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

Nagur Shareef Shaik
Georgia State University
Atlanta, GA 30303
nshaik3@student.gsu.edu

Teja Krishna Cherukuri
Georgia State University
Atlanta, GA 30303
tcherukuri1@student.gsu.edu

Adnan Masood
UST Global Inc
Aliso Viejo, CA 92656
adnan.massod@ust.com

Ehsan Adeli
Stanford University
Palo Alto, CA 94305
eadeli@stanford.edu

Dong Hye Ye
Georgia State University
Atlanta, GA 30303
dongye@gsu.edu

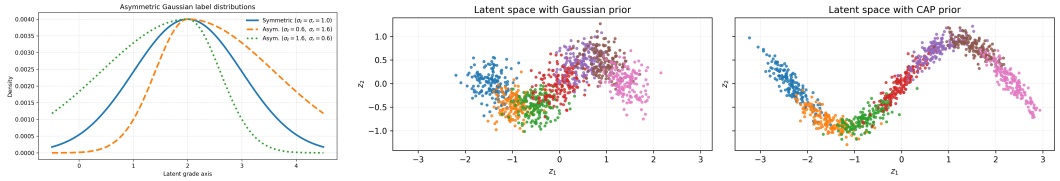
Abstract

Diabetic retinopathy grading is inherently ordinal and long-tailed, with minority stages being scarce, heterogeneous, and clinically critical to detect accurately. Conventional methods often rely on isotropic Gaussian priors and symmetric loss functions, misaligning latent representations with the task's asymmetric nature. We propose the Constrained Asymmetric Prior Wasserstein Autoencoder (CAP-WAE), a novel framework that addresses these challenges through three key innovations. Our approach employs a Wasserstein Autoencoder (WAE) that aligns its aggregate posterior with a *asymmetric prior*, preserving the heavy-tailed and skewed structure of minority classes. The latent space is further structured by a *Margin-Aware Orthogonality and Compactness (MAOC)* loss to ensure grade-ordered separability. At the supervision level, we introduce a direction-aware ordinal loss, where a lightweight head predicts asymmetric dispersions to generate soft labels that reflect clinical priorities by penalizing under-grading more severely. Stabilized by an adaptive multi-task weighting scheme, our end-to-end model requires minimal tuning. Across public DR benchmarks, CAP–WAE consistently achieves state-of-the-art Quadratic Weighted Kappa, accuracy, and macro-F1, surpassing both ordinal classification and latent generative baselines. t-SNE visualizations further reveal that our method reshapes the latent manifold into compact, grade-ordered clusters with reduced overlap.

1 Introduction

The steady rise of diabetes worldwide has precipitated a silent epidemic of preventable blindness, driven by *diabetic retinopathy (DR)*, a microvascular complication and leading cause of vision loss among working-age adults globally [16, 9]. Prolonged hyperglycemia damages retinal vasculature, progressing through stages defined by the International Clinical DR Disease Severity Scale, an ordinal taxonomy spanning five levels from no apparent retinopathy to proliferative DR [30]. Timely detection and accurate grading are critical, with estimates suggesting that up to 90% of severe vision impairment can be prevented through routine screening and treatment [9, 18]. However, manual grading is labor-intensive, prone to inter-grader variability, and difficult to scale as diabetes prevalence grows, motivating research into computer-aided DR grading from retinal fundus images [14, 8].

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.



(a) Asymmetric label distributions (b) Latent geometry: Gaussian vs. CAP prior

Figure 1: **(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.

Clinical deployment of automated systems faces two intertwined challenges. First, DR datasets are long-tailed: early and severe grades are underrepresented, biasing models toward majority grades and reducing sensitivity to critical minority cases [5, 4]. Second, subtle lesion-based differences between adjacent grades exacerbate misclassification, particularly under cross-center and cross-device domain shifts that degrade generalization [23, 22]. These factors highlight the need for methods that respect the *ordinal* nature of DR severity while handling imbalanced label distributions and domain variability [29]. Generative formulations that model latent structures aligned with disease progression offer a promising alternative to purely discriminative approaches, which often assume simplistic, symmetric class geometries [27].

Conventional approaches mitigate imbalance with reweighting or logit adjustments and handle domain shifts with adaptation techniques [17, 18], but they remain limited: decision boundaries are nudged toward majority grades, and one-hot supervision treats all errors equally, ignoring ordinal severity. Recent advances in *ordinal label-distribution learning (OLDL)* address this by predicting distributions over ordered grades and penalizing errors by distance along the severity scale [29]. Complementary studies show that asymmetric label distributions better capture clinical ambiguity than symmetric Gaussians, improving sensitivity at decision boundaries [27], while explicit ordering priors, such as language-driven alignment, enhance robustness under distribution shifts [28]. These insights, illustrated in Figure 1, suggest two gaps: (i) latent priors remain predominantly symmetric Gaussians that collapse minority structures, and (ii) supervision rarely accounts for directional clinical risks of over- vs. under-grading.

We propose the *Constrained Asymmetric Prior Wasserstein Autoencoder* (**CAP-WAE**), the first framework to jointly integrate *constrained asymmetric priors*, *direction-aware ordinal supervision*, and *Margin-Aware Orthogonality and Compactness* for robust DR grading. Unlike prior models that rely on symmetric Gaussian priors and one-hot supervision, CAP-WAE aligns the aggregate posterior to a skew- and tail-aware prior, preserving minority structures in the latent space. At the supervision level, it employs ordinal label distributions that penalize under-grading more severely, while adaptive multi-task weighting ensures stable, tuning/light training. Together, these innovations yield grade-ordered, well-separated manifolds and clinically aligned predictions that generalize across imbalanced and heterogeneous datasets.

The CAP-WAE framework is designed to address the following:

- **Constrained asymmetric prior with Margin-Aware Orthogonality and Compactness.** We introduce a Wasserstein autoencoder that aligns the aggregate posterior to a skew- and tail-aware *constrained asymmetric prior*, while enforcing geometric regularization to preserve minority structures and yield grade-ordered, well-separated manifolds.
- **Direction-aware ordinal supervision with tuning-light training.** We couple direction-aware ordinal label distributions with adaptive multi-task weighting, enabling clinically aligned grading and robust generalization without extensive hyperparameter tuning.

2 Preliminaries

Notation. Let $\mathcal{Y} = \{0, 1, \ldots, C-1\}$ denote ordered DR grades and Δ^{C-1} the probability simplex. We use $\mathbb{I}[\cdot]$ for indicator functions and write $q_\theta(x) \in \Delta^{C-1}$ for predicted label distributions.

2.1 Ordinal Label-Distribution Learning (OLDL)

Label Distribution Learning models a target distribution $d \in \Delta^{C-1}$ rather than a one-hot label and minimizes a divergence $\mathcal{D}(d, q_\theta(x))$; OLDL additionally encodes *order-awareness* so that larger grade gaps are penalized more strongly [29].

Order-aware divergences. Two widely used families are: (i) *Cumulative* distances on the discrete line [21]. With $F_d(k) = \sum_{i \le k} d_i$ and $F_q(k) = \sum_{i \le k} q_i$,

$$\mathrm{CAD}(d, q) = \sum_{k=0}^{C-1} \big|F_d(k) - F_q(k)\big|, \tag{1}$$

which coincides with the L_1 distance between discrete CDFs and is equivalent to the 1-Wasserstein distance on $\mathcal{Y}$. (ii) *Quadratic-form* distances that weight errors by grade gaps:

$$\mathrm{QFD}(d, q) = (d - q)^\top L (d - q), \qquad L = \mathrm{Diag}(W\mathbf{1}) - W, \;\; W_{ij} = \omega(|i - j|), \tag{2}$$

where L is the graph Laplacian induced by a nonnegative kernel ω (e.g., linear or quadratic in $|i - j|$) [29]. OLDL variants also use Jensen–Shannon-type divergences on cumulative distributions, improving rank-sensitive metrics.

2.2 Asymmetric Label Distributions in Medical Grading

Clinical ambiguity is often *directional*: the risk of over- vs. under-grading differs near boundaries. Medical imaging models therefore replace symmetric Gaussians with *asymmetric*, instance-wise distributions by predicting left/right spreads [27]. A standard two-sided Gaussian parameterization is

$$p(j \mid \mu, \sigma_\ell, \sigma_r) \;\propto\; \exp\!\left(- \frac{(j - \mu)^2}{2\big(\sigma_\ell^2 \, \mathbb{I}[j < \mu] + \sigma_r^2 \, \mathbb{I}[j \ge \mu]\big) + \varepsilon} \right), \quad j \in \mathcal{Y}, \tag{3}$$

with (σ_ℓ, σ_r) predicted per instance. This *direction-aware* spread calibrates decision boundaries and improves sensitivity under class imbalance and label ambiguity [27, 18].

2.3 Latent Distribution Learning

VAEs. Variational Autoencoders maximize the ELBO

$$\mathcal{L}_{\mathrm{ELBO}} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \mathrm{KL}(q_\phi(z|x) \,\|\, p(z)) \,, \tag{4}$$

typically with $p(z) = \mathcal{N}(0, I)$ [15, 20]. The *pointwise* KL term pushes every posterior $q_\phi(z|x)$ toward the spherical prior; under long-tailed regimes this can suppress non-Gaussian structure characteristic of minority classes.

WAEs and aggregate alignment. Wasserstein Autoencoders align the *aggregate* posterior $q_\phi(z) = \mathbb{E}_{p(x)}[q_\phi(z|x)]$ to a chosen prior $p(z)$ via an optimal-transport-style penalty [26]:

$$\mathcal{L}_{\mathrm{WAE}} = \mathbb{E}_{p(x)} \mathbb{E}_{q_\phi(z|x)}\big[\ell(x, \tilde{x})\big] + \lambda \, \mathcal{D}\big(q_\phi(z), p(z)\big), \quad \tilde{x} \sim p_\theta(x|z), \tag{5}$$

instantiated with Maximum Mean Discrepancy (MMD) using a characteristic kernel k [13]. Given samples $\{z_i\}_{i=1}^n \sim q$ and $\{\tilde{z}_j\}_{j=1}^m \sim p$, a standard (biased) estimator is

$$\mathrm{MMD}^2(q, p) = \frac{1}{n(n-1)} \sum_{i \ne j} k(z_i, z_j) + \frac{1}{m(m-1)} \sum_{i \ne j} k(\tilde{z}_i, \tilde{z}_j) - \frac{2}{nm} \sum_{i,j} k(z_i, \tilde{z}_j). \tag{6}$$

MMD-based aggregate matching is stable, sample-driven, and accommodates non-Gaussian/heavy-tailed priors, crucial for preserving minority structure in imbalanced settings. We adopt the same machinery in §3 (see Eq. (6)) for aligning to an asymmetric prior.

2.4 Asymmetric Generalized Gaussian (AGGD): Conceptual Background

To model skew and tail-heaviness in latents, we leverage the *asymmetric generalized Gaussian* family, which augments the generalized Gaussian with a location μ, a shape (tail) parameter β, and distinct left/right scales (α_ℓ, α_r) [12]. Intuitively, β tunes tail thickness (smaller $\beta \Rightarrow$ heavier tails), while (α_ℓ, α_r) control asymmetry around μ. We provide the formal pdf, moments and use an AGGD sampler to draw prior samples for MMD alignment.

3 Methodology

Our goal is to develop a robust framework for automated diabetic retinopathy (DR) grading that addresses long-tailed class distributions, ordinal severity scales, and clinical domain shifts. We propose the *Constrained Asymmetric Prior Wasserstein Autoencoder* (CAP-WAE), which integrates generative latent modeling with order-sensitive supervision and explicit geometric regularization. We formulate retinal grading as an ordinal multi-class problem on the ordered labels $\mathcal{Y} = \{0, 1, \ldots, C-1\}$. Given an image $x \in \mathbb{R}^{H\times W\times 3}$ with grade $y \in \mathcal{Y}$, our model learns: (i) a latent representation z whose aggregate distribution matches a data-informed asymmetric prior, preserving minority class structure; and (ii) an order-sensitive predictor trained with direction-aware supervision to produce clinically aligned and robust predictions.

3.1 Architecture

The CAP-WAE architecture consists of a deterministic encoder-decoder backbone with three lightweight heads operating on the shared latent representation.

Encoder-Decoder. The encoder $f_\phi : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^d$ is a VGG16 backbone (pre-trained on ImageNet) [24] followed by two fully connected layers, producing a latent vector $z = f_\phi(x)$. The decoder $g_\theta : \mathbb{R}^d \to \mathbb{R}^{H\times W\times 3}$ uses a symmetric structure of transposed convolutions to reconstruct the input image $\tilde{x} = g_\theta(z)$.

Prediction Heads. Three shallow MLP heads are attached to the latent vector z: a classifier head $h_{\psi_{\text{cls}}} : \mathbb{R}^d \to \mathbb{R}^C$ that outputs logits ℓ for class prediction; an Asymmetric-Gaussian (AG) head $a_{\psi_{\text{ag}}} : \mathbb{R}^d \to \mathbb{R}^2$ that predicts left/right log-dispersions $(\log\sigma_\ell, \log\sigma_r)$ used to construct direction-aware soft labels [27]; and an ordinal regression (ORM) head $r_{\psi_{\text{orm}}} : \mathbb{R}^d \to \mathbb{R}$ that predicts a scalar severity score s aligned with the grade index [25].

3.2 Constrained Asymmetric Prior

To capture the skewed, heavy-tailed nature of features in long-tailed medical datasets, we replace the standard spherical Gaussian prior with an *Asymmetric Generalized Gaussian Distribution (AGGD)*. The "constrained" aspect of our prior, which we term p_{cap}, refers to the fact that its parameters are fixed and empirically determined from the data distribution, providing a stable, informative target for latent space alignment.

Prior Definition (AGGD). We use a factorized prior $p_{\text{cap}}(z) = \prod_{j=1}^{d} p_{\text{cap}}(z_j)$, where each coordinate follows an AGGD with parameters $\eta_j = (\mu_j, \beta_j, \alpha_{\ell,j}, \alpha_{r,j})$. The coordinate-wise pdf is:

$$p_{\text{cap}}(u;\eta) = \frac{\beta}{(\alpha_\ell+\alpha_r)\,\Gamma(1/\beta)} \times \begin{cases} \exp\big(-\big(\frac{|\mu-u|}{\alpha_\ell}\big)^\beta\big), & u < \mu, \\ \exp\big(-\big(\frac{|u-\mu|}{\alpha_r}\big)^\beta\big), & u \geq \mu. \end{cases} \tag{7}$$

Here, μ is the location (center), β controls tail heaviness (β=2 Gaussian, β=1 Laplace, smaller β heavier-tailed), and α_ℓ, α_r are left/right scales governing asymmetry. Given latent vectors $\{z_i\}_{i=1}^N$, for each coordinate j,:

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

where N_ℓ, N_r are counts on each side of μ_j.

Aggregate Alignment via MMD. 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, as defined in Eq. (6). This distribution-level matching allows the model to preserve the non-Gaussian structures of minority classes, which would otherwise be lost under the instance-level KL divergence penalty in a standard VAE.

3.3 Order-Sensitive Supervision

To leverage the ordinal nature of DR grades, our supervision combines three complementary loss terms.

Hard Classification. A standard cross-entropy loss on the one-hot target y:

$$\mathcal{L}_{\mathrm{CE}} = -\log\big(\mathrm{softmax}_y(\ell)\big), \quad \text{where } \ell = h_{\psi_{\mathrm{cls}}}(z). \tag{8}$$

Asymmetric Gaussian Soft Labels (AG-soft). The AG head predicts instance-wise dispersions $(\sigma_\ell, \sigma_r) = \exp(a_{\psi_{\mathrm{ag}}}(z))$, which are clamped to a stable range $[\sigma_{\min}, \sigma_{\max}]$. These define a direction-aware target distribution:

$$p_{\mathrm{AG}}(j \mid y, \sigma_\ell, \sigma_r) \;\propto\; \exp\left(-\frac{(j-y)^2}{2\big(\sigma_\ell^2\,\mathbb{I}[j<y] + \sigma_r^2\,\mathbb{I}[j>y] + \sigma_{\mathrm{mid}}^2\,\mathbb{I}[j=y]\big) + \varepsilon}\right), \tag{9}$$

where $\sigma_{\mathrm{mid}} = (\sigma_\ell + \sigma_r)/2$ ensures a well-defined peak. This target is used in a soft cross-entropy loss, penalizing under- or over-grading differently based on the learned dispersions:

$$\mathcal{L}_{\mathrm{AG}} = -\sum_{j=0}^{C-1} p_{\mathrm{AG}}(j \mid y, \sigma_\ell, \sigma_r)\log\big(\mathrm{softmax}_j(\ell)\big). \tag{10}$$

Ordinal Regression (ORM). The ORM head predicts a scalar score $s = r_{\psi_{\mathrm{orm}}}(z)$, which is trained to align with the integer grade y using a Huber loss:

$$\mathcal{L}_{\mathrm{ORM}} = \mathrm{Huber}_\tau\big(s - y\big). \tag{11}$$

3.4 Latent Geometry Regularization

To enforce a structured latent space with well-separated and compact class clusters, we introduce a single geometric loss term, the **Margin-Aware Orthogonality and Compactness (MAOC)** loss. This loss combines two principles from prototype-based learning. Let μ_c be the running mean of latent vectors for class c, and $\hat{\mu}_c = \mu_c/\|\mu_c\|_2$.

$$\mathcal{L}_{\mathrm{MAOC}} = \underbrace{\frac{1}{|\mathcal{P}|}\sum_{c\neq c'}\big[\max\big(0,\, \hat{\mu}_c^\top \hat{\mu}_{c'} - \delta\big)\big]^2}_{\text{Margin-Aware Orthogonality}} + \gamma_{\mathrm{cmp}} \cdot \underbrace{\frac{1}{N}\sum_{i=1}^{N}\|z_i - \mu_{y_i}\|_2^2}_{\text{Intra-Class Compactness}}, \tag{12}$$

where $\mathcal{P}$ is the set of class pairs, δ is an angular margin, and γ_{cmp} is a hyperparameter balancing the two components. This encourages class prototypes to be nearly orthogonal while pulling individual samples toward their respective class centers.

3.5 Complete Objective and Training

The full CAP-WAE model is trained end-to-end by minimizing a composite objective. To balance the multiple supervised loss terms without extensive hyperparameter tuning, we use an uncertainty-based adaptive weighting scheme. Let Θ denote all network parameters and $s \in \mathbb{R}^3$ be learnable log-variance parameters for weighting. The final objective is:

$$\begin{aligned}\min_{\Theta, s}\quad \mathbb{E}_{(x,y)\sim p_{\mathrm{data}}}\Big[&\underbrace{\|\, g_\theta(f_\phi(x)) - x \,\|_2^2}_{\mathcal{L}_{\mathrm{recon}}} + \lambda_{\mathrm{reg}}\underbrace{\mathrm{MMD}^2\big(q_\phi(z),\, p_{\mathrm{cap}}\big)}_{\mathcal{L}_{\mathrm{reg}}} \\ &+ \lambda_{\mathrm{maoc}}\quad \underbrace{\mathcal{L}_{\mathrm{MAOC}}(Z)}_{\text{Geometric Regularization}} \quad + \underbrace{\sum_{k\in\{\mathrm{CE,\,AG,\,ORM}\}} \big(e^{-s_k}\mathcal{L}_k + s_k\big)}_{\text{Adaptive Ordinal Supervision}}\Big],\end{aligned} \tag{13}$$

where $z = f_\phi(x)$, Z represents the batch of latent vectors, and the $\mathcal{L}_k$ are defined in Eqs. (8), (10), and (11). The hyperparameters λ_{reg} and λ_{maoc} balance the generative and geometric regularizers against the reconstruction and supervised losses. The model is trained using the AdamW optimizer with gradient clipping to ensure stability.

4 Experimental Results

4.1 Datasets

We conduct a comprehensive evaluation on four public DR datasets to test CAP-WAE's ordinal grading performance and robustness to real-world domain shifts. These corpora utilize two distinct clinical grading taxonomies and differ in camera hardware (Zeiss, Kowa, Topcon) and patient populations, providing a rigorous testbed for generalization. The primary taxonomy is a fine-grained 7-grade scale used by Zenodo-DR-7 [3], which includes C_0: No DR, C_1: Mild Non-Proliferative DR (NPDR), C_2: Moderate NPDR, C_3: Severe NPDR, C_4: Very Severe NPDR, C_5: Proliferative DR (PDR), and C_6: Advanced PDR. The other datasets adhere to the more common international standard 5-grade scale: 0: No DR, 1: Mild NPDR, 2: Moderate NPDR, 3: Severe NPDR, and 4: Proliferative DR (PDR). Key characteristics of each dataset are summarized in Table 1.

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

Dataset	Camera system	# Classes	Train	Test	Total
Zenodo-DR-7 [3]	Zeiss Visucam 500	7	605	152	757
IDRiD [19]	Kowa VX-10a	5	413	103	516
APTOS-2019 [2]	Heterogeneous	5	2,929	733	3,662
Messidor-2 [1]	Topcon TRC-NW6	5	1,398	350	1,748

Zenodo-DR-7. Our primary benchmark, featuring a fine-grained 7-level severity scale and an explicit long-tail distribution, making it ideal for evaluating rare-stage separability. We use the official train/test split.

IDRiD. Sourced from an Indian population, this dataset allows us to assess generalization to a different camera system and demographic. We use the official 5-grade labels and data split.

APTOS-2019. A large-scale dataset captured under varied clinical conditions, serving as a standard benchmark for model robustness against noise and acquisition variability.

Messidor-2. A widely used clinical benchmark that is crucial for evaluating cross-device generalization from the Zeiss and Kowa systems used in the other datasets.

4.2 Implementation Details

All models share one PyTorch pipeline for fairness. Images are resized to 224×224 (bicubic), normalized to ImageNet mean/std, and passed to an ImageNet-initialized VGG16 encoder; a symmetric transposed-conv decoder reconstructs $\tilde{x}$. Three lightweight heads on the latent z output (i) logits, (ii) AG-soft dispersions (σ_ℓ, σ_r), and (iii) an ordinal score. Evaluation reports Accuracy, macro-F1 and Quadratic Weighted Kappa (QWK). Hyperparameters: latent dimension 512, batch size 32, epochs 100; AdamW with learning rate 1×10^{-4} and weight decay 1×10^{-5}. All hyperparameters were selected empirically and tuned over multiple runs to balance reconstruction fidelity, latent regularization, and downstream predictive performance. The objective is the sum of pixel MSE (reconstruction), MMD (prior alignment), MAOC (geometry), and supervised terms (CE, AG-soft CE with softly clamped σ_ℓ, σ_r, Smooth-L1 ORM). Loss weights: λ_{reg}=0.1, λ_{MAOC}=0.05; if adaptive weighting is enabled, CE/AG/ORM are combined with learned log-variances, else fixed weights $\{1.0, 1.0, 0.5\}$ are used. The asymmetric prior sampler uses β=1.2 (tail-heaviness); MAOC uses margin 0.1 and compactness 0.5; AG-soft clamps $\sigma \in [0.2, 5.0]$. Optimization uses AMP, grad-norm clip 1.0, and `ReduceLROnPlateau` on validation QWK (factor 0.2, patience 7); the best QWK checkpoint is reported.

4.3 Quantitative Evaluation

The comparative study in Table 2 highlights three key observations. First, conventional imbalance-aware classifiers (CE, Focal, LDAM-DRW) improve balanced accuracy, but struggle to respect

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 [6]	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 [7]	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) [29]	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) [29]	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 [15]	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 [26]	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† [11]	–	84.61	83.19	–	83.22	67.83	–	76.79	61.47	–	61.17	46.18
GCG† [10]	0.931	90.13	88.49	–	85.29	70.57	–	80.23	73.85	–	72.14	68.34
DGN [27]	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 [27]	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**

the ordinal continuum of DR severity, leading to inconsistent macro-F1 across datasets. Ordinal-specific baselines (CORN, OLDL) mitigate label noise by leveraging rank information, achieving stronger QWK and F1, though still limited by rigid discriminative training. Second, latent generative approaches (VAE–KL, WAE–MMD) provide more stable calibration and smoother feature distributions, but their reliance on symmetric priors or loose Wasserstein alignment constrains separation in long-tailed classes. Third, attention- and discriminator-based models (e.g., GCG, AGDGN) report competitive accuracy but underperform in F1 on minority and severe classes, reflecting limited ordinal inductive bias.

CAP-WAE consistently outperforms all baselines across four benchmarks, with gains most pronounced on Zenodo-DR-7 (QWK $+3.2$ over AGDGN+OLDL) and IDRiD (macro-F1 $+2.6$ over the strongest baseline). The improvements stem from the synergy of generative alignment, asymmetric priors, and ordinal/geometry-aware supervision: Wasserstein matching yields smoother latent coverage; the asymmetric prior reduces bias toward majority grades; AG-Soft and ORM explicitly enforce ordinal calibration; and the proposed MAOC prior reshapes the latent space into severity-ordered, nearly orthogonal bands, improving both separability and robustness to class imbalance.

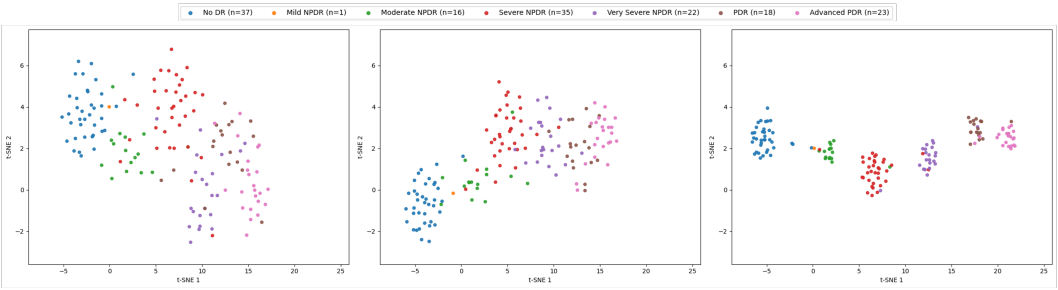
The ablation in Table 3 corroborates this progression. Adding asymmetric priors (+0.8–1.1 QWK) mitigates misclassification of rare severe grades. Wasserstein alignment and AG-Soft further improve class-level balance by pulling clusters closer to an ordinal axis. The ORM head boosts macro-F1 through structured ordinal regression, while MAOC provides the most distinct leap, translating to sharper separation and improved QWK on all datasets. Notably, the full CAP-WAE surpasses prior generative or ordinal baselines not by marginal tuning, but by systematically integrating imbalance-, ordinality-, and geometry-aware design choices. These results justify the rationale behind each component, introduced to address a concrete limitation observed in prior baselines (imbalance, lack of ordinality, poor latent geometry), and their cumulative effect produces state-of-the-art performance that is consistent across heterogeneous DR benchmarks.

4.4 Qualitative Evaluation

The t-SNE visualizations in Fig. 2 provide qualitative evidence of how the latent space evolves across model variants. The baseline VAE–KL produces overlapping clusters with no clear ordering, reflecting its difficulty in modeling ordinal transitions. Introducing Wasserstein alignment (WAE–MMD) tightens clusters and yields partial separation, but cross-class mixing remains prevalent, consistent with the moderate gains in QWK and macro-F1. In contrast, the full CAP–WAE arranges clusters into

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**



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

Figure 2: t-SNE latent-space evolution across ablation. Left: *VAE–KL* baseline shows overlapping clusters and weak ordinal structure. Middle: *WAE–MMD* improves separation but retains cross-class mixing. Right: *Full CAP–WAE (ours)* exhibits compact, well-ordered bands along disease severity with clearer orthogonal separation.

compact, nearly disjoint bands that follow the severity progression from No DR through Advanced PDR. This ordered, orthogonally separated structure directly corresponds to the effect of the MAOC prior and ORM head, which explicitly encourage severity-aware alignment in the latent space. The qualitative alignment of visual cluster geometry with ordinal disease progression validates the quantitative improvements, confirming that CAP–WAE does not merely boost classification metrics but reshapes the latent representation into a semantically meaningful space.

5 Conclusion

We introduced CAP–WAE, a generative–discriminative framework that unifies constrained asymmetric priors, adaptive ordinal supervision, and manifold-aware latent regularization to jointly address class imbalance, disease severity ordering, and clinical asymmetry in diabetic retinopathy grading. Across four benchmarks, CAP–WAE consistently outperforms prior baselines in QWK, accuracy, and macro-F1, while t-SNE analyses show a progression from entangled to well-ordered, semantically meaningful latent manifolds. Despite these advances, the approach assumes stationarity of training distributions when fixing asymmetric priors, and the manifold orthogonality constraint (MAOC) enforces local alignment without theoretical guarantees of global separability. Moreover, performance depends on reliable ordinal annotations, which in clinical practice can be noisy or inconsistent. Future work should investigate adaptive prior learning under domain shift, integrate self-supervised or multi-modal representations to reduce annotation dependence, and extend to longitudinal modeling of progression. Beyond DR, the proposed framework offers a general template for ordinal, imbalanced medical grading tasks (e.g., cancer staging, fibrosis scoring), with the potential to mitigate under-grading errors and improve triage decisions in clinical workflows.

References

[1] M. D. Abràmoff, J. C. Folk, D. P. Han, J. D. Walker, D. F. Williams, S. R. Russell, P. Massin, B. Cochener, P. Gain, L. Tang, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA ophthalmology*, 131(3):351–357, 2013.

[2] Asia Pacific Tele-Ophthalmology Society (APTOS). Aptos 2019 blindness detection. `https://www.kaggle.com/competitions/aptos2019-blindness-detection`, 2019. Kaggle competition dataset, accessed 2025-08-17.

[3] V. E. C. Benítez, I. C. Matto, J. C. M. Román, J. L. V. Noguera, M. García-Torres, J. Ayala, D. P. Pinto-Roa, P. E. Gardel-Sotomayor, J. Facon, and S. A. Grillo. Dataset from fundus images for the study of diabetic retinopathy. *Data in brief*, 36:107068, 2021.

[4] J. D. Bodapati, N. S. Shaik, and V. Naralasetti. Composite deep neural network with gated-attention mechanism for diabetic retinopathy severity classification. *Journal of Ambient Intelligence and Humanized Computing*, 12(10):9825–9839, 2021.

[5] J. D. Bodapati, N. S. Shaik, and V. Naralasetti. Deep convolution feature aggregation: an application to diabetic retinopathy severity level prediction. *Signal, Image and Video Processing*, 15(5):923–930, 2021.

[6] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. URL `https://proceedings.neurips.cc/paper/2019/hash/621461af90cadfdaf0e8d4cc25129f91-Abstract.html`.

[7] W. Cao, V. Mirjalili, and S. Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35-2, pages 1062–1070, 2021. URL `https://ojs.aaai.org/index.php/AAAI/article/view/16177`.

[8] I. Castiglioni, L. Rundo, M. Codari, G. Di Leo, C. Salvatore, M. Interlenghi, F. Gallivanone, A. Cozzi, N. C. D'Amico, and F. Sardanelli. Ai applications to medical images: From machine learning to deep learning. *Physica medica*, 83:9–24, 2021.

[9] X. Chen, X. Wang, K. Zhang, K.-M. Fung, T. C. Thai, K. Moore, R. S. Mannel, H. Liu, B. Zheng, and Y. Qiu. Recent advances and clinical applications of deep learning in medical image analysis. *Medical image analysis*, 79:102444, 2022.

[10] T. K. Cherukuri, N. S. Shaik, and D. H. Ye. Guided context gating: Learning to leverage salient lesions in retinal fundus images. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 3098–3104. IEEE, 2024.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

[12] T. Elguebaly and N. Bouguila. Finite asymmetric generalized gaussian mixture models learning for infrared object detection. *Computer Vision and Image Understanding*, 117(12):1659–1671, 2013.

[13] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

[14] A. He, T. Li, N. Li, K. Wang, and H. Fu. Cabnet: Category attention block for imbalanced diabetic retinopathy grading. *IEEE Transactions on Medical Imaging*, 40(1):143–153, 2020.

[15] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

[16] T. Li, W. Bo, C. Hu, H. Kang, H. Liu, K. Wang, and H. Fu. Applications of deep learning in fundus images: A review. *Medical Image Analysis*, 69:101971, 2021.

[17] B. Lin, D. Nie, X. Wu, X. Shen, and C. Yang. G2c-net: Grade-skewed domain adaptation network with coordinate and category attention for diabetic retinopathy grading. *Biomedical Signal Processing and Control*, 110:108203, 2025.

[18] Y. Ma, Y. Gu, S. Guo, X. Qin, S. Wen, N. Shi, W. Dai, and Y. Chen. Grade-skewed domain adaptation via asymmetric bi-classifier discrepancy minimization for diabetic retinopathy grading. *IEEE Transactions on Medical Imaging*, 44(3):1115–1129, Mar. 2025. doi: 10.1109/TMI.2024.3485064.

[19] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, and F. Meriaudeau. Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018.

[20] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning (ICML)*, pages 1278–1286, 2014.

[21] T. Sakai. On variants of root normalised order-aware divergence and a divergence based on kendall's tau. *arXiv preprint arXiv:2204.07304*, 2022.

[22] N. S. Shaik and T. K. Cherukuri. Lesion-aware attention with neural support vector machine for retinopathy diagnosis. *Machine Vision and Applications*, 32(6):126, 2021.

[23] N. S. Shaik and T. K. Cherukuri. Hinge attention network: A joint model for diabetic retinopathy severity grading. *Applied Intelligence*, 52(13):15105–15121, 2022.

[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[25] W. Tang, Z. Yang, and Y. Song. Disease-grading networks with ordinal regularization for medical imaging. *Neurocomputing*, 545:126245, 2023.

[26] I. O. Tolstikhin, O. Bousquet, S. Gelly, and B. Schölkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations (ICLR)*, 2018.

[27] V. M. Vargas, P. A. Gutiérrez, R. Rosati, L. Romeo, E. Frontoni, and C. Hervás-Martínez. Disease-grading networks with asymmetric gaussian distribution for medical imaging. *IEEE Transactions on Medical Imaging*, 2025. Accepted; early access DOI: 10.1109/TMI.2025.3575402.

[28] R. Wang, P. Li, H. Huang, C. Cao, R. He, and Z. He. Learning-to-rank meets language: Boosting language-driven ordering alignment for ordinal classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. L2RCLIP preprint/code: https://github.com/raywang335/L2RCLIP.

[29] C. Wen, X. Zhang, X. Yao, and J. Yang. Ordinal label distribution learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23481–23491, 2023.

[30] C. P. Wilkinson, F. L. Ferris III, R. E. Klein, P. P. Lee, C. D. Agardh, M. Davis, D. Dills, A. Kampik, R. Pararajasegaram, J. T. Verdaguer, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, 110(9):1677–1682, 2003.