Ordinal Label-Distribution Learning with Constrained Asymmetric Priors for Imbalanced Retinal Grading

Nagur Shareef Shaik[1], Teja Krishna Cherukuri[1], Adnan Masood[2], Ehsan Adeli[3], Dong Hye Ye[1]

[1]*Georgia State University, Atlanta, GA 30303;* [2]*UST Global Inc., Aliso Viejo, CA 92656*

[3]*Stanford University, Palo Alto, CA 94305*

NEURAL INFORMATION PROCESSING SYSTEMS

The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance – NeurIPS 2025

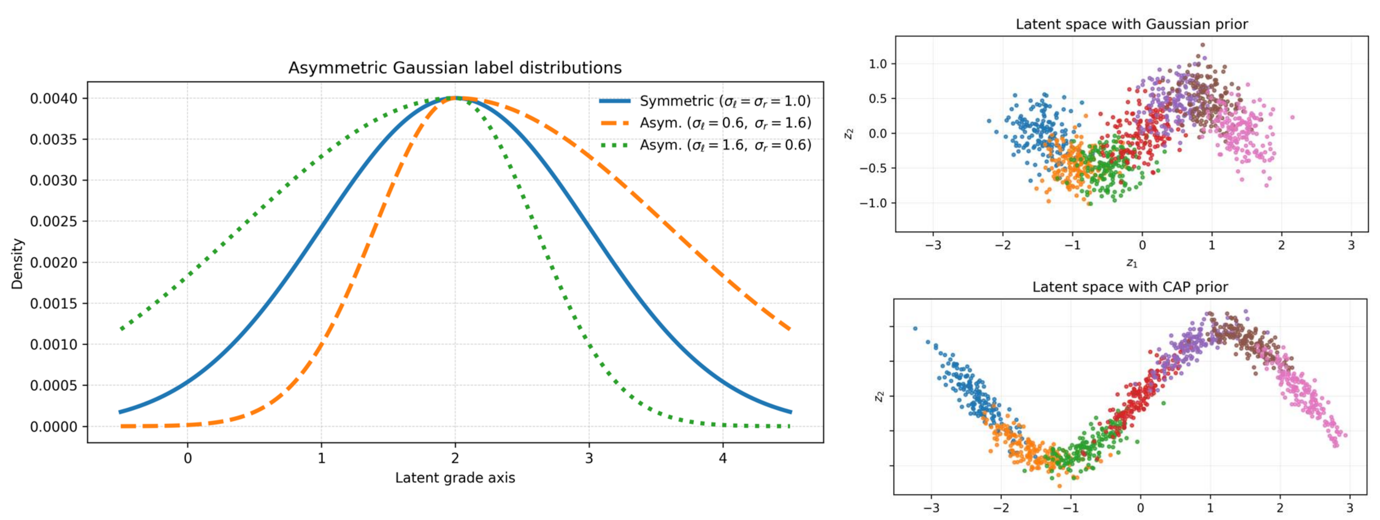
1. Introduction

1.1 Motivation

- Diabetic Retinopathy (DR) grading is **ordinal** and **severely imbalanced**.
- Real-world labels display **asymmetric ambiguity** (under-grading risk > over-grading risk).
- Standard VAEs with symmetric Gaussian priors **collapse minority modes**.
- Purely discriminative models fail to maintain **ordinal geometry** of disease progression.
- Need a framework that:
  ✓ preserves minority structure
  ✓ respects ordering
  ✓ handles clinical asymmetry
  ✓ generalizes across domains

1.2 Key Observations

- DR datasets follow a **long-tail** distribution → rare severe stages.
- Latent spaces from conventional VAEs are **spherical, overlapping**, and not severity-aligned.
- Asymmetric label distributions reflect clinical decision boundaries more realistically.
- Ordinal label-distribution learning (OLDL) outperforms one-hot classification but **still assumes symmetric distributions** and lacks a geometry prior.



(a) Direction-aware ordinal label distributions allocate asymmetric probability mass around the reference grade. (b) Latent manifolds: a spherical Gaussian prior (left) contracts minority modes and overlaps grades, whereas a *constrained asymmetric prior* (right) preserves skew/tails and yields grade-ordered separability.

1.3 Contributions

- **Constrained Asymmetric Prior (AGGD)**
  A data-estimated, heavy-tailed, left–right asymmetric prior
- **Wasserstein Aggregate Alignment**
  Uses MMD to align latent distribution to the asymmetric prior.
- **Direction-Aware Ordinal Supervision**
  Predicts left/right spreads (σ_l, σ_r) → generates **asymmetric soft labels**
- **Margin-Aware Orthogonality & Compactness (MAOC)**
  Encourages class prototypes to be nearly orthogonal & shrinks intra-class variance

Together → a generative-discriminative, ordinal-aware, clinically aligned framework.

GeorgiaState University · TReNDS Translational Research in Neuroimaging & Data Science · UST · Stanford University

2. Methodology

2.1 Architecture Overview

- **Encoder:** VGG16 backbone + 2 FC layers → latent vector z (dim 512)
- **Decoder:** symmetric transposed-conv decoder
- **Heads:**
  - **Classification head:** logits for 5 - 7 severity levels
  - **AG-Soft head:** predicts left/right dispersions (log σ_l, log σ_r)
  - **Ordinal regression head:** scalar severity score

Goal: Representations must respect **disease severity ordering**, **minority structure**, and **clinical asymmetry**.

2.2 Constrained Asymmetric Prior (AGGD)

We replace the Gaussian prior with an **Asymmetric Generalized Gaussian Distribution** (AGGD), factorized across latent dimensions.

Per-dimension prior $p_{\text{cap}}(u;\eta) = \dfrac{\beta}{(\alpha_\ell + \alpha_r)\,\Gamma(1/\beta)} \times \begin{cases} \exp\left(-\left(\frac{|\mu - u|}{\alpha_\ell}\right)^\beta\right), & u < \mu, \\ \exp\left(-\left(\frac{|u - \mu|}{\alpha_r}\right)^\beta\right), & u \geq \mu. \end{cases}$

$$\mu_j = \frac{1}{N}\sum_i z_{i,j}, \qquad \alpha_{\ell,j} = \sqrt{\frac{1}{N_\ell}\sum_{i:z_{i,j}<\mu_j}(\mu_j - z_{i,j})^2}, \qquad \alpha_{r,j} = \sqrt{\frac{1}{N_r}\sum_{i:z_{i,j}\geq\mu_j}(z_{i,j} - \mu_j)^2},$$

Captures **skew**, **asymmetry**, and **heavy-tails**, which occur in severe DR stages. β controls tail heaviness (β=2 Gaussian, β=1 Laplace, smaller β heavier-tailed).

We align the aggregate posterior $q_\phi(z) = \mathbb{E}_{p(x)}[q_\phi(z|x)]$ to our fixed prior p_{cap} using the Maximum Mean Discrepancy (MMD) penalty,

$$\text{MMD}^2(q,p) = \frac{1}{n^2}\sum_{i,j} k(z_i, z_j) + \frac{1}{m^2}\sum_{i,j} k(\tilde{z}_i, \tilde{z}_j) - \frac{2}{nm}\sum_{i,j} k(z_i, \tilde{z}_j)$$

2.3 Order-Sensitive Supervision

1. Hard Classification: $\mathcal{L}_{\text{CE}} = -\log\left(\text{softmax}_y(\ell)\right), \quad \text{where } \ell = h_{\psi_{\text{cls}}}(z).$
2. Asymmetric Gaussian Soft Labels (AG-Soft): instance-wise dispersions

$$p_{\text{AG}}(j \mid y, \sigma_\ell, \sigma_r) \;\propto\; \exp\left(-\frac{(j-y)^2}{2\left(\sigma_\ell^2\,\mathbb{I}[j<y] + \sigma_r^2\,\mathbb{I}[j>y] + \sigma_{\text{mid}}^2\,\mathbb{I}[j=y]\right) + \varepsilon}\right)$$

$$\mathcal{L}_{\text{AG}} = -\sum_{j=0}^{C-1} p_{\text{AG}}(j \mid y, \sigma_\ell, \sigma_r)\log\left(\text{softmax}_j(\ell)\right).$$

3. Ordinal Regression (ORM): $\mathcal{L}_{\text{ORM}} = \text{Huber}_\tau(s - y)$
4. Latent Geometry Regularization:

$$\mathcal{L}_{\text{MAOC}} = \underbrace{\frac{1}{|\mathcal{P}|}\sum_{c\neq c'}\left[\max\left(0,\, \hat{\mu}_c^\top \hat{\mu}_{c'} - \delta\right)\right]^2}_{\text{Margin-Aware Orthogonality}} + \gamma_{\text{cmp}} \cdot \underbrace{\frac{1}{N}\sum_{i=1}^{N}\|z_i - \mu_{y_i}\|_2^2}_{\text{Intra-Class Compactness}},$$

5. Complete Objective and Training:

$$\min_{\Theta,s}\; \mathbb{E}_{(x,y)\sim p_{\text{data}}}\Big[\underbrace{\|\, g_\theta(f_\phi(x)) - x\,\|_2^2}_{\mathcal{L}_{\text{recon}}} + \lambda_{\text{reg}}\underbrace{\text{MMD}^2\left(q_\phi(z),\, p_{\text{cap}}\right)}_{\mathcal{L}_{\text{reg}}} + \lambda_{\text{maoc}}\underbrace{\mathcal{L}_{\text{MAOC}}(Z)}_{\text{Geometric Regularization}} + \underbrace{\sum_{k\in\{\text{CE, AG, ORM}\}}\left(e^{-s_k}\mathcal{L}_k + s_k\right)}_{\text{Adaptive Ordinal Supervision}}\Big],$$

3. Experimental Results

3.1 Dataset Details

Table 1: Summary of diabetic retinopathy datasets used for evaluation.

Dataset	**Camera system**	**# Classes**	**Train**	**Test**	**Total**
Zenodo-DR-7	Zeiss Visucam 500	7	605	152	757
IDRiD	Kowa VX-10a	5	413	103	516
APTOS-2019	Heterogeneous	5	2,929	733	3,662
Messidor-2	Topcon TRC-NW6	5	1,398	350	1,748

3.2 Quantitative Results

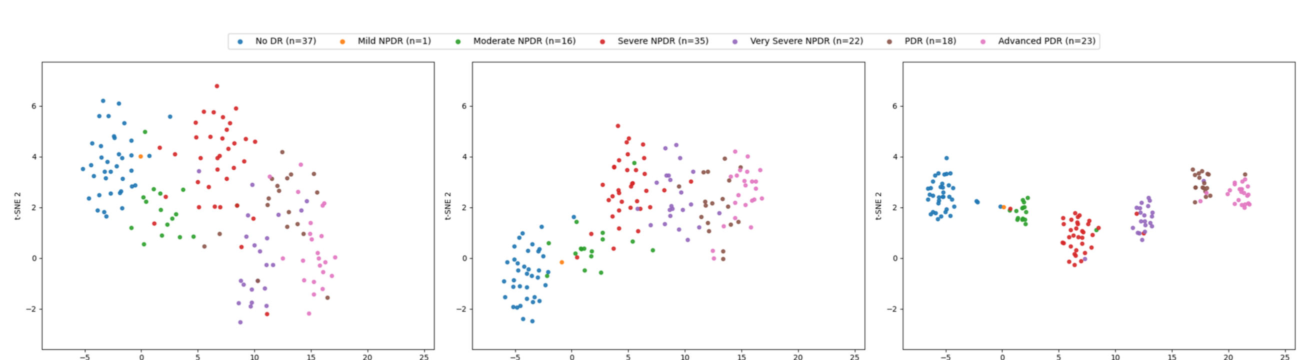
Table 2: **Main comparisons** on four DR benchmarks. Quadratic Weighted Kappa (QWK), Accuracy (Acc, %), and Macro-F1 (F1, %). †: values reported by original papers; Best in **bold**; Methods are grouped as imbalance/ordinal baselines, latent generative baselines, attention/gating models, and discriminator-based grading networks for clarity

	Zenodo-DR-7 (7)			**APTOS-2019** (5)			**Messidor-2** (5)			**IDRiD** (5)		
Method	QWK	Acc	F1	QWK	Acc	F1	QWK	Acc	F1	QWK	Acc	F1
CE (balanced)	0.83	86.10	83.00	0.82	82.50	76.20	0.81	81.00	76.80	0.78	78.80	74.20
Focal (γ=2)	0.85	87.20	84.10	0.83	83.80	77.70	0.82	82.00	77.80	0.79	79.80	75.10
Logit-Adj	0.86	87.90	84.80	0.84	84.90	78.90	0.83	82.80	78.60	0.80	80.60	75.90
LDAM-DRW	0.87	88.50	85.60	0.85	84.60	79.80	0.84	83.50	79.40	0.81	81.50	76.80
CORN	0.88	89.10	86.20	0.86	84.20	80.40	0.85	84.10	80.10	0.82	82.30	77.60
OLDL (S)	0.89	89.70	86.80	0.87	84.70	81.10	0.86	84.60	80.70	0.83	82.90	78.30
OLDL (AS)	0.90	90.00	87.50	0.88	85.90	82.30	0.87	85.30	81.60	0.84	83.80	79.10
VAE–KL	0.84	86.70	83.50	0.83	83.40	77.90	0.82	82.20	77.40	0.79	80.20	75.40
WAE–MMD	0.86	88.00	84.90	0.85	85.00	79.50	0.84	83.60	79.00	0.81	81.60	76.50
ViT†	–	84.61	83.19	–	83.22	67.83	–	76.79	61.47	–	61.17	46.18
GCG†	0.931	90.13	88.49	–	85.29	70.57	–	80.23	73.85	–	72.14	68.34
DGN	0.87	88.30	85.20	0.84	84.60	78.40	0.83	83.00	78.80	0.80	80.80	76.20
AGDGN	0.89	89.50	86.50	0.86	86.10	80.10	0.85	84.30	80.40	0.82	82.10	77.80
AGDGN+OLDL	0.90	89.90	87.20	0.87	86.80	81.50	0.86	84.90	81.10	0.83	82.70	78.60
CAP-WAE	**0.94**	**91.80**	**89.90**	**0.90**	**87.14**	**83.64**	**0.89**	**86.90**	**83.00**	**0.87**	**86.10**	**81.20**

Table 3: **Ablation study** of CAP-WAE across four DR benchmarks. We progressively add key components: asymmetric prior (AS), Wasserstein/MMD alignment, AG-Soft supervision, ORM head, and latent geometry priors (MAOC).

	Zenodo-DR-7 (7)			**APTOS-2019** (5)			**Messidor-2** (5)			**IDRiD** (5)		
Variant	QWK	Acc	F1	QWK	Acc	F1	QWK	Acc	F1	QWK	Acc	F1
VAE–KL	0.84	86.70	83.50	0.83	83.40	77.90	0.82	82.20	77.40	0.79	80.20	75.40
+ AS. (KL)	0.85	87.40	84.30	0.84	84.20	78.50	0.83	83.10	78.20	0.80	81.10	76.00
WAE–MMD	0.86	88.00	84.90	0.85	85.00	79.50	0.84	83.60	79.00	0.81	81.60	76.50
+ AS. (MMD)	0.88	89.00	86.00	0.86	86.00	80.60	0.85	84.60	80.20	0.82	82.50	77.40
+ AG-Soft	0.89	89.50	86.60	0.87	86.50	81.10	0.86	85.30	81.00	0.83	83.20	78.20
+ ORM	0.90	90.00	87.20	0.88	86.90	81.70	0.87	85.80	81.50	0.84	83.70	78.80
+ MAOC	0.91	91.00	88.50	0.89	87.00	82.20	0.88	86.20	82.40	0.85	84.40	79.50
CAP-WAE	**0.94**	**91.80**	**89.90**	**0.90**	**87.14**	**83.64**	**0.89**	**86.90**	**83.00**	**0.87**	**86.10**	**81.20**

3.3 Qualitative Results



(a) VAE-KL (b) WAE-MMD (c) CAP-WAE

References

[1] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schölkopf. *Wasserstein Auto-Encoders.* ICLR, 2018.
[2] C. Wen, X. Zhang, X. Yao, and J. Yang. *Ordinal Label Distribution Learning.* ICCV, 2023.
[3] T. Elguebaly and N. Bouguila. *Finite Asymmetric Generalized Gaussian Mixture Models Learning.* Computer Vision and Image Understanding, 2013.
[4] V. M. Vargas, P. A. Gutiérrez, R. Rosati, et al. *Disease-Grading Networks with Asymmetric Gaussian Distribution for Medical Imaging.* IEEE Transactions on Medical Imaging, 2025.
[5] W. Cao, V. Mirjalili, and S. Raschka. *Rank-Consistent Ordinal Regression for Neural Networks.* AAAI, 2021.