TOPOLOGY-AWARE LATENT DIFFUSION WITH DISC-CENTERED POLAR POOLING FOR DIABETIC RETINOPATHY GRADING

Anonymous Author(s)

ABSTRACT

We address diabetic retinopathy grading from a single color fundus photograph, a five-stage ordinal task in which lesion interpretation depends on disc-centered vascular topology and quadrant extent. Current CNN/ViT systems often underuse such cues and are brittle to acquisition variability and class imbalance near referral thresholds. We propose a topology-aware latent diffusion framework operating in a compact latent space that fuses global context, lesion evidence, and a differentiable multi-scale vesselness map. A learnable disc-centered polar pooling module yields an anatomy-aware conditioning vector that summarizes vessel structure and disc and quadrant extent in line with ETDRS, and conditional latent diffusion refines task embeddings to sharpen ordinal separation. A classifier with momentum-updated class centers and joint global/local distribution alignment regularizes the latent space, and deterministic sampling conditioned on the anatomy vector stabilizes predictions at inference. By explicitly encoding vascular topology and disc- and quadrant-derived extent, the approach increases robustness to acquisition variability and better supports guideline-based referral decisions from a single fundus image. On APTOS 2019 under a stratified split and shared evaluation protocol against CNN and ViT baselines, our method improves ordinal agreement and class-aware performance.

1 INTRODUCTION

Automated analysis of retinal fundus photographs has become a cornerstone of population-scale screening in computer vision for healthcare. Diabetic retinopathy (DR) grading is a five-stage ordinal task in which severity is determined by the spatial extent of lesions in relation to the retinal vasculature. Clinical guidelines, notably ETDRS, define referral thresholds by quadrant-referenced burden and emphasize vascular structure around the optic disc for decision making Early Treatment Diabetic Retinopathy Study Research Group (1991). The concrete problem we address is to learn representations that couple lesion appearance with anatomy-aware cues—disc- and quadrant-referenced extent and vessel topology—while remaining reliable under common acquisition variability.

Deep convolutional networks and vision transformers have driven progress in DR grading by capturing global and local appearance, often augmented with attention or alignment objectives Gulshan et al. (2016); Ting et al. (2017); Dosovitskiy et al. (2021). Diffusion models further introduce generative priors that can regularize features and improve data efficiency Ho et al. (2020); Rombach et al. (2022); Yang et al. (2023). Despite these advances, existing models are predominantly texture-biased: they recognize lesion morphology but only weakly encode clinical structure such as extent relative to the optic disc and quadrants, or the branching geometry of vessels. This gap is most consequential near referral thresholds, where separating advanced non-proliferative from proliferative disease depends on lesion distribution along the vascular tree rather than appearance alone. Moreover, routine shifts in acquisition—optical blur, motion, uneven illumination and color casts, variation in field of view, and disc–fovea positioning—alter perceived contrast and disrupt quadrant-based assessment, degrading calibration and minority-grade recall. The ordinal label structure and class imbalance further make both optimization and evaluation sensitive to representation quality.

Addressing these limitations is important for safe and equitable screening. Accurate and calibrated decisions reduce missed sight-threatening disease while avoiding unnecessary referrals, a balance that hinges on sensitivity to minority grades and stability around decision boundaries. Robust generalization across devices, sites, and preprocessing pipelines is essential for deployment at scale. Anatomy-aware signals provide a principled path toward these goals: vessel topology is a stable geometric anchor under color and illumination shifts; disc- and quadrant-referenced extent encodes the clinical semantics used by experts; and embeddings that align appearance with vascular context can improve ordinal separability and calibration under acquisition variability. Bridging appearance and anatomy thus has direct impact on both performance and trustworthiness.

We propose a vessel-topology aware latent diffusion framework for five-class DR grading from color fundus photographs. The core idea is to integrate global context and lesion appearance with an explicit vesselness signal through an anatomy-aware conditioning vector that summarizes vascular topology and disc- and quadrant-referenced extent in line with ETDRS. Conditional latent diffusion then refines task embeddings so that texture cues are interpreted relative to vascular geometry, encouraging ordinally consistent representations and reducing sensitivity to routine shifts in acquisition. We evaluate the approach against CNN and ViT baselines under class imbalance, and conduct ablations that isolate the contribution of anatomy-aware conditioning and diffusion on ordinal agreement, minority-grade performance, and calibration across acquisition variability.

Our main contributions are as follows.

- We introduce a vessel-topology aware latent diffusion framework that unifies global context, lesion appearance, and explicit vascular structure for ordinal DR grading.
- We design disc- and quadrant-aware conditioning that encodes vascular topology and lesion extent in accordance with ETDRS, guiding representation learning toward clinically meaningful structure.
- We demonstrate improved ordinal agreement, minority-grade recall, and calibration robustness under routine acquisition variability compared with strong CNN and ViT baselines.
- We provide a systematic evaluation with ablations that quantify the individual effects of anatomy-aware conditioning and diffusion priors on performance and stability.

2 Related Work

2.1 Diffusion Models, Efficient Solvers, and Conditional Guidance

Diffusion probabilistic models cast generation as reverse-time denoising of a fixed noising process and link variational inference, score matching, and SDE formulations Sohl-Dickstein et al. (2015); Ho et al. (2020); Song et al. (2021). Notable developments include DDPMs with noise-prediction objectives and practical training Ho et al. (2020), continuous-time score-based modeling with probability-flow ODEs that enable stable deterministic sampling Song et al. (2021), and latent diffusion, which scales to high-resolution images by operating in a perceptual latent space Rombach et al. (2022). Inference has been sped up by non-Markovian, often deterministic trajectories (DDIM) that reduce the number of steps while approximately preserving DDPM marginals Song et al. (2020), and by specialized ODE solvers that cut function evaluations with minor quality loss (e.g., DPM-Solver) Lu et al. (2022). Conditioning has evolved from classifier guidance, which modifies the reverse dynamics using gradients of conditional log-likelihoods Dhariwal & Nichol (2021), to classifier-free guidance that mixes conditional and unconditional predictions without an auxiliary classifier Ho & Salimans (2022), and to structured-edit methods that inject spatial priors during sampling (e.g., RePaint) Lugmayr et al. (2022). These trends have shifted practice from many-step pixel-space diffusion toward more efficient, often deterministic sampling with stronger and more flexible conditioning.

We adopt DDPM-style training and use DDIM- and ODE-based accelerations for few-step, stable sampling Ho et al. (2020); Song et al. (2020); Lu et al. (2022), operate in latent space for efficiency Rombach et al. (2022), and use simple conditioning aligned to the task, inspired by classifier-free guidance, to direct denoising toward clinically relevant structures, with guidance strength tuned for deterministic solvers Dhariwal & Nichol (2021); Ho & Salimans (2022).

2.2 Medical Imaging for DR: Attention, Alignment, and Clinically Grounded Priors

Early DR screening achieved strong CNN baselines trained end to end on fundus photographs Gulshan et al. (2016); Ting et al. (2017). Attention and saliency methods (e.g., Grad-CAM) improved interpretability but often produced coarse, texture-biased explanations Selvaraju et al. (2017). Transformer-based backbones strengthened long-range context modeling Dosovitskiy et al. (2021), and distribution-alignment objectives such as MMD mitigated cross-dataset shifts Gretton et al. (2012). Clinical heuristics formalized by ETDRS emphasize vessel-level signs (e.g., venous beading, IRMA) and neovascularization near the optic disc (NVD) or elsewhere (NVE), underscoring the importance of topology- and location-aware cues for grading Early Treatment Diabetic Retinopathy Study Research Group (1991). Classical multiscale vesselness filters (e.g., Frangi) enhance tubular structures and provide anatomy-aware priors Frangi et al. (1998). Recent diffusion-enabled classifiers combine dual guidance and alignment to connect global context with local evidence (e.g., DiffMIC) Yang et al. (2023). Breakdown of the blood–retinal barrier causes edema and exudation that confound texture-based cues Cunha-Vaz (2010), and microglia-mediated neurovascular interactions influence angiogenesis and tuft dynamics in DR, motivating priors focused on vascular topology and disc-/quadrant-aware context Hu et al. (2024).

We combine attention-informed conditioning with clinically grounded priors. A differentiable multiscale vesselness channel and disc-/quadrant-aware pooling emphasize vascular topology and NVD/NVE localization within a latent diffusion pipeline, and class-aware alignment stabilizes the integration of global and local cues for cross-dataset robustness Frangi et al. (1998); Early Treatment Diabetic Retinopathy Study Research Group (1991); Gretton et al. (2012); Yang et al. (2023). This design targets fine-grained lesion geometry and neurovascular context beyond texture cues Cunha-Vaz (2010); Hu et al. (2024).

2.3 Latent Representation Refinement, Metric Learning, and Topology-Aware Constraints

Metric learning shapes latent spaces with center- and margin-based objectives to reduce intra-class variance and enlarge inter-class margins (e.g., Center Loss, ArcFace) Wen et al. (2016); Deng et al. (2019). Proxy-based variants improve efficiency and stability under multimodality (e.g., Proxy-Anchor) Kim et al. (2020). Long-tail settings common in DR benefit from logit adjustment, which incorporates label-frequency priors to recalibrate decision boundaries without heavy reweighting Menon et al. (2020). To align multi-branch or cross-domain representations, kernel MMD and deep adaptation (DAN) align feature distributions in reproducing kernel Hilbert spaces or via joint feature adaptation Gretton et al. (2012); Long et al. (2015). Topology-aware objectives aim to preserve connectivity in tubular anatomy (e.g., clDice) and penalize topological errors (TopoLoss), and topological autoencoders align latent metrics with data topology Shit et al. (2021); Clough et al. (2020); Moor et al. (2019). Recent diffusion-based classifiers use generative likelihoods and denoising for robust prediction Chen et al. (2024).

We integrate center-based regularization with class-aware MMD alignment inside a latent diffusion framework to refine representations during denoising, improving compactness and separation under distribution shift Wen et al. (2016); Gretton et al. (2012); Long et al. (2015). Rather than adding segmentation losses, we introduce topology through differentiable vesselness and anatomically structured pooling to bias features toward vascular geometry, and we apply long-tail calibration to improve minority-class recall in DR Frangi et al. (1998); Menon et al. (2020); Shit et al. (2021); Deng et al. (2019). In our design, diffusion refines latent features alongside discriminative heads, and we couple denoised features with metric and alignment constraints instead of treating the model as a standalone generative classifier Chen et al. (2024).

3 Methodology

Following ETDRS practice, where lesion burden is assessed relative to disc-centered quadrants and vascular topology, the architecture encodes vasculature and region-wise extent and then refines task-specific embeddings with conditional diffusion. We propose a vessel-topology aware latent diffusion framework for five-class diabetic retinopathy (DR) grading from a single RGB fundus image.

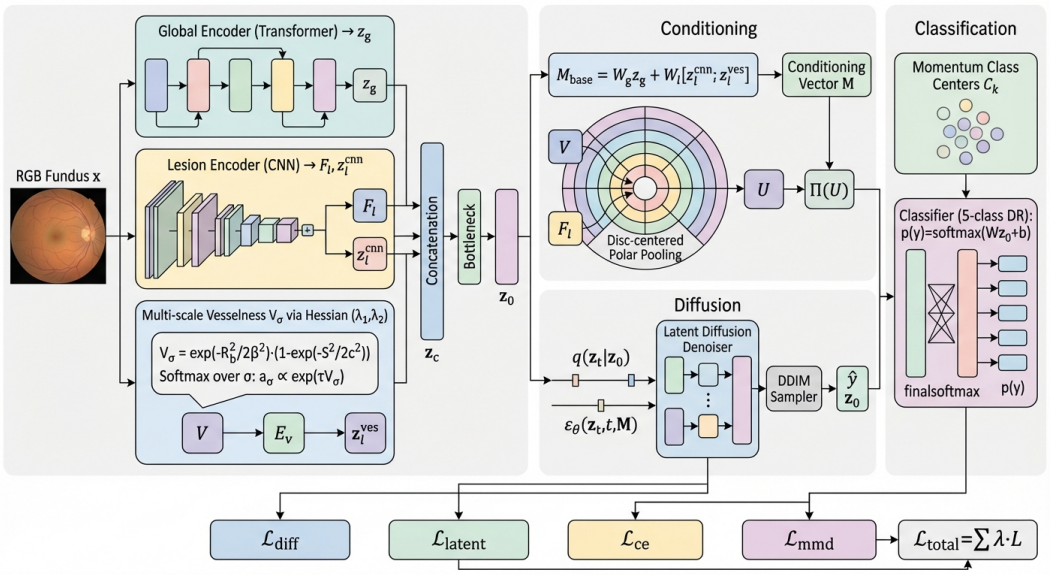

Figure 1: Architecture of the vessel-topology aware latent diffusion framework for five-class diabetic retinopathy grading from a single fundus image: a global encoder, a lesion encoder, and a multi-scale vesselness branch fuse into a compact latent; a disc-centered polar pooling module constructs an anatomical conditioning vector that captures vessel structure and quadrant extent to guide the denoiser; a center-regularized classifier produces the final grade.

The framework operates in a compact latent space and combines a topology-aware representation that fuses global context, lesion appearance, and a differentiable multi-scale vesselness channel, a latent diffusion denoiser conditioned on an anatomically based vector constructed via learnable disc-centered polar pooling, and a classification head regularized by momentum-updated class centers. The design embeds retinal vasculature and quadrant-wise extent into the conditioning pathway and refines task-specific latents through conditional diffusion to improve class separability and robustness.

3.1 Vessel-Topology Representation

We first construct a compact latent that consolidates global appearance, lesion cues, and vessel morphology. Let $x \in \mathbb{R}^{B\times 3\times H\times W}$ denote a minibatch of fundus images. A global encoder $E_g(\cdot)$ (e.g., transformer-based Dosovitskiy et al. (2021)) yields a holistic embedding $z_g = E_g(x) \in \mathbb{R}^{B\times d_g}$. A convolutional encoder $E_l(\cdot)$ produces lesion-level features; we use its penultimate feature map $F_l \in \mathbb{R}^{B\times C\times H'\times W'}$ and its pooled descriptor $z_l^{\mathrm{cnn}} \in \mathbb{R}^{B\times d_l}$. To encode vascular morphology explicitly, we construct a differentiable multi-scale vesselness channel. Let $L_\sigma = G_\sigma * x$ be a Gaussian scale-space at scale σ, and $H_\sigma(p)$ the Hessian at pixel p with ordered eigenvalues $|\lambda_{1,\sigma}(p)| \le |\lambda_{2,\sigma}(p)|$. We define the per-scale vesselness

$$
\begin{aligned}
R_b(p,\sigma) &= \frac{|\lambda_{1,\sigma}(p)|}{|\lambda_{2,\sigma}(p)| + \varepsilon}, \quad S(p,\sigma) = \sqrt{\lambda_{1,\sigma}(p)^2 + \lambda_{2,\sigma}(p)^2}, \\
V_\sigma(p) &= \exp\Big(-\frac{R_b(p,\sigma)^2}{2\beta^2}\Big)\Big(1 - \exp\Big(-\frac{S(p,\sigma)^2}{2c^2}\Big)\Big),
\end{aligned}
\tag{1}
$$

where $\beta, c > 0$ are hyperparameters and $\varepsilon > 0$ ensures numerical stability Frangi et al. (1998). To aggregate across scales, we adopt softmax pooling with temperature $\tau > 0$,

$$
a_\sigma(p) = \frac{\exp\big(\tau V_\sigma(p)\big)}{\sum_{\sigma'} \exp\big(\tau V_{\sigma'}(p)\big)}, \quad V(p) = \sum_\sigma a_\sigma(p)\, V_\sigma(p), \tag{2}
$$

yielding a differentiable vesselness map $V \in \mathbb{R}^{B\times 1\times H\times W}$. A shallow encoder $E_v(\cdot)$ maps V to $z_l^{\text{ves}} = E_v(V) \in \mathbb{R}^{B\times d_v}$. We construct the clean latent via a bottleneck projection,

$$z_c = \big[\, z_g;\, z_l^{\text{cnn}};\, z_l^{\text{ves}} \,\big] \in \mathbb{R}^{B\times(d_g+d_l+d_v)}, \quad z_0 = f_{\text{bottleneck}}(z_c) \in \mathbb{R}^{B\times d}, \tag{3}$$

which serves as the target latent for diffusion and the input to the classifier. This fused latent preserves global context while adding vessel-aware cues, providing the basis for anatomy-aware conditioning.

3.2 Anatomical Conditioning Vector

To guide denoising toward anatomically plausible and quadrant-aware embeddings, we build a dense conditioning vector $M \in \mathbb{R}^{B\times d_c}$ that summarizes global, local, and disc-centered spatial evidence. We first form a base vector from global and local descriptors,

$$M_{\text{base}} = W_g z_g + W_l\big[\, z_l^{\text{cnn}};\, z_l^{\text{ves}} \,\big], \quad W_g \in \mathbb{R}^{d_c\times d_g},\ W_l \in \mathbb{R}^{d_c\times(d_l+d_v)}. \tag{4}$$

To encode region- and extent-aware anatomy aligned with ETDRS practice, we introduce learnable disc-centered polar pooling over spatial maps. The polar origin $o = (o_x, o_y)$ is predicted by a small head $g_o(\cdot)$ acting on z_g, and approximates the optic disc center to anchor quadrants. Let $\rho(p) = \|p - o\|_2$ and $\theta(p) = \operatorname{atan2}(p_y - o_y, p_x - o_x)$ be polar coordinates for pixel p. For R concentric rings with centers μ_r and widths σ_r, and Q angular sectors with centers ϕ_q and concentration κ, we define soft memberships

$$\begin{aligned} s_r(p) &= \exp\Big(-\frac{\big(\rho(p)-\mu_r\big)^2}{2\sigma_r^2}\Big), \quad s_q(p) = \frac{\exp(\kappa\cos(\theta(p)-\phi_q))}{2\pi I_0(\kappa)}, \\ w_{r,q}(p) &= \frac{s_r(p)\, s_q(p)}{\sum_{p'} s_r(p')\, s_q(p') + \varepsilon}, \end{aligned} \tag{5}$$

where $I_0(\cdot)$ is the modified Bessel function of the first kind and $\varepsilon > 0$. We pool vesselness and lesion features within each polar cell via soft averages,

$$u_{r,q}^{(V)} = \sum_p w_{r,q}(p)\, V(p), \quad u_{r,q}^{(F)}(c) = \sum_p w_{r,q}(p)\, F_l(c,p), \tag{6}$$

and form U by concatenating all $\{u_{r,q}^{(V)}\}$ and channel-wise statistics of $\{u_{r,q}^{(F)}(c)\}$. The final conditioning vector is

$$M = M_{\text{base}} + \Pi(U), \quad \Pi : \mathbb{R}^{B\times d_U} \to \mathbb{R}^{B\times d_c}. \tag{7}$$

These soft rings and sectors capture both radial extent and quadrant-wise distribution of lesions and vessels, mirroring ETDRS disc-centered quadrants while remaining fully differentiable. To promote consistency between global and local descriptors, we regularize with a kernel two-sample objective,

$$\mathcal{L}_{\text{mmd}} = \frac{1}{n(n-1)}\sum_{i\neq j} k(z_{g,i}, z_{g,j}) + \frac{1}{n(n-1)}\sum_{i\neq j} k(z_{l,i}, z_{l,j}) - \frac{2}{n^2}\sum_{i,j} k(z_{g,i}, z_{l,j}), \tag{8}$$

where $z_l = \big[z_l^{\text{cnn}}; z_l^{\text{ves}}\big]$, $k(a,b) = \exp\big(-\|a-b\|_2^2/(2\sigma^2)\big)$, and n is the batch size Gretton et al. (2012). This alignment stabilizes the conditioning signal before it guides diffusion.

3.3 Latent Diffusion Denoiser

We adopt a latent (not pixel-space) diffusion process to refine z_0, as operating in a compact latent focuses capacity on clinically salient structure and is computationally more efficient and robust to acquisition variability than high-dimensional pixel diffusion. We adopt a T-step variance-preserving forward process in latent space Ho et al. (2020); Rombach et al. (2022). Let $\{\beta_t\}_{t=1}^T$ be a noise schedule with $\alpha_t = 1 - \beta_t$ and $\bar\alpha_t = \prod_{s=1}^t \alpha_s$. The forward diffusion is

$$\begin{aligned} q(z_t \mid z_0) &= \mathcal{N}\big(\sqrt{\bar\alpha_t}\, z_0,\, (1-\bar\alpha_t)\mathbf{I}\big), \quad t = 1,\dots,T, \\ z_t &= \sqrt{\bar\alpha_t}\, z_0 + \sqrt{1-\bar\alpha_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0,\mathbf{I}). \end{aligned} \tag{9}$$

The reverse model conditions on M via a parameterized noise predictor $\epsilon_\theta(z_t, t, M)$,

$$p_\theta(z_{t-1} \mid z_t, t, M) = \mathcal{N}(\mu_\theta(z_t, t, M),\, \sigma_t^2 \mathbf{I}),$$
$$\mu_\theta(z_t, t, M) = \frac{1}{\sqrt{\alpha_t}}\Big(z_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\,\epsilon_\theta(z_t, t, M)\Big), \tag{10}$$

with fixed variance schedule σ_t^2. The denoising objective minimizes the conditional noise prediction error,

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t, z_0, \epsilon}\Big[\big\|\epsilon - \epsilon_\theta(z_t, t, M)\big\|_2^2\Big], \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{11}$$

Conditioning on M guides the reverse trajectory toward anatomically consistent, class-separable latents that are more stable near referral thresholds.

3.4 Center-Regularized Classification

To structure the latent space for five-class DR grading, we maintain momentum-updated class centers $\{C_k\}_{k=1}^5$ with $C_k \in \mathbb{R}^d$. Given a batch with index sets $S_k = \{i : y_i = k\}$ and cardinalities $n_k = |S_k|$, the centers are updated via

$$\bar{C}_k = \frac{1}{\max(1, n_k)} \sum_{i \in S_k} z_{0,i}, \quad C_k \leftarrow \beta\, C_k + (1-\beta)\, \bar{C}_k, \quad \beta \in [0, 1). \tag{12}$$

We penalize intra-class dispersion and encourage inter-class margins,

$$\mathcal{L}_{\text{latent}} = \sum_i \big\|z_{0,i} - C_{y_i}\big\|_2^2 + \lambda_{\text{inter}} \sum_{k \neq j} \max\big(0,\, m - \|C_k - C_j\|_2\big), \tag{13}$$

and compute class probabilities with a linear classifier,

$$\ell_i = W z_{0,i} + b, \quad p(y_i = k \mid z_{0,i}) = \frac{\exp(\ell_{i,k})}{\sum_{j=1}^5 \exp(\ell_{i,j})},$$
$$\mathcal{L}_{\text{ce}} = -\sum_i \log p(y_i \mid z_{0,i}), \tag{14}$$

where $W \in \mathbb{R}^{5 \times d}$, $b \in \mathbb{R}^5$, and $m > 0$ is the margin. In practice, the momentum coefficient β controls the speed of center updates and the margin m governs separation; careful tuning and normalization of latents help stabilize center dynamics, which is particularly relevant for minority grades.

3.5 Deterministic Latent Sampling

At inference, we employ a DDIM-style deterministic sampler Song et al. (2020) conditioned on M. This avoids sampling variance inherent to stochastic reverse processes and improves run-to-run stability of predictions. With a monotone subset of steps $S = T > T_{S-1} > \cdots > T_1 \geq 1$ and initialization $z_T \sim \mathcal{N}(0, \mathbf{I})$, we iterate

$$\hat{\epsilon}_\theta = \epsilon_\theta(z_s, s, M), \quad \hat{z}_0 = \frac{z_s - \sqrt{1-\bar{\alpha}_s}\,\hat{\epsilon}_\theta}{\sqrt{\bar{\alpha}_s}},$$
$$z_{s-1} = \sqrt{\bar{\alpha}_{s-1}}\,\hat{z}_0 + \sqrt{1-\bar{\alpha}_{s-1}}\,\hat{\epsilon}_\theta, \quad s \in \{T, \ldots, 1\}, \tag{15}$$

to obtain $\hat{z}_0$, which is then fed to the classifier to produce the five-class posterior.

3.6 Optimization Objectives

The model is trained end-to-end with a weighted sum of losses,

$$\mathcal{L}_{\text{total}} = \lambda_{\text{diff}}\, \mathcal{L}_{\text{diff}} + \lambda_{\text{latent}}\, \mathcal{L}_{\text{latent}} + \lambda_{\text{ce}}\, \mathcal{L}_{\text{ce}} + \lambda_{\text{mmd}}\, \mathcal{L}_{\text{mmd}}, \tag{16}$$

where the weights $\{\lambda_{\text{diff}}, \lambda_{\text{latent}}, \lambda_{\text{ce}}, \lambda_{\text{mmd}}\}$ balance conditional denoising, latent structuring, classification, and global-local alignment. To assess whether anatomy-aware conditioning and latent diffusion translate into improved ordinal grading and robustness, the next section evaluates this framework under a unified protocol on APTOS 2019 with strong CNN/ViT baselines and ablations that selectively disable each component.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

4.1.1 DATASETS AND EVALUATION PROTOCOLS

Because DR grading is ordinal and clinical use prioritizes calibrated agreement, we evaluate on the APTOS 2019 Blindness Detection dataset, which provides single-modality RGB fundus photographs annotated with five ordinal grades of diabetic retinopathy (DR): 0 (No DR), 1 (Mild), 2 (Moderate), 3 (Severe non-proliferative DR), and 4 (Proliferative DR). Images are indexed by unique identifiers and accompanied by diagnosis labels in a standardized CSV. We fix the random seed (42) and create a stratified split of the original training set into 80% training and 20% validation subsets, preserving the label distribution in both splits. We do not use external modalities or labels.

Preprocessing includes resizing images to 224×224, standard normalization, and moderate data augmentation (horizontal flip, small rotations, and color jitter). Because the data are imbalanced, we use class-aware sampling during training via a WeightedRandomSampler with weights from empirical class frequencies in the training split, and we set class weights in the cross-entropy objective proportional to inverse class frequencies.

At each epoch we report quadratic weighted kappa (QWK) as the primary metric, which measures ordinal agreement across the five grades. We also report accuracy (ACC), macro-F1, and macro-averaged AUC computed in a multiclass one-vs-rest manner. In addition, we compute per-class precision, recall, and F1 and a confusion matrix to analyze class-specific behavior, with particular focus on grades 3 and 4, where minority effects are strongest. The same protocol is applied to all baselines and the proposed method.

4.1.2 BASELINES SETTING

To evaluate the effectiveness of the proposed method, we compare it against several widely used convolutional neural network architectures commonly adopted for medical image classification. Specifically, we consider Inception-v3, MobileNet-V3, ResNet-50, and VGG-16 as baseline models. These architectures cover a diverse spectrum of design philosophies, including lightweight networks, deep residual architectures, and classical convolutional backbones. All baselines are trained under the same experimental protocol to ensure a fair comparison. The models are trained for 50 epochs, and the checkpoint achieving the best validation accuracy is selected for evaluation. Our method is implemented under the same training setting and evaluated on the same test split as the baseline models.

4.1.3 IMPLEMENTATION DETAILS

Training runs for 50 epochs using Cosine Annealing with $T_{\max} = 50$. We optimize the encoders, fusion layers, diffusion eps-prediction model, and classification head with AdamW, and we optimize CenterLoss parameters with SGD. Automatic mixed precision and gradient clipping (max_norm $= 1.0$) are used in all experiments. The best model checkpoint is selected based on validation QWK. For each epoch, we record QWK (primary), ACC, Macro-F1, AUC, per-class metrics, and the confusion matrix. A DDIM sampler is used at inference to provide deterministic latent sampling conditioned on M.

4.2 LEARNING DYNAMICS AND CONVERGENCE

We examine learning behavior over the 50-epoch schedule to study optimization stability and the progression of ordinal calibration. Figure 2 shows the training loss and validation accuracy. During the early training stage, all evaluation metrics improve rapidly. For example, the model progresses from 0.7366 QWK and 0.5333 accuracy at epoch 5 to 0.7662 QWK and 0.5933 accuracy at epoch 10, accompanied by consistent gains in F1 score and AUC, indicating that both ordinal consistency and classification quality improve as the representation stabilizes. Performance continues to increase around epoch 15, reaching 0.7769 QWK and 0.6067 accuracy, with corresponding improvements in precision, recall, and F1 score, suggesting better class discrimination across DR severity levels. The model achieves its best performance near epoch 19–20, where QWK peaks at 0.8328, accuracy

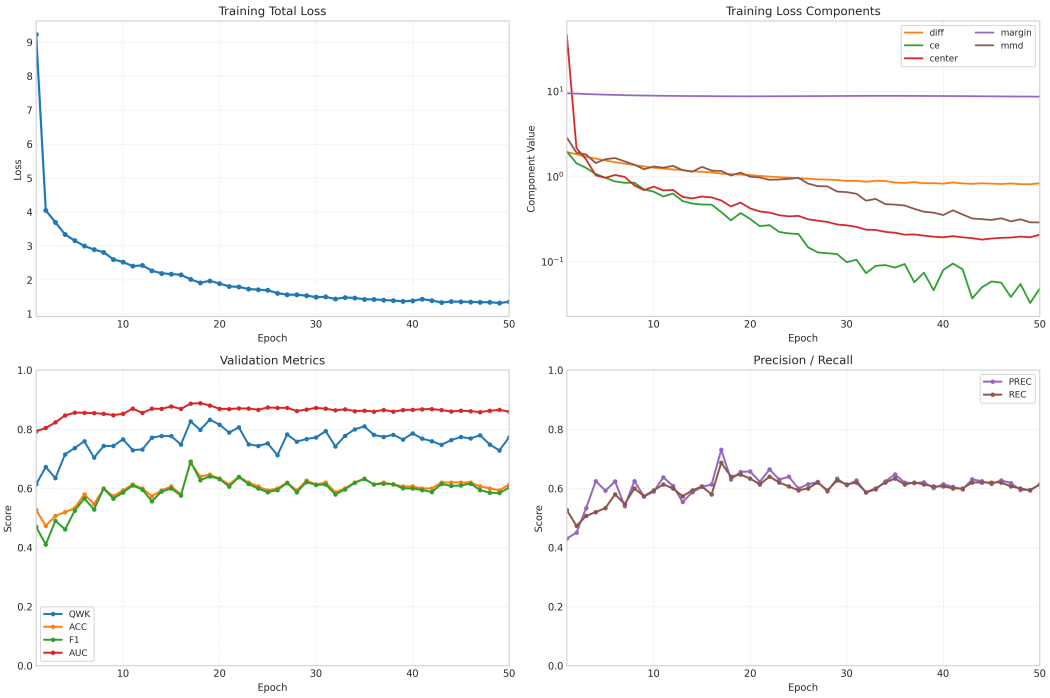

Figure 2: Learning dynamics over 50 epochs.

Table 1: Comparison with standard CNN baselines on the DR grading task. The best results are shown in **bold**.

Method	QWK ↑	ACC ↑	Precision ↑	Recall ↑	F1 ↑	AUC ↑
Inception-v3	0.7889	0.6600	0.6681	0.6600	0.6559	0.8586
MobileNet-v3	0.7766	0.6600	0.6652	0.6600	0.6587	**0.8911**
ResNet-50	0.8090	0.6467	0.6592	0.6467	0.6477	0.8656
VGG-16	0.8248	0.6600	0.6893	0.6600	0.6621	0.8589
Ours	**0.8267**	**0.6867**	**0.7306**	**0.6867**	**0.6905**	0.8867

reaches 0.6467, and both F1 score and AUC attain their highest values, demonstrating strong ordinal agreement and balanced classification performance. After this stage, the metrics fluctuate within a relatively narrow range while maintaining high values across QWK, accuracy, F1, and AUC, indicating stable convergence without noticeable overfitting. Overall, the results show that the proposed framework converges reliably within the first 20 epochs while sustaining consistent performance across multiple evaluation metrics.

4.3 Main Performance Comparison

Table 1 summarizes the quantitative comparison between the proposed method and the baseline architectures. Overall, the proposed method achieves the best performance on the primary metric (QWK), obtaining 0.8267, which surpasses all baseline models. The closest competitor is VGG-16, which achieves 0.8248, while ResNet-50, Inception-v3, and MobileNet-V3 obtain 0.8090, 0.7889, and 0.7766, respectively. The improvement over ResNet-50 and Inception-v3 indicates that the proposed design better captures the ordinal structure of diabetic retinopathy grading. In terms of classification accuracy, our method reaches 68.7 %, outperforming all baseline models that remain at 66 % or below. This improvement demonstrates that the proposed architecture yields more reliable predictions across severity levels. The proposed approach also achieves the highest precision (0.7306) and F1-score (0.6905) among all compared methods, suggesting a better balance between sensitivity and specificity. Although MobileNet-V3 obtains the highest AUC (0.8911), its QWK and

accuracy remain significantly lower, indicating that strong ranking performance does not necessarily translate into better ordinal classification. Overall, these results demonstrate that the proposed method provides more consistent improvements across multiple evaluation metrics, while achieving the best agreement with ground-truth grading according to QWK, which is the most critical metric for this task.

4.4 Visualization Comparison

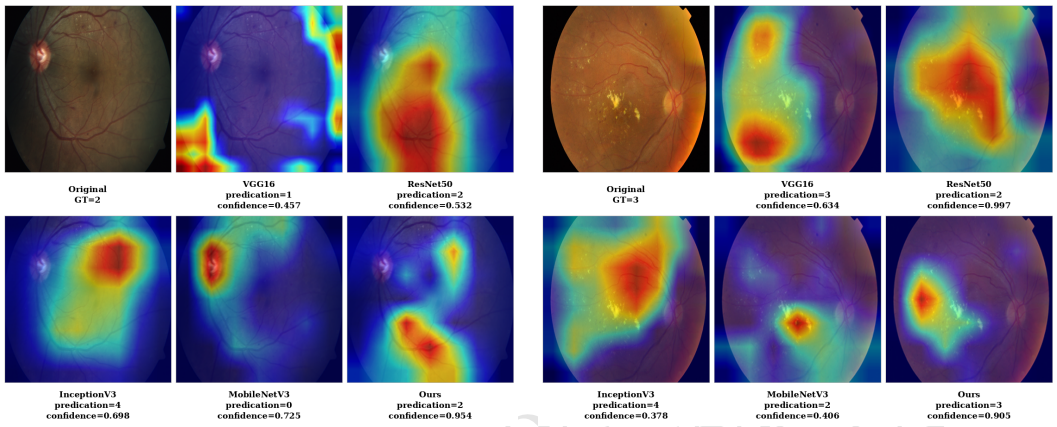

Figure 3: Visualization results from GradCAM between our method and the other baseline methods.

To better understand the decision behavior of different models, we visualize class activation maps (CAMs) for two representative retinal images, as shown in Fig. 3. In the first case (true label: moderate DR), several baseline models produce inconsistent predictions, including underestimation by VGG16 (grade 1), severe overestimation by InceptionV3 (grade 4), and a false negative prediction by MobileNetV3 (grade 0). In contrast, our method correctly predicts grade 2 with substantially higher confidence (0.95). The corresponding activation map concentrates on clinically relevant lesion regions, while competing models exhibit scattered or misplaced attention. A similar trend is observed in the second example (true label: severe DR). Although VGG16 predicts the correct grade, its activation remains relatively diffuse, whereas ResNet50 and MobileNetV3 underestimate the severity (grade 2), and InceptionV3 overestimates it (grade 4). Our model again yields the correct prediction with high confidence (0.91) and produces a more localized activation pattern aligned with pathological structures. Overall, the visualizations suggest that our model attends more consistently to lesion-related regions, leading to more reliable grading decisions compared with conventional CNN baselines.

4.5 Ablation Studies

We perform module-wise ablations to analyze the contribution of each component of the proposed vessel-topology aware latent diffusion framework. Removing the latent diffusion refinement reduces QWK from 0.8267 to 0.8103 and accuracy from 0.6867 to 0.6733, indicating that diffusion-based latent refinement improves ordinal consistency. Disabling the momentum-updated class centers leads to a larger drop to 0.7774 QWK and 0.6133 ACC, suggesting that class-center regularization is important for maintaining inter-class separability. Similarly, removing MMD alignment further decreases QWK to 0.7622, demonstrating the role of distribution alignment in stabilizing representation learning. We also analyze the topology-aware representation and conditioning mechanisms. Using only the global branch achieves 0.7806 QWK, while the local lesion/vessel branch alone reaches 0.8069, indicating that lesion-level cues provide stronger signals for DR severity estimation. Removing anatomical conditioning significantly degrades performance to 0.7569 QWK, confirming the importance of disc- and quadrant-aware conditioning. The full model combining all components consistently achieves the best performance (0.8267 QWK, 0.6867 ACC).

Table 2: Module-wise ablation of the proposed framework. TopoRep denotes the topology-aware representation, Diffusion denotes latent diffusion refinement, Center denotes momentum-updated class centers, and Cond. denotes disc- and quadrant-aware conditioning.

Variant	TopoRep	Diffusion	Center	MMD	Cond.	QWK↑	ACC↑	Prec↑	Rec↑	F1↑	AUC↑
Ours	✓	✓	✓	✓	✓	**0.8267**	**0.6867**	**0.7306**	**0.6867**	**0.6905**	**0.8867**
no_diffusion	✓	×	✓	✓	✓	0.8103	0.6733	0.6750	0.6733	0.6701	0.8676
no_center	✓	✓	×	✓	✓	0.7774	0.6133	0.6122	0.6133	0.6112	0.8495
no_mmd	✓	✓	✓	×	✓	0.7622	0.6400	0.6448	0.6400	0.6356	0.8487
global_only	×	✓	✓	✓	✓	0.7806	0.6667	0.6680	0.6667	0.6642	0.8617
local_only	×	✓	✓	✓	✓	0.8069	0.6600	0.6697	0.6600	0.6625	0.8508
M_zero	✓	✓	✓	✓	×	0.7569	0.6533	0.6655	0.6533	0.6551	0.8611
M_global_only	✓	✓	✓	✓	✓	0.7300	0.6333	0.6308	0.6333	0.6278	0.8629
M_local_only	✓	✓	✓	✓	✓	0.8090	0.6733	0.6929	0.6733	0.6767	0.8848

4.6 Analysis and Interpretation

The results show three consistent patterns. Accuracy rises early, indicating rapid learning of coarse separability, while ordinal agreement improves later and peaks. Minority-class performance remains weak despite class-aware sampling and weighted losses, as reflected by low Macro-F1; stronger lesion-aware conditioning and more stable center dynamics are expected to better recover minority signals, particularly for grades 3 and 4 where referral decisions are most sensitive. Ablation results identify the current center formulation as a bottleneck: removing it markedly improves both ordinal and class-aware metrics and enables the diffusion denoiser to contribute more effectively. These findings point to the importance of attention-guided conditioning, EMA-normalized center dynamics, and refined latent inference in capturing the ordinal structure of DR grading and turning representational quality into calibrated predictions, under a protocol that enforces parity across baselines for a fair assessment.

5 Conclusion

We address five-stage diabetic retinopathy grading from single fundus images under ETDRS, which relies on lesion burden within disc-centered anatomy. We propose a topology-aware latent diffusion model fusing global context, lesion cues, and multi-scale vesselness. Disc- and quadrant-aligned polar pooling yields anatomy-aware conditioning that guides the denoiser toward ordinal consistency. Under a common training protocol, anatomy-aware conditioning and conditional diffusion improve ordinal agreement and robustness to acquisition shifts; ablations show that removing unstable center loss unlocks diffusion gains and improves sensitivity to minority grades, while dual-branch fusion and cross-branch alignment remain necessary. Despite interpretable cues, macro-level performance remains limited and minority-grade detection near referral thresholds can degrade in a single-dataset setting. Future work will replace center loss with EMA-normalized prototypes, enrich conditioning with supervised disc localization and graph-based vessel encodings, and evaluate cross-dataset generalization via calibration and external validation.

6 Ethics Statement

This research study was conducted retrospectively using human subject data made available in open access by the APTOS 2019 Blindness Detection competition Aravind Eye Hospital and PG Institute of Ophthalmology (2019) on Kaggle, sponsored by Aravind Eye Hospital & PG Institute of Ophthalmology (India). Ethical approval was not required as confirmed by the license attached with the open access data.

References

Aravind Eye Hospital and PG Institute of Ophthalmology. APTOS 2019 Blindness Detection. https://www.kaggle.com/competitions/aptos2019-blindness-detection, 2019. Kaggle Competition.

Tianyu Chen, Zhaoyang Wang, Yixiao Li, et al. A robust diffusion classifier. *arXiv preprint arXiv:2402.17139*, 2024.

James R Clough, Ilkay Oksuz, Nick Byrne, Julia A Schnabel, and Andrew P King. A topological loss function for deep-learning based image segmentation using persistent homology. *arXiv preprint arXiv:2009.13107*, 2020.

José Cunha-Vaz. The blood-retinal barrier in retinal disease. *European Journal of Ophthalmology*, 20(suppl 6):S71–S74, 2010.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4690–4699, 2019.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Early Treatment Diabetic Retinopathy Study Research Group. Grading diabetic retinopathy from stereoscopic color fundus photographs—an extension of the modified airlie house classification: Etdrs report number 10. *Ophthalmology*, 98(5):786–806, 1991.

Alejandro Frangi, Wiro Niessen, Koen Vincken, and Max Viergever. Multiscale vessel enhancement filtering. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 130–137. Springer, 1998.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alex Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Varun Gulshan, Lily Peng, Marc Coram, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22): 2402–2410, 2016.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Wei Hu, Yao Li, and Xin Zhang. Microglia in diabetic retinopathy: Pathophysiology and therapeutic opportunities. *Cells*, 13(2):345, 2024.

Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3238–3247, 2020.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 97–105. PMLR, 2015.

Cheng Lu, Yu Zhou, Jianfei Bao, Jianmin Chen, Jun Zhu, and Wenqiang Zhao. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Andreas Lugmayr, Martin Danelljan, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11461–11471, 2022.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Michael Moor, Max Horn, Bastian Rieck, and Karsten Borgwardt. Topological autoencoders. In *International Conference on Artificial Neural Networks (ICANN)*, pp. 65–76. Springer, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

Suprosanna Shit, Alejandro Gomez, Anindo Sekuboyina, et al. cldice—a novel topology-preserving loss function for tubular structure segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 107–117. Springer, 2021.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 2256–2265. PMLR, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Daniel SW Ting, Carol Y Cheung, Gilbert Lim, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*, 318(22):2211–2223, 2017.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision (ECCV)*, pp. 499–515. Springer, 2016.

Xinyu Yang, Zhiqiang Li, Lei Zhang, et al. Diffmic: Dual-guidance diffusion model for medical image classification. *arXiv preprint arXiv:2306.00986*, 2023.