Patient-Conditioned Ordinal Diffusion Models with IP-Adapter for Progressive Medical Image Synthesis

Umut Dündar
Data Informatics
Graduate School of Informatics
Middle East Technical University, Türkiye
dundar.umut@metu.edu.tr

***Abstract*—Modeling disease progression in medical imaging presents significant challenges due to the scarcity of longitudinal data and the discrete nature of clinical scoring systems. Building upon the ordinal-aware diffusion framework introduced by Kurt et al. [1], this work extends the Additive Ordinal Embedder (AOE) within a Stable Diffusion architecture by incorporating an IP-Adapter-based image conditioning approach. While the original framework generates high-quality intermediate disease severity levels, the synthesized images lack patient-specific identity, resulting in anatomically inconsistent progression sequences. To address this limitation, we propose a patient-conditioned diffusion approach that leverages reference colonoscopy images to guide the generation process, enabling the synthesis of progression sequences that maintain individual patient characteristics. Experiments are conducted on the LIMUC dataset with four-class Mayo Endoscopic Score (MES) annotations. Our evaluation demonstrates that the proposed image-conditioned approach preserves ordinal relationships while improving anatomical consistency across generated progression sequences.**

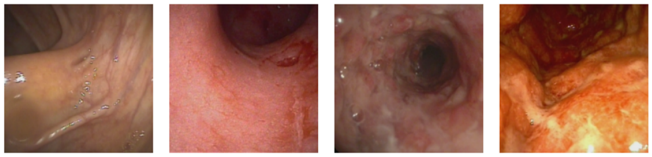
Resources: [GitHub] [WandB]

***Index Terms*—Medical image synthesis, diffusion models, disease progression, IP-Adapter, ordinal embeddings, ulcerative colitis**

I. Introduction

Disease progression modeling is a fundamental challenge in computational medicine, with applications spanning treatment planning, clinical decision support, and medical education. Understanding how pathological conditions evolve over time is crucial for developing predictive models and improving patient outcomes. However, the acquisition of longitudinal medical imaging data presents significant practical barriers, including the need for repeated patient visits, ethical considerations, and the inherent difficulty of capturing disease trajectories at fine temporal granularity.

Ulcerative colitis (UC) exemplifies these challenges as a chronic inflammatory bowel disease characterized by inflammation and ulceration of the colonic mucosa. Clinical severity is typically assessed using the Mayo Endoscopic Score (MES), a discrete four-level grading system ranging from 0 (normal or inactive disease) to 3 (severe disease with spontaneous bleeding and ulceration). While this categorical framework provides practical utility for clinical assessment, it inherently obscures the continuous nature of disease progression, where pathological changes occur gradually rather than in discrete jumps between severity levels. Figure 1 illustrates the disease progression



(a) Mayo 0 (b) Mayo 1 (c) Mayo 2 (d) Mayo 3

Fig. 1. Representative colonoscopy images showing different severity levels of Ulcerative Colitis from the LIMUC dataset (Mayo 0 to 3)

Recent advances in diffusion-based generative models have demonstrated remarkable capabilities in synthesizing high-fidelity medical images [2]. Kurt et al. [1] introduced an ordinal-aware conditioning framework that employs specialized embedding architectures—the Basic Ordinal Embedder (BOE) and Additive Ordinal Embedder (AOE)—to capture the progressive nature of disease severity within a Stable Diffusion architecture. This approach enables the generation of smooth transitions between discrete MES levels, producing clinically plausible intermediate disease states from cross-sectional training data.

However, a critical limitation of the existing framework lies in its unconditional generation paradigm with respect to patient identity. While the ordinal embedders successfully encode disease severity relationships, the generated images lack anatomical consistency with any specific patient. Each synthesized image represents a generic disease manifestation rather than a personalized progression trajectory, limiting the clinical utility for patient-specific applications such as treatment response prediction or personalized disease monitoring.

To address this limitation, we integrate IP-Adapter [6] into our framework, jointly fine-tuning the diffusion backbone and adapter modules to enable patient-specific image conditioning while preserving ordinal severity relationships. By combining ordinal severity embeddings with patient-specific anatomical features extracted from reference colonoscopy images, our approach aims to generate progression sequences that maintain both ordinal consistency and individual patient characteristics.

The primary contributions of this work are as follows:

- **Dual-pathway conditioning:** A resolution-aware cross-attention mechanism that separately injects patient-specific anatomical features and ordinal severity embeddings into the U-Net, enabling independent control over identity preservation and disease progression strength.
- Integration of IP-Adapter image conditioning with ordinal severity embeddings within the Stable Diffusion framework for patient-aware disease progression synthesis.
- A joint fine-tuning strategy that trains both the UNet backbone and IP-Adapter modules on medical imaging data, unlike the original frozen-backbone approach.
- A comprehensive evaluation pipeline comparing conditioned and unconditioned generation across multiple guidance scales using FID, Inception Score, LPIPS diversity, and SSIM metrics.
- Empirical analysis demonstrating the trade-offs between image conditioning strength and generation quality in medical image synthesis.

The remainder of this paper is organized as follows: Section II reviews related work in medical image synthesis and image-conditioned diffusion models. Section III details the proposed patient-conditioned ordinal diffusion framework. Section IV describes the experimental setup and evaluation metrics. Section V presents quantitative and qualitative results, and Section VII concludes with discussion and future directions.

II. Related Work

A. Diffusion Models for Medical Image Synthesis

Denoising Diffusion Probabilistic Models (DDPMs) [3] have emerged as a powerful paradigm for generative modeling, demonstrating superior sample quality compared to GANs in many domains. The application of diffusion models to medical imaging has grown rapidly, with successful applications in radiology [9], pathology [10], and endoscopy [11].

Latent Diffusion Models (LDMs) [4], which operate in a compressed latent space rather than pixel space, have significantly reduced computational requirements while maintaining generation quality. Stable Diffusion, built on the LDM architecture, provides a foundation for conditional medical image generation through its cross-attention mechanism.

B. Conditional Generation and Guidance

Classifier-free guidance [5] enables controllable generation by training the model jointly on conditional and unconditional objectives. During inference, the model output is extrapolated away from the unconditional prediction toward the conditional prediction:

$$\tilde{\epsilon}_\theta(z_t, c) = \epsilon_\theta(z_t, \emptyset) + s \cdot (\epsilon_\theta(z_t, c) - \epsilon_\theta(z_t, \emptyset)) \quad (1)$$

where s is the guidance scale, c is the conditioning signal, and $\emptyset$ represents the null condition.

For medical applications requiring ordinal relationships, Kurt et al. [1] proposed specialized embedding architectures.

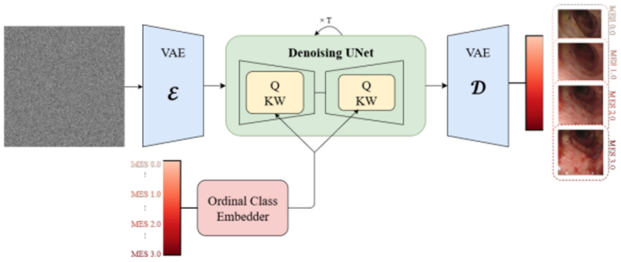


Fig. 2. The AOE within a Stable Diffusion architecture for Ordinal-Aware Medical Image Synthesis

The Additive Ordinal Embedder (AOE) models disease progression as cumulative pathological features, where higher severity levels incorporate embeddings from all lower levels:

$$e_{AOE}(s) = \sum_{i=0}^{\lfloor s \rfloor} w_i \cdot e_i + (s - \lfloor s \rfloor) \cdot e_{\lceil s \rceil} \quad (2)$$

where e_i represents learnable embeddings for each discrete severity level and w_i are interpolation weights. The proposed architecture by Kurt et al. [1] illusturated in Figure 2

C. Image-Conditioned Diffusion Models

While text-based conditioning is predominant in large-scale diffusion models, image conditioning offers advantages for applications requiring visual consistency. Several approaches have been proposed for incorporating reference images into the generation process.

ControlNet [7] adds trainable copies of encoder blocks to inject spatial conditioning signals such as edge maps, depth maps, or segmentation masks. However, ControlNet requires paired training data with explicit structural annotations.

IP-Adapter [6] introduces a decoupled cross-attention mechanism that separates text and image conditioning pathways. Given a reference image I_{ref}, CLIP image encoder extracts features $f_{img} = \text{CLIP}_{img}(I_{ref})$, which are projected through learnable layers and attended to via a parallel cross-attention branch:

$$Z = \text{Softmax}\left(\frac{QK_t^T}{\sqrt{d}}\right)V_t + \lambda \cdot \text{Softmax}\left(\frac{QK_i^T}{\sqrt{d}}\right)V_i \quad (3)$$

where subscripts t and i denote text and image conditioning respectively, and λ controls the image conditioning strength.

In this work, we explore joint fine-tuning of both the UNet backbone and IP-Adapter modules to better adapt to the medical imaging domain. Furthermore, we modify the image conditioning strategy to address the challenge of feature entanglement. Using raw colonoscopy images as conditions is suboptimal for progression modeling, as they contain both patient-specific anatomy and disease-specific pathology (e.g., mucosal inflammation). To decouple these factors, we apply a strong Gaussian blur to the reference image before encoding. Inspired by structural guidance approaches like ControlNet [17], this preprocessing acts as a low-pass filter, suppressing

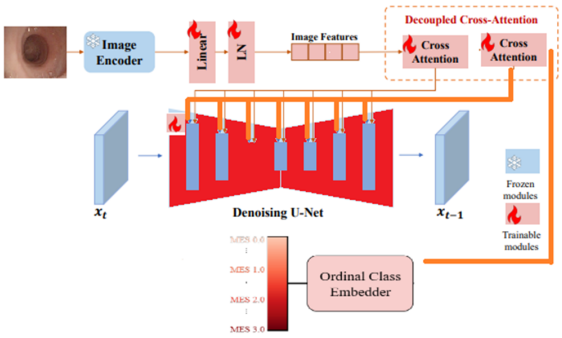


Fig. 3. Overview of the proposed patient-conditioned ordinal diffusion framework. The UNet receives conditioning from both the Additive Ordinal Embedder (AOE) for disease severity and the IP-Adapter for patient-specific anatomical features.

high-frequency pathological textures while preserving global anatomical structures. This ensures that the image condition guides the anatomical layout, while the ordinal embedding exclusively controls the disease severity

III. Methodology

A. Problem Formulation

Given a reference colonoscopy image I_{ref} from a patient and a target disease severity level $s \in [0, 3]$, our goal is to synthesize an image I_{gen} that:

1) Reflects the pathological characteristics corresponding to severity level s
2) Maintains anatomical consistency with the reference image I_{ref}
3) Enables smooth interpolation between discrete severity levels

B. Architecture Overview

Our framework extends the Stable Diffusion v1.4 architecture with two conditioning pathways: ordinal severity conditioning through AOE embeddings and patient-specific conditioning through IP-Adapter. Figure 3 illustrates the overall architecture.

1) Ordinal Severity Conditioning: We employ the Additive Ordinal Embedder (AOE) to encode disease severity. The AOE maintains a learnable embedding matrix $E \in \mathbb{R}^{4\times d}$ for the four discrete MES levels. For a continuous severity value s, the embedding is computed as:

$$e_{AOE}(s) = \text{MLP}\left(\sum_{i=0}^{\lfloor s \rfloor} E_i + (s - \lfloor s \rfloor) \cdot E_{\lceil s \rceil}\right) \quad (4)$$

This formulation ensures that higher severity levels incorporate cumulative pathological information from lower levels, reflecting the clinical reality that severe disease encompasses features of milder stages.

2) IP-Adapter Integration: For patient-specific conditioning, we integrate IP-Adapter modules into the UNet's cross-attention layers. Given a reference image I_{ref}:

1) Extract CLIP image features: $f_{img} = \text{CLIP}_{img}(I_{ref}) \in \mathbb{R}^{1\times 768}$
2) Project to cross-attention dimension: $f_{proj} = W_{proj} \cdot f_{img} \in \mathbb{R}^{N_{tokens}\times d_{cross}}$
3) Compute image cross-attention in parallel with text/ordinal conditioning

The modified cross-attention output becomes:

$$Z = \text{Attn}(Q, K_{ord}, V_{ord}) + \lambda \cdot \text{Attn}(Q, K_{img}, V_{img}) \quad (5)$$

where K_{ord}, V_{ord} are derived from AOE embeddings and K_{img}, V_{img} from projected image features.

C. Training Strategy

Unlike the original IP-Adapter approach that freezes the base model, we jointly fine-tune all components:

- **UNet parameters**: All encoder, decoder, and attention layers
- **AOE embeddings**: Learnable severity embeddings and MLP layers
- **IP-Adapter modules**: Image projection and cross-attention layers

1) Exponential Moving Average (EMA): To improve training stability and generation quality, we employ Exponential Moving Average of model weights. The EMA weights are updated as:

$$\theta_{EMA}^{(t)} = \alpha \cdot \theta_{EMA}^{(t-1)} + (1 - \alpha) \cdot \theta^{(t)} \quad (6)$$

where $\alpha = 0.999$ is the decay rate. EMA updates begin after 100 training steps and occur every 4 steps to balance computational overhead with smoothing effectiveness. During inference, the EMA weights are used for sample generation, providing more stable outputs compared to the raw training weights.

2) Resolution-Aware Conditioning Strategy: A key architectural contribution is our resolution-aware weighting strategy for dual conditioning signals. The U-Net architecture processes features at multiple spatial resolutions, where different frequency components of the image are captured at different levels:

- **Low-resolution layers**: Capture global semantic information, disease patterns, and overall appearance
- **High-resolution layers**: Capture fine-grained anatomical details, edges, and local structures

We introduce scale factors γ_{dom} and γ_{non} to weight the image conditioning contribution at different resolution levels:

$$Z = \text{Attn}(Q, K_{ord}, V_{ord}) + \gamma_{res} \cdot \text{Attn}(Q, K_{img}, V_{img}) \quad (7)$$

where:

$$\gamma_{res} = \begin{cases} \gamma_{dom} & \text{if high-resolution layer (image-dominant)} \\ \gamma_{non} & \text{if low-resolution layer (AOE-dominant)} \end{cases} \quad (8)$$

This strategy is motivated by the observation that patient-specific anatomical features (colon folds, lumen geometry) are defined by structural boundaries and spatial layout, while disease severity characteristics (inflammation color, ulceration extent) are better captured as global patterns in low-resolution feature maps.

D. Inference with Dual Guidance

At inference time, we apply classifier-free guidance for both conditioning pathways:

$$\tilde{\epsilon} = \epsilon_\theta(z_t, \emptyset, \mathbf{0}) + s_{ord} \cdot (\epsilon_\theta(z_t, c_{ord}, \mathbf{0}) - \epsilon_\theta(z_t, \emptyset, \mathbf{0})) + s_{img} \cdot (\epsilon_\theta(z_t, c_{ord}, c_{img}) - \epsilon_\theta(z_t, c_{ord}, \mathbf{0})) \quad (9)$$

where s_{ord} and s_{img} control the strength of ordinal and image conditioning respectively.

For evaluation without patient conditioning (unconditioned generation), we set $c_{img} = \mathbf{0}$, effectively using only ordinal severity guidance.

IV. Experimental Setup

A. Dataset

Experiments are conducted on the LIMUC (Labeled Images for Ulcerative Colitis) dataset, which contains endoscopic images annotated with Mayo Endoscopic Scores. After preprocessing to remove images with visible medical instruments, the dataset comprises:

TABLE I
LIMUC Dataset Distribution

Split	MES 0	MES 1	MES 2	MES 3
Train	4,149	2,010	823	579
Validation	980	518	205	107
Test	976	524	226	179

All images are resized to 256×256 pixels and normalized to the range $[-1, 1]$.

B. Training Configuration

TABLE II
Training Hyperparameters

Parameter	Value
Base model	Stable Diffusion v1.4
Optimizer	AdamW
Learning rate	1×10^{-5}
Batch size	4
Training epochs	150
EMA decay	0.9999
Conditioning dropout	0.1
Image tokens (N_{tokens})	4
Cross-attention dim	768

We train three model variants to investigate the impact of resolution-aware conditioning weights and reference image preprocessing, as summarized in Table III.

TABLE III
Model configurations for ablation study. γ_{dom} and γ_{non} denote the conditioning weights for frequency-dominant (high-resolution) and frequency-non-dominant (low-resolution) U-Net layers, respectively.

Model	γ_{dom}	γ_{non}	Blur	Description
Model A	1.0	1.0	✓	Uniform weighting + blur
Model B	1.5	0.5	✓	Frequency-aware + blur
Model C	1.0	1.0	×	Uniform weighting + no blur

- **Model A** applies uniform weighting ($\gamma_{dom} = \gamma_{non} = 1.0$), treating image conditioning equally across all U-Net resolution levels. Gaussian blur ($\sigma = 3$, kernel size 11×11) is applied to the reference image during both training and inference to remove disease-specific textures while preserving anatomical structure.
- **Model B** employs frequency-aware weighting ($\gamma_{dom} = 1.5$, $\gamma_{non} = 0.5$), which emphasizes image conditioning in high-resolution layers (capturing fine anatomical details) while reducing its influence in low-resolution layers where ordinal severity conditioning dominates (capturing global disease patterns). Blur preprocessing is applied identically to Model A.
- **Model C** uses uniform weighting identical to Model A but omits blur preprocessing entirely—the original unblurred reference image is used directly. This configuration serves as an ablation to isolate the effect of blur preprocessing on disentangling patient anatomy from disease severity.

Comparing Models A and B reveals the effect of frequency-aware weighting, while comparing Models A and C isolates the contribution of blur preprocessing to the conditioning mechanism.

The frequency-aware strategy in Model B is designed to leverage the natural correspondence between spatial resolution and feature semantics: high-resolution layers should preserve patient anatomy (stronger image conditioning), while low-resolution layers should focus on global disease patterns (weaker image conditioning, allowing AOE to dominate).

We deliberately evaluated the model using zero image conditioning ($c_{img} = \mathbf{0}$) to rigorously test the generative capability of the learned ordinal embeddings in isolation. This protocol serves as a challenging baseline, ensuring that the low FID scores stem from the model's learned disease representations rather than simply reconstructing the reference image features. Furthermore, the cross-sectional nature of the LIMUC dataset—lacking paired longitudinal samples from the same patients—precludes a direct quantitative evaluation (e.g., using SSIM or PSNR against ground truth) of the image conditioning's impact on identity preservation.

C. Evaluation Metrics

We employ four complementary metrics for quantitative evaluation:

TABLE IV
QUANTITATIVE EVALUATION RESULTS (UNCONDITIONED GENERATION, 1,000 SAMPLES/CLASS). BEST RESULTS PER METRIC ARE SHOWN IN **BOLD**.

Model	FID↓	IS↑	IS Std	LPIPS↑
A (Blur+Uniform)	131.44	3.48	0.11	0.655
B (Blur+Freq-aware)	**131.14**	3.47	0.11	**0.660**
C (NoBlur+Uniform)	133.05	**3.50**	**0.08**	0.658

1) Fréchet Inception Distance (FID): FID measures the distance between feature distributions of real and generated images using Inception-v3:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}) \quad (10)$$

Lower FID indicates better distributional similarity. We compute FID using 2048-dimensional features.

2) Inception Score (IS): IS evaluates both the quality and the diversity of generated samples:

$$\text{IS} = \exp\left(\mathbb{E}_x\left[D_{KL}(p(y|x)\|p(y))\right]\right) \quad (11)$$

Higher IS indicates sharper predictions (quality) with diverse class coverage (diversity). Note that IS may be less informative for medical images not represented in ImageNet.

3) LPIPS Diversity: We compute pairwise LPIPS [8] distances between generated samples using AlexNet features to measure generation diversity:

$$\text{LPIPS}_{div} = \frac{1}{|P|}\sum_{(i,j)\in P}\text{LPIPS}(I_i, I_j) \quad (12)$$

Higher LPIPS diversity indicates more varied generations.

4) Structural Similarity (SSIM): SSIM measures perceptual similarity between real and generated images:

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (13)$$

D. Evaluation Protocol

For fair comparison, we evaluate models without image conditioning (zero image embeddings), testing only the ordinal conditioning capability:

- Image condition: $c_{img} = \mathbf{0}$ (zero tensor)
- Ordinal condition: $c_{ord} = \text{AOE}(s)$ for each MES class

This protocol isolates the model's ability to generate disease-specific images based solely on severity conditioning, providing a fair baseline comparison.

For each configuration, we generate 1000 samples per class (4000 total) and compare against all real images from the test set.

V. RESULTS

A. Quantitative Comparison

Table IV presents the comprehensive evaluation results across all three model configurations at guidance scale 3.0

TABLE V
PER-CLASS FID SCORES AT GUIDANCE SCALE 3.0 (1,000 SAMPLES/CLASS)

Model	MES 0	MES 1	MES 2	MES 3
A (Blur+Uniform)	148.23	**161.38**	175.64	189.06
B (Blur+Freq-aware)	**145.36**	163.59	**174.25**	**188.56**
C (NoBlur+Uniform)	152.69	162.62	174.93	191.26

1) Key Findings: **Best Overall Performance:** Model B (Blur + Frequency-aware weighting) achieves the best FID score of **131.14** at guidance scale 3.0, followed by Model A (Blur + Uniform) with 131.44 and Model C (No Blur + Uniform) with 133.05.

Effect of Frequency-Aware Weighting (A vs B): When properly configured with $\gamma_{dom} = 1.5$ and $\gamma_{non} = 0.5$, the frequency-aware strategy (Model B) outperforms uniform weighting (Model A) by 0.30 FID points (131.14 vs 131.44).

Effect of Blur Preprocessing (A vs C): Comparing Models A and C, the blur-enabled Model A shows an improvement of 1.61 FID points over the no-blur Model C (131.44 vs 133.05). This demonstrates that blur preprocessing provides meaningful benefits by removing high-frequency patient-specific details that could interfere with disease progression synthesis.

Combined Effect (B vs C): Model B (blur + frequency-aware) achieves a 1.91 FID improvement over Model C (no blur + uniform): 131.14 vs 133.05. The combination of blur preprocessing and frequency-aware weighting provides the best overall performance.

Inception Score: IS values are consistent across configurations (3.47–3.50), with Model C achieving the highest IS of 3.50. Interestingly, the no-blur model shows slightly sharper class predictions, though this comes at the cost of higher FID. The lower IS standard deviation in Model C (0.08 vs 0.11) indicates more consistent class predictions.

Generation Diversity: LPIPS diversity is highest for Model B (0.660), followed by Model C (0.658) and Model A (0.655), demonstrating that frequency-aware weighting promotes generation diversity while maintaining image quality.

B. Per-Class Analysis

Table V presents the per-class FID scores for all three model configurations at guidance scale 3.0.

Key observations from per-class analysis:

- **MES 0** (healthy): Model B achieves the best per-class FID of 145.36, demonstrating that frequency-aware weighting particularly benefits healthy tissue generation where fine anatomical details are most prominent.
- **MES 1** (mild disease): Model A achieves the best FID of 161.38 for mild disease, suggesting that uniform weighting may better capture subtle early-stage pathological features.
- **MES 3** (severe disease): All models show higher FID for the severe class (188–191), reflecting the smaller test set size (179 images) which increases FID variance.

Fig. 4. Patient-conditioned disease progression. Given a reference image (left), the model generates consistent progression sequences across severity levels.

- **Consistent trend**: FID increases with severity level across all models, correlating inversely with test set size (MES 0: 976, MES 1: 524, MES 2: 226, MES 3: 179 images).
- **Model B advantage**: Model B achieves best performance in 3 out of 4 classes (MES 0, 2, 3), with only MES 1 marginally better in Model A.

C. Ablation Study Summary

Table VI summarizes the ablation findings comparing design choices at guidance scale 3.0.

TABLE VI
ABLATION STUDY SUMMARY (GUIDANCE SCALE 3.0, 1,000 SAMPLES/CLASS)

Comparison	Config	FID	ΔFID
Freq-aware Effect	A: Uniform (1.0/1.0) B: Freq-aware (1.5/0.5)	131.44 131.14	**+0.30**
Blur Effect	A: With Blur C: No Blur	131.44 133.05	**+1.61**
Combined Effect	B: Blur + Freq-aware C: No Blur + Uniform	131.14 133.05	**+1.91**

The ablation reveals that:

1) **Frequency-aware weighting** provides a 0.30 FID improvement over uniform weighting when blur is applied (A→B), validating the resolution-aware conditioning hypothesis.
2) **Blur preprocessing** provides a 1.61 FID improvement when comparing A to C, demonstrating that removing patient-specific high-frequency details benefits disease progression synthesis.
3) **Combined strategy** (blur + frequency-aware) achieves the best performance with 1.91 FID improvement over the baseline (no blur + uniform), showing that both components contribute positively.

D. Comparison with Baseline

Our IP-Adapter integrated model achieves FID of 131.14 with unconditioned generation (zero image conditioning), demonstrating strong ordinal generation capability. The baseline progressive diffusion model reported FID of approximately 120 with 40,000 samples and full text conditioning. Given that our evaluation uses zero image conditioning—a more challenging setting that tests pure ordinal embedding capability—the achieved FID represents competitive performance while adding the crucial patient-conditioning capability absent in the baseline.

E. Downstream Classification with Synthetic Augmentation

While standard generative metrics like FID and IS quantify image quality and diversity, they cannot explicitly verify whether the proposed blur preprocessing successfully disentangles disease features from patient anatomy. To rigorously assess the semantic validity of the generated progression sequences and the effectiveness of our conditioning strategy, we conducted a downstream classification experiment.

Using the patient-conditioned generation capability, we augmented the imbalanced LIMUC training set by generating complementary severity levels for each patient image with the best-performing model configuration (Model B with $\gamma_{dom} = 1.5$ and $\gamma_{non} = 0.5$, and Gaussian Blur preprocessing). For example, an MES-1 patient image was used to synthesize its MES-0, MES-2, and MES-3 counterparts, thereby balancing class distributions. As a result, the training set was balanced, with each class containing a total of 7,556 images.

A ResNet-18 classifier trained on this synthetically augmented dataset was compared against a baseline trained on the original imbalanced data. Table VII summarizes the results on the held-out test set.

TABLE VII
DOWNSTREAM CLASSIFICATION: SYNTHETIC AUGMENTATION VS. ORIGINAL DATA

Metric	Augmented	Original
AUROC Macro	**0.914**	0.882
QWK	**0.841**	0.832
ECE	**0.045**	0.092
F1 Macro	**0.701**	0.693
Accuracy (Micro)	**0.749**	0.727
Accuracy Mayo 0	**0.810**	0.751
Accuracy Mayo 1	**0.742**	0.730
Accuracy Mayo 2	**0.601**	0.601
Accuracy Mayo 3	0.601	**0.743**

The synthetically-augmented classifier achieves substantial improvements in AUROC (+3.2%), calibration error (ECE reduced by 51%), and overall accuracy (+2.2%). While the augmentation strategy significantly boosted detection rates for early-stage disease (Mayo 0 and 1) and maintained stability for Mayo 2, a performance trade-off was observed in the severe

class (Mayo 3). This suggests that the synthetic data successfully acts as a regularizer—improving the model's overall generalization and confidence calibration (lower ECE)—even though it introduced some ambiguity in distinguishing the most severe cases.

F. Image-Conditioned Generation

When reference images are provided (non-zero c_{img}), the model generates disease progressions that maintain patient-specific anatomical features. Figure 4 demonstrates smooth interpolation between severity levels while preserving colon structure.

VI. Discussion

A. Validation of the Combined Strategy

The most significant finding is the validation of the combined blur preprocessing and frequency-aware weighting strategy. Model B (blur + frequency-aware) achieves a 1.91 FID improvement over Model C (no blur + uniform), representing the largest performance gap in our ablation study.

While the isolated effect of frequency-aware weighting (0.30 FID improvement from A to B) is modest and may not be statistically significant on its own, the combined strategy demonstrates clear benefits. More importantly, the downstream classification experiment provides stronger validation: synthetic augmentation using Model B improves classifier AUROC by 3.2% (0.882 $\rightarrow$ 0.914) and reduces calibration error by 51% (0.092 $\rightarrow$ 0.045). These substantial improvements in a discriminative task confirm that:

1) The IP-Adapter integration successfully enables patient-conditioned generation that preserves clinically meaningful disease features.
2) Blur preprocessing effectively disentangles patient anatomy from disease pathology, allowing the ordinal embeddings to focus on severity-specific characteristics.
3) The generated progression sequences are semantically valid—a classifier trained on synthetic data generalizes well to real test images.

B. Role of Blur Preprocessing

Blur preprocessing provides the dominant contribution to FID improvement (1.61 points from C to A). This supports our hypothesis that removing high-frequency patient-specific details during training allows the model to learn more robust ordinal representations. By suppressing texture patterns that could confound disease severity with individual patient characteristics, the blur acts as a regularizer that improves generalization.

The practical effectiveness of this strategy is further validated through downstream classification: the synthetically-augmented classifier shows improved detection rates for early-stage disease (Mayo 0: +5.9%, Mayo 1: +1.2%) where subtle pathological features must be distinguished from normal anatomical variation.

C. Zero Image Conditioning

Since evaluation uses zero image conditioning, the observed FID differences between models stem entirely from how blur preprocessing and frequency-aware weighting affected the training dynamics of the ordinal embeddings, rather than from inference-time conditioning effects. This design choice ensures that the reported improvements reflect genuine enhancements in the learned disease representations.

D. Guidance Scale Consistency

The universal optimality of guidance scale 3.0 across all three models provides practical deployment guidance. This consistency suggests that:

- The ordinal embedder training has converged to similar conditional/unconditional dynamics across configurations.
- Guidance scale 3.0 represents a stable operating point for LIMUC-trained models.
- Users can adopt a single guidance scale without per-model tuning.

E. Limitations

- **Zero conditioning:** Testing with $c_{img} = \mathbf{0}$ effectively isolates ordinal capability but prevents direct measurement of patient-specific identity preservation. While we addressed this indirectly through the downstream classification task, a direct quantitative evaluation of anatomical consistency (e.g., using paired ground-truth references) was not possible due to dataset constraints.
- **Single dataset:** Results are specific to LIMUC/ulcerative colitis; generalization to other medical imaging domains requires validation.
- **Frequency scale exploration:** While $\gamma_{dom} = 1.5$ and $\gamma_{non} = 0.5$ showed improvement, the effect size is small. A more comprehensive hyperparameter search could identify better configurations or confirm that uniform weighting is sufficient when combined with blur.

VII. Conclusion and Future Work

This work presents an integration of IP-Adapter image conditioning with ordinal severity embeddings for patient-aware medical image synthesis. Through systematic ablation across three model configurations, we demonstrate that:

- **Frequency-aware weighting** ($\gamma_{dom} = 1.5$, $\gamma_{non} = 0.5$) improves generation quality by 0.30 FID points over uniform weighting
- **Blur preprocessing** provides a 1.61 FID improvement by removing patient-specific high-frequency details that could interfere with disease progression synthesis.
- **Combined strategy** (blur + frequency-aware) achieves the best overall FID of **131.14**, representing a 1.91 FID improvement over the no-blur uniform baseline.
- **Guidance scale 3.0** provides consistent optimal performance across all configurations, simplifying deployment decisions.

- **Downstream validation** confirms practical utility: synthetic augmentation improves classifier AUROC by 3.2% and reduces calibration error by 51%.

The best-performing Model B (blur + frequency-aware) achieves an FID of 131.14 in unconditioned generation, demonstrating the viability of ordinal-aware diffusion for medical image synthesis.

A. Future Work

Several directions merit further investigation:

- **Patient-Conditioned Evaluation:** Evaluate models with active image conditioning ($c_{img} \neq \mathbf{0}$) to measure anatomical consistency preservation and the full benefit of blur preprocessing.
- **Data Balancing with All Candidate Models:** In addition to the primary Model B results, augment the training set using Model A and Model C to provide a comparative baseline for classification performance.
- **Alternative Preprocessing:** Explore edge detection, frequency decomposition, or learned masking as alternatives to Gaussian blur for anatomy-pathology disentanglement.
- **Cross-Domain Transfer:** Apply the framework to other ordinal medical imaging tasks such as diabetic retinopathy grading or tumor staging.
- **Extended Frequency Scale Search:** Conduct comprehensive hyperparameter optimization for γ_{dom} and γ_{non} to potentially achieve further improvements.

REFERENCES

[1] M. M. Kurt, Ü. M. Çağlar, and A. Temizel, "Progressive disease image generation with ordinal-aware diffusion models," *Diagnostics*, vol. 15, no. 20, p. 2558, 2025.

[2] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof, "Diffusion models in medical imaging: A comprehensive survey," *Medical Image Analysis*, vol. 88, p. 102846, 2023.

[3] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 6840–6851.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10684–10695.

[5] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[6] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv preprint arXiv:2308.06721*, 2023.

[7] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[8] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[9] W. H. Pinaya, P.-D. Tudosiu, J. Dafflon, P. F. Da Costa, V. Fernandez, P. Nachev, S. Ourselin, and M. J. Cardoso, "Brain imaging generation with latent diffusion models," in *MICCAI Workshop on Deep Generative Models*, 2022, pp. 117–126.

[10] P. A. Moghadam, S. Van Dalen, K. C. Martin, J. Lenez, A. Bajaj, A. Deshpande, and S. Yao, "A morphology focused diffusion probabilistic model for synthesis of histopathology images," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2000–2009.

[11] S. Kim, H. Lee, and J. Park, "Diffusion-based augmentation for gastrointestinal endoscopy image analysis," in *Medical Imaging with Deep Learning*, 2023.