Huo, and Jia-Qi Lv. Semi-supervised adaptive label distribution learning for facial age estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *European conference on computer vision*, pp. 662–679. Springer, 2016.

Qiang Li, Jingjing Wang, Zhaoliang Yao, Yachun Li, Pengju Yang, Jingwei Yan, Chunmao Wang, and Shiliang Pu. Unimodal-concentrated loss: Fully adaptive label distribution learning for ordinal regression. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20481–20490. IEEE, 2022.

James E Matheson and Robert L Winkler. Scoring rules for continuous probability distributions. *Management science*, 22(10):1087–1096, 1976.

Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4920–4928, 2016.

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

Benjamin A Richardson and Katherine J Kuchenbecker. Learning to predict perceptual distributions of haptic adjectives. *Frontiers in neurorobotics*, 13:116, 2020.

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.

Kartik Waghmare and Johanna Ziegel. Proper scoring rules for estimation and forecast evaluation. *Annual Review of Statistics and Its Application*, 13, 2025.

Jinhong Wang, Yi Cheng, Jintai Chen, TingTing Chen, Danny Chen, and Jian Wu. Ord2seq: Regarding ordinal regression as label sequence prediction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5842–5852. IEEE, 2023.

Changsong Wen, Xin Zhang, Xingxu Yao, and Jufeng Yang. Ordinal label distribution learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23424–23434. IEEE, 2023.

Chaoxiang Ye, Guido de Croon, and Salua Hamaza. Whisker-based tactile flight for tiny drones. *Nature Communications*, 17(1):9743, 2026.

A Complete Proof of Cumulative Strict Propriety

Fix $q \in \Delta^{K-1}$, let $Y \sim q$, and define

$$B_k = \mathbb{I}[Y \leq k].$$

Then

$$\mathbb{E}[S_\ell(p, Y)] = \frac{1}{K-1} \sum_{k=0}^{K-2} \mathbb{E}[\ell(P_k, B_k)] \tag{18}$$

$$= \frac{1}{K-1} \sum_{k=0}^{K-2} \left[Q_k \ell(P_k, 1) + (1 - Q_k)\ell(P_k, 0) \right]. \tag{19}$$

Because ℓ is strictly proper for Bernoulli outcomes, each term is uniquely minimized at

$$P_k = Q_k.$$

Therefore the full expected score is uniquely minimized when all nontrivial cumulative probabilities match.

The mapping from the cumulative vector to the categorical PMF is injective:

$$p_0 = P_0, \qquad p_j = P_j - P_{j-1} \quad (1 \leq j \leq K-2),$$

and

$$p_{K-1} = 1 - P_{K-2}.$$

Hence $P_k = Q_k$ for all k uniquely implies $p = q$, proving strict propriety over categorical distributions.

For Brier loss,

$$\mathbb{E}[(P - B)^2] = Q(P-1)^2 + (1-Q)P^2 \tag{20}$$

$$= (P - Q)^2 + Q(1-Q), \tag{21}$$

where the second term is independent of P. Thus minimizing expected binary Brier loss is equivalent to minimizing $(P-Q)^2$. Summing over ordinal thresholds gives Equation (4) up to a target-dependent constant.

For binary log loss,

$$\mathbb{E}[-B\log P-(1-B)\log(1-P)]=-Q\log P-(1-Q)\log(1-P),$$

which yields Equation (5) exactly.

B Gradient Derivations

The softmax Jacobian is

$$\frac{\partial p_i}{\partial a_j}=p_i(\mathbb{I}[i=j]-p_j).$$

Since

$$P_k=\sum_{i\le k}p_i,$$

we obtain

$$\frac{\partial P_k}{\partial a_j}=\begin{cases}p_j(1-P_k), & j\le k,\\ -p_jP_k, & j>k.\end{cases} \tag{22}$$

For cumulative log scoring,

$$\ell_{\text{CDF-CE},k}=-Q_k\log P_k-(1-Q_k)\log(1-P_k),$$

and therefore

$$\frac{\partial \ell_{\text{CDF-CE},k}}{\partial P_k}=\frac{P_k-Q_k}{P_k(1-P_k)}.$$

Summing over the logits on the left side of threshold k,

$$\sum_{j\le k}\frac{\partial \ell_{\text{CDF-CE},k}}{\partial a_j}=\frac{P_k-Q_k}{P_k(1-P_k)}\sum_{j\le k}p_j(1-P_k) \tag{23}$$

$$=\frac{P_k-Q_k}{P_k(1-P_k)}P_k(1-P_k) \tag{24}$$

$$=P_k-Q_k. \tag{25}$$

For the cumulative squared score,

$$\ell_{\text{EMD}^2,k}=(P_k-Q_k)^2,$$

so

$$\frac{\partial \ell_{\text{EMD}^2,k}}{\partial P_k}=2(P_k-Q_k),$$

and

$$\sum_{j\le k}\frac{\partial \ell_{\text{EMD}^2,k}}{\partial a_j}=2(P_k-Q_k)\sum_{j\le k}p_j(1-P_k) \tag{26}$$

$$=2(P_k-Q_k)P_k(1-P_k). \tag{27}$$

Gradient attenuation. Combining the two expressions gives

$$G_k^{\text{sq}}=2P_k(1-P_k)\,G_k^{\log}.$$

Since $P(1-P)$ is maximized at $P=1/2$,

$$0<2P_k(1-P_k)\le\frac{1}{2}\qquad\text{for } P_k\in(0,1).$$

The factor converges to zero as $P_k\to 0$ or $P_k\to 1$, establishing Equation (9).

Interior score geometry. For $Q \in (0,1)$,

$$\ell_{\log}(P,Q) - \ell_{\log}(Q,Q) = -Q \log \frac{P}{Q} - (1-Q) \log \frac{1-P}{1-Q} \tag{28}$$

$$= D_{\mathrm{KL}}(\mathrm{Ber}(Q) \,\|\, \mathrm{Ber}(P)), \tag{29}$$

which proves Equation (10). The first derivative with respect to P vanishes at $P = Q$, while

$$\left.\frac{\partial^2 \ell_{\log}(P,Q)}{\partial P^2}\right|_{P=Q} = \frac{1}{Q(1-Q)}. \tag{30}$$

A second-order Taylor expansion around $P = Q$ therefore gives

$$\ell_{\log}(P,Q) - \ell_{\log}(Q,Q) = \frac{(P-Q)^2}{2Q(1-Q)} + o\big((P-Q)^2\big), \tag{31}$$

whereas

$$\ell_{\mathrm{sq}}(P,Q) - \ell_{\mathrm{sq}}(Q,Q) = (P-Q)^2 \tag{32}$$

holds exactly.

Boundary case for hard targets. For $Q \in \{0,1\}$, the common optimum lies on the boundary rather than in the interior. For example, if $Q = 1$, then

$$\ell_{\log}(P,1) = -\log P, \qquad \ell_{\mathrm{sq}}(P,1) = (1-P)^2,$$

both uniquely minimized as $P \to 1$. The case $Q = 0$ is symmetric. Hence the interior curvature analysis above applies specifically to soft cumulative targets.

C Geometry Validity and Invariances

Order preservation and unimodality. Because

$$g_k = (K-1)\,\mathrm{softmax}(u)_k > 0, \qquad \sum_{k=0}^{K-2} g_k = K-1,$$

the cumulative construction in Equation (12) gives

$$0 = z_0 < z_1 < \cdots < z_{K-1} = K-1.$$

For the adaptive Gaussian target,

$$q_j^{(y)} \propto \exp\left[-\frac{(z_j - z_y)^2}{2\tau^2}\right],$$

the squared distance from z_y is uniquely minimized at $j = y$. Hence $q_y^{(y)}$ is the unique maximum. Moreover,

$$|z_i - z_y| < |z_j - z_y| \quad \Longrightarrow \quad q_i^{(y)} > q_j^{(y)},$$

so the target probability decreases with distance from the observed class on either side.

Affine invariance. The adaptive target depends on z and τ only through the standardized distances

$$\frac{z_j - z_y}{\tau}.$$

Consequently, for any $c > 0$ and $b \in \mathbb{R}$,

$$(z, \tau) \mapsto (cz + b, c\tau)$$

leaves every target distribution $q^{(y)}$ unchanged. Fixing

$$z_0 = 0, \qquad z_{K-1} = K-1$$

removes this global translation and scale ambiguity from the class geometry. The unconstrained gap logits retain the usual common-shift invariance

$$\mathrm{softmax}(u + c\mathbf{1}) = \mathrm{softmax}(u),$$

which leaves g, z, and the resulting target distribution unchanged.

Table 3: Controlled method taxonomy.

Method	Target / geometry	Score or head	Correspondence
M0 Hard CE	one-hot	categorical CE	standard classification
Hard EMD^2	one-hot	cumulative squared	Hou et al. (Hou et al., 2016)
Hard CDF-CE	one-hot	cumulative log	scoring counterpart
M1 Gaussian CE	fixed $z_k = k$	categorical CE	SORD (Diaz & Marathe, 2019)
M2 Gaussian EMD^2	same as M1	cumulative squared	target-controlled EMD^2
M3 Gaussian CDF-CE	same as M1	cumulative log	target-controlled CDF-CE
M4 learned z	ordered adaptive z	CDF-CE	ours
M5 learned (z, τ)	local + global adaptation	CDF-CE	ours
M6 direct CDF	same target as M5	monotone CDF head	representation diagnostic

Table 4: Evaluation settings for the five datasets.

Dataset	K	Units	Backbone	Optimizer / budget	Select	τ_0
AADB	10	5 seeds	ResNet-18	AdamW, 20 ep.	ρ	1.0
Adience	8	5 folds	ResNet-18	AdamW, 20 ep.	Acc.	1.0
HCI	5	10 splits	VGG-16	Adam, 1700 steps	MAE-EV	0.5
PHAC-2	5	10 adj. means (5 folds each)	multimodal CNN	AdamW, $\leq$100 ep.	MAE^M	1.0
Whisker	9	5 seeds	MLP-32	Adam, 200 ep.	last epoch	1.0

For PHAC-2, fold-level results are first averaged within each adjective and then macro-averaged over the ten adjectives.

D Validity of the Direct CDF Parameterization

The direct cumulative head is constructed such that

$$0 < P_0 < P_1 < \cdots < P_{K-2} < 1.$$

The corresponding categorical probabilities are

$$p_0 = P_0, \qquad p_j = P_j - P_{j-1} \quad (1 \leq j \leq K-2), \qquad p_{K-1} = 1 - P_{K-2}.$$

Strict monotonicity gives

$$p_j > 0 \qquad \text{for all } j,$$

and the probabilities telescope:

$$\sum_{j=0}^{K-1} p_j = P_0 + \sum_{j=1}^{K-2} (P_j - P_{j-1}) + (1 - P_{K-2}) \tag{33}$$

$$= 1. \tag{34}$$

Hence the direct-CDF parameterization always induces a valid categorical distribution.

E Controlled Method Taxonomy

The complete controlled taxonomy is reproduced below for reference.

F Detailed Experimental Protocol

All learned-geometry variants use $\lambda_z = 0.1$. M5–M6 additionally use $\lambda_\tau = 1$ and $R = 2$, which constrains $\tau/\tau_0 \in (0.5, 2)$. We set $\tau_0 = 0.5$ for HCI and $\tau_0 = 1$ for all other datasets.

AADB. We use the predefined data split with 10 ordinal classes and an ImageNet-pretrained ResNet-18 backbone. Images are resized/cropped to 224×224. Models are trained for 20 epochs with AdamW at a learning rate of 3×10^{-4}, weight decay 10^{-4}, a cosine learning-rate schedule, and batch size 64. Checkpoint selection is based on validation Spearman correlation. Results are reported over five random seeds.

Adience. We use the five official subject-exclusive folds and formulate age prediction as an eight-class ordinal problem. The backbone is an ImageNet-pretrained ResNet-18 with 224×224 inputs. Training uses AdamW with learning rate 3×10^{-4}, weight decay 10^{-4}, cosine scheduling, batch size 64, and a 20-epoch budget. Checkpoint selection uses validation accuracy. The fold index is also used as the random seed.

HCI. We evaluate on ten fixed train/validation/test splits with five ordinal classes. The model uses an ImageNet-pretrained VGG-16 backbone. Training images are augmented by random 224×224 cropping and horizontal flipping, while evaluation uses a center crop. Optimization uses Adam with learning rate 10^{-4}, no weight decay or learning-rate scheduler, and an effective batch size of 32 obtained from a physical batch size of 4 with gradient accumulation over 8 steps. Models are trained for 1700 optimizer steps, evaluated every 100 steps, and selected according to validation MAE-EV. The split index is used as the random seed.

PHAC-2. We consider ten haptic adjectives, each evaluated with five object-disjoint folds and discretized into five ordinal rating classes. The input comprises the BioTac PDC, PAC, TDC, TAC, and electrode signals, with each sequence represented at length 512. The multimodal CNN uses modality-specific branches of width 64, 128-dimensional branch representations, a 256-dimensional fusion layer, and dropout 0.1. Training uses AdamW with learning rate 3×10^{-4}, weight decay 10^{-4}, cosine scheduling, and batch size 32. Models are trained for at most 100 epochs, with early stopping enabled after epoch 15 using patience 12 and a minimum improvement of 10^{-4}. Checkpoint selection is based on validation MAE^M. Metrics are first averaged across the five folds within each adjective and then macro-averaged across the ten adjectives.

Whisker. We use Train1–Train9 for training and Test1–Test3 for evaluation, with a contact threshold of 10. Length-5 rolling windows are used only to define temporally aligned valid endpoints: the target depth is taken at the final step, while the MLP receives only the corresponding three-channel pressure measurement x_t. Input channels are standardized using statistics computed from the training set, and depth is discretized into uniform 10-mm bins, yielding nine ordinal classes. The MLP contains three hidden layers of width 32 with dropout 0.1. Training uses Adam with learning rate 10^{-3}, weight decay 10^{-3}, batch size 32, and a fixed budget of 200 epochs. Results are reported over five matched random seeds using the model at the end of the fixed training budget. The continuous-regression control M7 uses the same input representation, hidden layers, optimizer, training budget, and seeds, replacing the ordinal output with a scalar prediction trained using mean squared error.

G Matched-Unit Paired Comparisons

Paired comparison. Matched-unit comparisons show the same overall direction under Gaussian supervision: the direction-corrected $\mathrm{CDF\text{-}CE} - \mathrm{EMD}^2$ difference favors CDF-CE on all five datasets. The 95% paired interval excludes zero on PHAC-2, where CDF-CE improves MAE^M by .0105; intervals on AADB, Adience, HCI, and Whisker overlap zero despite sharing the same mean direction. Under hard targets, PHAC-2 shows the opposite effect, with EMD^2 improving MAE^M over CDF-CE by .0170. The remaining hard-target differences are small relative to their matched variability. Together, these paired comparisons reinforce the factorial result that the relative behavior of the two cumulative scores depends on the target distribution being scored.

H Detailed Evaluation Results

Results are reported as mean $\pm$ population standard deviation over the evaluation units defined in Section F. **Bold** marks the best mean on the pre-specified task-primary metric among M0–M5. M6 is separated because it changes the output representation, and M7 is the continuous-regression control used only for Whisker. Complementary metrics are reported without additional highlighting.

Table 5: AADB results over five seeds. Spearman's ρ is the task-primary metric.

Method	Spearman↑	MAE-EV↓	RMSE-EV↓	Acc.↑	One-off↑
M0	$.648 \pm .014$	$1.311 \pm .081$	$1.627 \pm .111$	$.222 \pm .019$	$.610 \pm .028$
M1	$.675 \pm .006$	$1.228 \pm .013$	$1.511 \pm .011$	$.239 \pm .009$	$.669 \pm .009$
M2	$.674 \pm .006$	$1.232 \pm .020$	$1.519 \pm .017$	$.237 \pm .014$	$.659 \pm .015$
M3	$.676 \pm .009$	$1.219 \pm .015$	$1.506 \pm .014$	$.239 \pm .009$	$.675 \pm .014$
M4	$\mathbf{.678 \pm .007}$	$1.217 \pm .007$	$1.506 \pm .010$	$.239 \pm .006$	$.665 \pm .007$
M5	$.671 \pm .011$	$1.231 \pm .014$	$1.520 \pm .021$	$.241 \pm .009$	$.662 \pm .008$
M6	$.682 \pm .009$	$1.203 \pm .016$	$1.492 \pm .017$	$.251 \pm .007$	$.678 \pm .010$

Table 6: Adience results over the five official subject-exclusive folds. Exact-match accuracy is the task-primary metric.

Method	Acc.↑	One-off↑	MAE-EV↓	RMSE-EV↓	Spearman↑
M0	$.578 \pm .058$	$.914 \pm .017$	$.520 \pm .067$	$.824 \pm .058$	$.888 \pm .020$
M1	$.616 \pm .040$	$.955 \pm .013$	$.535 \pm .034$	$.730 \pm .049$	$.913 \pm .014$
M2	$.616 \pm .041$	$.946 \pm .012$	$.555 \pm .036$	$.744 \pm .048$	$.911 \pm .016$
M3	$\mathbf{.620 \pm .036}$	$.953 \pm .011$	$.541 \pm .039$	$.732 \pm .050$	$.913 \pm .015$
M4	$.615 \pm .040$	$.955 \pm .014$	$.548 \pm .033$	$.732 \pm .047$	$.913 \pm .015$
M5	$.612 \pm .039$	$.948 \pm .013$	$.539 \pm .026$	$.727 \pm .042$	$.911 \pm .016$
M6	$.605 \pm .031$	$.952 \pm .013$	$.543 \pm .030$	$.726 \pm .048$	$.912 \pm .015$

I Representation and Continuous-Regression Controls

M6 tests whether the effects of ordinal supervision persist when the categorical softmax is replaced by a direct monotone-CDF representation. The representation change is beneficial on some ordinal metrics: M6 reaches .682 Spearman on AADB and .768 QWK on Whisker, exceeding the corresponding categorical models on these measures. At the same time, it changes other properties of the predictive distribution; for example, on AADB, NLL/ECE increase from $1.886/.096$ for M3 to $2.068/.186$ for M6. We therefore treat output representation as a separate design axis rather than folding M6 into the target-geometry comparison.

For Whisker, we additionally compare against continuous MSE regression using the same MLP backbone and matched seeds. The regression control obtains 7.560 ± 0.581 mm MAE and 9.824 ± 0.595 mm RMSE, whereas Gaussian CDF-CE achieves 6.778 ± 0.231 mm MAE and 9.612 ± 0.180 mm RMSE. Thus, the ordinal formulation achieves lower mean physical error on both MAE and RMSE despite using discretized supervision.

J Representation and Calibration Diagnostics

Table 7: HCI results over ten fixed splits. MAE-EV is the task-primary metric.

Method	MAE-EV↓	RMSE-EV↓	Spearman↑	QWK↑	Acc.↑	One-off↑
M0	$.715 \pm .040$	$1.078 \pm .058$	$.682 \pm .035$	$.640 \pm .043$	$.532 \pm .019$	$.807 \pm .024$
M1	$.685 \pm .040$	$.982 \pm .041$	$.719 \pm .016$	$.678 \pm .029$	$.564 \pm .032$	$.835 \pm .030$
M2	$.718 \pm .065$	$1.002 \pm .080$	$.694 \pm .048$	$.659 \pm .048$	$.525 \pm .044$	$.823 \pm .035$
M3	$.693 \pm .046$	$.970 \pm .046$	$.718 \pm .030$	$.702 \pm .036$	$.544 \pm .029$	$.854 \pm .026$
M4	$\mathbf{.666 \pm .030}$	$.954 \pm .037$	$.727 \pm .021$	$.708 \pm .027$	$.565 \pm .020$	$.859 \pm .017$
M5	$.689 \pm .048$	$.977 \pm .056$	$.715 \pm .035$	$.697 \pm .042$	$.536 \pm .030$	$.852 \pm .035$
M6	$.690 \pm .057$	$.962 \pm .069$	$.716 \pm .039$	$.704 \pm .051$	$.516 \pm .034$	$.866 \pm .034$

Table 8: PHAC-2 results. Fold-level metrics are first averaged within each adjective and then summarized across the ten adjective-level means. MAE^M is the task-primary metric.

Method	MAE^M ↓	Acc.↑	One-off↑	CDF-MAE↓	Mean MAE↓	Mean RMSE↓
M0	$.689 \pm .155$	$.627 \pm .188$	$.874 \pm .089$	$.144 \pm .053$	$.519 \pm .199$	$.690 \pm .227$
M1	$.632 \pm .150$	$.639 \pm .175$	$.900 \pm .078$	$.149 \pm .040$	$.568 \pm .158$	$.683 \pm .205$
M2	$.629 \pm .155$	$.612 \pm .192$	$.889 \pm .090$	$.148 \pm .042$	$.560 \pm .162$	$.680 \pm .212$
M3	$.618 \pm .148$	$.643 \pm .175$	$.897 \pm .084$	$.146 \pm .040$	$.553 \pm .157$	$.674 \pm .206$
M4	$.605 \pm .160$	$.643 \pm .180$	$.897 \pm .083$	$.143 \pm .043$	$.539 \pm .170$	$.662 \pm .218$
M5	$\mathbf{.593 \pm .166}$	$.650 \pm .174$	$.893 \pm .096$	$.140 \pm .045$	$.529 \pm .177$	$.656 \pm .221$
M6	$.594 \pm .167$	$.622 \pm .193$	$.904 \pm .086$	$.142 \pm .045$	$.541 \pm .182$	$.660 \pm .223$

Table 9: Whisker results over five matched seeds. Depth MAE is the task-primary metric. M7 is a continuous-regression control; its categorical metrics are obtained by rebinning continuous predictions into the same 10-mm ordinal bins.

Method	Depth MAE↓	Depth RMSE↓	QWK↑	Acc.↑	One-off↑
M0	$7.928 \pm .572$	$10.780 \pm .329$	$.693 \pm .016$	$.412 \pm .034$	$.809 \pm .024$
M1	$7.143 \pm .508$	$10.176 \pm .384$	$.733 \pm .016$	$.477 \pm .022$	$.851 \pm .012$
M2	$7.064 \pm .494$	$9.517 \pm .388$	$.696 \pm .074$	$.382 \pm .101$	$.766 \pm .077$
M3	$\mathbf{6.778 \pm .231}$	$9.612 \pm .180$	$.753 \pm .009$	$.525 \pm .032$	$.841 \pm .006$
M4	$7.182 \pm .189$	$9.837 \pm .193$	$.743 \pm .011$	$.510 \pm .025$	$.843 \pm .008$
M5	$7.114 \pm .171$	$9.799 \pm .161$	$.746 \pm .011$	$.528 \pm .032$	$.842 \pm .003$
M6	$7.293 \pm .312$	$9.765 \pm .230$	$.768 \pm .006$	$.491 \pm .041$	$.859 \pm .009$
M7 reg.	$7.560 \pm .581$	$9.824 \pm .595$	$.753 \pm .028$	$.450 \pm .042$	$.854 \pm .029$

Table 10: **Calibration diagnostics for the Gaussian-CDF-CE, adaptive-geometry, and direct-CDF variants.** NLL and ECE are complementary diagnostics and are not used for checkpoint selection or primary method comparison. M3–M5 retain the categorical softmax representation, whereas M6 replaces it with a direct monotone-CDF output. For PHAC-2, values are summarized over the ten adjective-level means after fold averaging. Lower is better for both metrics.

Dataset	Method	NLL↓	ECE↓
AADB	M3 Gaussian CDF-CE	$1.886 \pm .016$	$.096 \pm .009$
	M4 learned z	$1.915 \pm .024$	$.116 \pm .007$
	M5 learned (z, τ)	$1.993 \pm .026$	$.138 \pm .009$
	M6 direct CDF	$2.068 \pm .034$	$.186 \pm .007$
Adience	M3 Gaussian CDF-CE	$1.130 \pm .047$	$.232 \pm .027$
	M4 learned z	$1.106 \pm .046$	$.212 \pm .037$
	M5 learned (z, τ)	$1.078 \pm .047$	$.187 \pm .033$
	M6 direct CDF	$1.060 \pm .050$	$.155 \pm .026$
HCI	M3 Gaussian CDF-CE	$1.246 \pm .068$	$.110 \pm .017$
	M4 learned z	$1.217 \pm .065$	$.101 \pm .029$
	M5 learned (z, τ)	$1.262 \pm .091$	$.131 \pm .027$
	M6 direct CDF	$1.376 \pm .075$	$.167 \pm .019$
PHAC-2	M3 Gaussian CDF-CE	$1.017 \pm .292$	$.247 \pm .083$
	M4 learned z	$1.010 \pm .309$	$.244 \pm .061$
	M5 learned (z, τ)	$.981 \pm .301$	$.228 \pm .058$
	M6 direct CDF	$1.004 \pm .322$	$.224 \pm .045$
Whisker	M3 Gaussian CDF-CE	$1.479 \pm .010$	$.232 \pm .032$
	M4 learned z	$1.477 \pm .009$	$.214 \pm .025$
	M5 learned (z, τ)	$1.464 \pm .007$	$.225 \pm .032$
	M6 direct CDF	$1.432 \pm .015$	$.153 \pm .043$