Disentangled Anatomy-Disease Diffusion (DADD) for Controllable Ulcerative Colitis Progression Synthesis

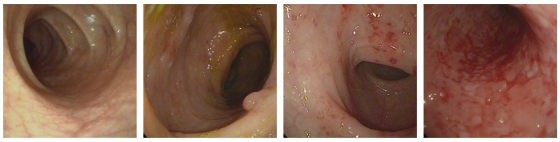
Umut Dundar, Alptekin Temizel
Graduate School of Informatics
Middle East Technical University, Turkey
dundar.umut@metu.edu.tr, atemizel@metu.edu.tr

Abstract

Synthesizing longitudinal medical images at controllable disease stages while preserving patient-specific anatomy is hindered by the entanglement of pathological textures and structural features. We address this challenge for ulcerative colitis (UC) endoscopy, where severity follows a continuous ordinal progression along the Mayo Endoscopic Score (MES). Our framework, Disentangled Anatomy-Disease Diffusion (DADD), conditions a latent diffusion model on two complementary embeddings: a pretrained image encoder for patient anatomy and a separately trained ordinal embedder for cumulative disease severity. Since image embeddings inevitably capture disease information, we introduce a Feature Purifier, a cross-attention-based erasure mechanism that identifies and suppresses disease-correlated channels, yielding purified anatomical representations. These cleaned anatomy tokens and target disease tokens are injected into the denoising network via a Triple-Pathway Cross-Attention mechanism with resolution-dependent routing gates. This architecture leverages the U-Net hierarchy, in which different network depths encode global structure versus fine-grained pathological texture. Furthermore, we introduce Delta Steering, a training-free directional signal derived from the ordinal embeddings that enables explicit, single-pass control over disease transitions at inference without requiring additional forward passes. Validated on the LIMUC dataset, our approach produces high-fidelity images across all severity levels and effectively rebalances skewed class distributions, enhancing performance for downstream classification tasks. The code is available at github.com/umutdundar99/progressive-stable-diffusion

1. Introduction

Generative models for image synthesis are advancing rapidly, with significant impact in the medical field, where data scarcity and privacy constraints often limit access to large-scale datasets. Traditionally, Generative Adversarial Networks (GANs) [6, 36] have been widely adopted to synthesize medical images and augment training sets for downstream clinical tasks. However, recent focus has shifted towards Diffusion Models [17], which offer superior training stability, better mode coverage, and higher image quality. A key advantage of modern diffusion architectures is their highly adaptable U-Net backbones, which readily integrate diverse conditioning signals (e.g., text prompts, clinical labels, or reference images [28, 35]), enabling precise, controllable generation. A compelling clinical application of these models is the synthesis of gastrointestinal disease images, such as ulcerative colitis (UC). UC severity is assessed using the Mayo Endoscopic Score (MES) [30], an ordinal scale ranging from 0 (normal) to 3 (severe) (Fig. 1). Existing datasets, such as the Labeled Images for Ulcerative Colitis (LIMUC) dataset [25], provide valuable resources but suffer from severe class imbalance and lack longitudinal follow-up due to the invasive nature of repeated endoscopies. Synthesizing subject-specific longitudinal data computationally would enable visualization of disease progression and provide targeted augmentation for downstream classification tasks.



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

Figure 1. **Sample labeled images from LIMUC [25]. (a)** normal mucosa with visible vascular pattern. **(b)** mild erythema, decreased vascular pattern. **(c)** marked erythema, friability, erosions. **(d)** spontaneous bleeding and ulceration.

Recent work effectively models continuous MES progression using an Additive Ordinal Embedder (AOE) [18]; however, because it synthesizes samples from random noise, it inherently alters patient-specific anatomy. While IP-Adapter [35] can condition generation on reference images using the Contrastive Language-Image Pretraining (CLIP)

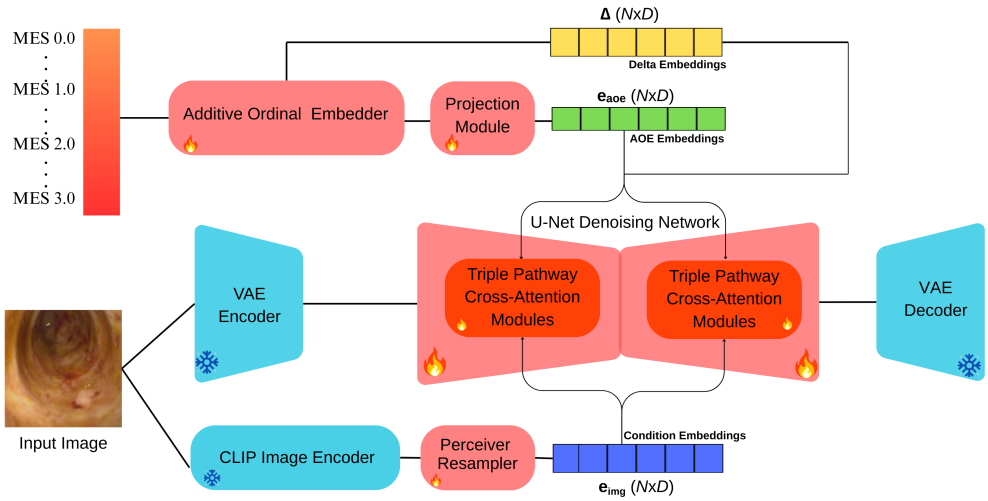


Figure 2. **Framework overview.** The input image **I** is mapped into the latent space by the frozen encoder of the Variational Autoencoder (VAE) [28]. Simultaneously, a frozen CLIP image encoder (ViT-L/14) [27] extracts feature tokens from **I**, which are then compressed into N anatomy tokens via a trainable Perceiver Resampler [15]. The AOE [18] and Projection Module encode the target MES as a cumulative ordinal embedding into N disease tokens. Delta embeddings are extracted as the post-projection of two MES scores. All three token streams are injected into the U-Net denoising network via Triple Pathway Cross-Attention (as detailed in Fig. 3). During inference, the VAE's frozen decoder maps the denoised latent representations back to pixel space to generate the final synthetic image. Snowflake and flame icons indicate frozen and trainable components, respectively.

image encoder [27], it inevitably entangles patient-specific anatomy (e.g., mucosal folds) with source disease pathology (e.g., ulceration). This creates a direct semantic conflict when attempting to alter the target severity.

We ground the development and evaluation of our DADD framework in the LIMUC dataset, the current benchmark for labeled UC endoscopic images. Our key insight is to force anatomy and disease embeddings to interact via cross-attention, explicitly erasing disease information from the image representation. By querying image tokens with the AOE, pathological channels are identified and gated out, yielding a purified anatomical representation. We inject these tokens into the U-Net via frequency-aware routing gates, thereby confining disease edits to fine-texture scales while preserving patient-specific structural identity. Finally, we introduce *delta steering*, a single-pass signal derived from the projected AOE difference that provides magnitude-proportional control over changes in severity.

Our contributions are as follows:

- Cross-attention-based disease erasure that purifies image tokens using ordinal embeddings, disentangling anatomy from pathology.
- Frequency-aware split-injection attention with fixed routing gates across U-Net layers.
- Delta steering, a single-pass alternative to Classifier-free Guidance (CFG) [13], providing signed, magnitude-proportional control of disease changes. This approach reduces inference latency by $\approx 2\times$ compared to standard CFG.
- Empirical evaluation showing improved generation fidelity and enhanced downstream classification performance via targeted augmentation.

2. Related Work

Latent diffusion models [28] have largely replaced GANs for high-fidelity synthesis [5] and are increasingly adopted in medical imaging, including chest X-rays [4] and histopathology [1]. In GI endoscopy, recent work demonstrated that ordinal embeddings in a fine-tuned stable diffusion model can successfully capture ulcerative colitis progression [18]. However, while these models generate visually plausible disease states, they struggle to preserve instance-specific anatomical structures across varying severity levels.

To preserve structural identity, a parallel line of work conditions diffusion models on reference images. While methods like ControlNet [37] and T2I-Adapter [22] enforce strict spatial alignment using dense representations such as edge maps or depth, they impose heavy constraints on the localized structural deformations needed to model severe pathologies, such as ulceration. Instead, semantic conditioning approaches such as IP-Adapter [35] inject CLIP image features [27] via decoupled cross-attention, and its Plus variant adds a Perceiver Resampler [15] for richer spatial encoding.

To tackle the entanglement between style and content within these semantic spaces, methods like DEADiff [26] and Inv-Adapter [34] employ frequency-band Q-Formers

and inverted split-injection, respectively. We adopt a similar decomposition philosophy; however, rather than separating generic artistic styles from natural images, we specifically target the disentanglement of pathological disease texture from patient anatomy in the medical domain.

Concept erasure methods employ diverse mechanisms to remove unwanted attributes: LEACE [2] utilizes fixed linear projections; recent methods such as SuMa [23] and GrOCE [8] achieve erasure through target subspace mapping and graph-guided semantic severing, respectively. Unlike these approaches, our Feature Purifier learns a *conditional* gate that dynamically adapts the degree of erasure to the disease severity present in the source image. At the attention level, methods like Prompt-to-Prompt [11], CASteer [7], and NASA [24] steer generation by intervening in the cross-attention mechanism, such as by modifying attention maps or injecting steering vectors into hidden representations. Our delta steering shares this fundamental principle but operates uniquely within an ordinal embedding space specifically designed for monotonic disease progression, offering a single-pass alternative to CFG.

Synthetic augmentation is a promising remedy for endemic class imbalance in medical datasets, traditionally explored with GANs [31] and more recently with diffusion models [33]. In contrast to generic generation, our work emphasizes preserving patient-specific anatomy while enabling controlled, monotonic edits of disease severity. This targeted severity control produces longitudinally plausible samples, directly addressing the skewed distributions and subsequently improving downstream classification performance.

3. Method

We present a conditional latent diffusion framework designed to disentangle instance-specific anatomy from disease severity, enabling controlled, monotonic progression synthesis from a single endoscopic image (Fig. 2).

3.1. Motivation: The Entanglement Bottleneck

To condition the generation process on both patient anatomy and target disease severity, a naive approach would combine an image prompt with an ordinal severity embedding. Given a source image $\mathbf{I}$ with MES $y_s \in \{0, 1, 2, 3\}$, IP-Adapter [35] extracts condition $\boldsymbol{e}_{\text{img}} = \mathcal{E}_{img}(\mathbf{I})$ via a pre-trained CLIP image encoder $\mathcal{E}_{img}$ and Perceiver Resampler. Simultaneously, the target severity y_t is encoded via AOE [18] into disease tokens $\boldsymbol{e}^{\text{t}}_{\text{aoe}} = \text{AOE}(y_t)$. Ideally, $\boldsymbol{e}_{\text{img}}$ strictly encodes anatomical identity A, while $\boldsymbol{e}^{\text{t}}_{\text{aoe}}$ dictates target pathology. In practice, CLIP embeddings entangle anatomy and source disease D_s, yielding $\boldsymbol{e}_{\text{img}} \approx A \oplus D_s$.

When attempting to generate a counterfactual progression where $y_t \neq y_s$, injecting both $\boldsymbol{e}_{\text{img}}$ and $\boldsymbol{e}^{\text{t}}_{\text{aoe}}$ into the U-Net's cross-attention layers creates a direct semantic conflict. The network receives competing pathological signals: the residual source disease D_s from the image tokens and the intended target disease D_t from the AOE. This entanglement prevents monotonic disease progression and leads to structural artifacts, as the model attempts to render two distinct severity levels simultaneously.

3.2. Proposed Architecture

To resolve this entanglement, the disease information D_s must be explicitly identified and erased from $\boldsymbol{e}_{\text{img}}$ prior to the denoising process. To this end, our proposed framework comprises three key components:

(1) **Additive Ordinal Embedder (AOE):** Encodes disease severity (y_t) as cumulative pathological tokens $\boldsymbol{e}^{\text{t}}_{\text{aoe}}$ (Sec. 3.4).

(2) **Feature Purifier:** A cross-attention-based mechanism that removes disease-correlated information (D_s) from $\boldsymbol{e}_{\text{img}}$, yielding purified anatomy tokens $\boldsymbol{e}_{\text{clean}}$ (Sec. 3.5).

(3) **Triple-Pathway Cross-Attention & Delta Steering:** A split-injection attention mechanism with resolution-aware routing gates, coupled with delta steering for inference-time severity control. (Secs. 3.6 and 3.7).

3.3. Framework Overview

The overall architecture of the proposed framework is shown in Fig. 2. To condition the generation process on both anatomy and disease severity, our architecture first combines an IP-Adapter with an AOE. To resolve the entanglement problem defined in Sec. 3.1, we integrate three novel components into this combined baseline: a Feature Purifier, Triple-Pathway Cross-Attention modules, and a delta steering mechanism.

During training, both the VAE encoder/decoder [28] and the CLIP ViT-L/14 [27] image encoder are kept frozen. Given an endoscopic image $\mathbf{I} \in \mathbb{R}^{3\times256\times256}$, we extract CLIP features and project them to N=16 tokens via a trainable Perceiver Resampler, yielding $\boldsymbol{e}_{\text{img}} \in \mathbb{R}^{N\times D}$ with D=768. These tokens carry both anatomical structure and the disease state.

Simultaneously, the target Mayo Endoscopic Score y_t is encoded by the AOE into disease tokens $\boldsymbol{e}^{\text{t}}_{\text{aoe}} \in \mathbb{R}^{N\times D}$. The complete conditioning context fed to each U-Net cross-attention layer concatenates three token groups:

$$\mathbf{C} = \left[\ \underbrace{\boldsymbol{e}^{\text{t}}_{\text{aoe}}}_{N} \parallel \underbrace{\boldsymbol{e}_{\text{clean}}}_{N} \parallel \underbrace{\boldsymbol{\Delta}}_{N}\ \right] \in \mathbb{R}^{3N\times D} \tag{1}$$

where $\boldsymbol{e}_{\text{clean}}$ denotes the purified anatomy tokens obtained via the proposed feature purifier (Sec. 3.5) and $\boldsymbol{\Delta}$ encodes the directional severity change. During training, both source and target labels are identical (y_s=y_t), so $\boldsymbol{\Delta}$=$\mathbf{0}$ and the model learns standard image reconstruction. At inference, a different target activates the delta pathway to steer disease generation (Sec. 3.7).

3.4. Additive Ordinal Embedder (AOE)

AOE mirrors the clinical observation that a higher MES level exhibits all features of lower levels plus additional pathology and represents severity as a cumulative sum of learnable increments (Eq. (2)).

$$E[k] = \boldsymbol{b} + \sum_{i=0}^{k-1} \boldsymbol{\delta}_i, \quad k \in \{0, \ldots, K-1\} \tag{2}$$

where $\boldsymbol{b} \in \mathbb{R}^D$ is a learnable base vector representing healthy mucosa and each $\boldsymbol{\delta}_i$ captures the incremental change associated with increasing severity (initialized with positive mean to encourage monotonic progression). Continuous interpolation $E(y) = (1-\alpha)\, E[\lfloor y \rfloor] + \alpha\, E[\lceil y \rceil]$ enables generation at arbitrary intermediate severity levels. The resulting embedding is projected to N tokens via a two-layer MLP with GELU activation and LayerNorm.

3.5. Feature Purifier

The CLIP image tokens $\boldsymbol{e}_{\text{img}}$ encode both the patient's anatomy and the current disease state. When the target severity differs from the source, this entanglement introduces conflicting signals: the model would simultaneously try to preserve the existing disease appearance and generate a different one. The Feature Purifier addresses this by selectively removing disease-correlated information while retaining anatomical structure (Fig. 3) in three stages. First, a cross-attention layer queries the image tokens against the source disease embedding ($\boldsymbol{e}^{\text{s}}_{\text{aoe}}$) to identify the disease component in this particular image (Eq. (3)), where LN is Layer Normalization.

$$\boldsymbol{d} = \text{CrossAttn}\big(Q{=}\text{LN}(\boldsymbol{e}_{\text{img}}),\ K{=}V{=}\text{LN}(\boldsymbol{e}^{\text{s}}_{\text{aoe}})\big) \tag{3}$$

Next, a gating MLP takes the concatenation of the detected disease component and the original tokens, producing a per-dimension mask that controls how much of each feature dimension to suppress (Eq. (4)). Finally, gated residual subtraction yields the cleaned tokens (Eq. (5)).

$$\boldsymbol{g} = \sigma\big(\text{MLP}_{2D\to D}([\boldsymbol{d}\,;\, \text{LN}(\boldsymbol{e}_{\text{img}})])\big) \in (0,1)^{N\times D} \tag{4}$$

$$\boldsymbol{e}_{\text{clean}} = \text{LN}(\boldsymbol{e}_{\text{img}} - \boldsymbol{g} \odot \boldsymbol{d}) \tag{5}$$

Since training uses $y_s{=}y_t$, the purifier learns to identify disease-correlated dimensions from the reconstruction signal alone. At inference, providing the true source label enables disease removal regardless of the target. Our current formulation assumes that the source MES label y_s is available at inference time.

3.6. Triple Pathway Cross-Attention

We replace standard U-Net cross-attention with a split-injection processor, using dedicated pathways rather than concatenating all conditioning tokens (Fig. 3). The U-Net hidden states $\mathbf{h} \in \mathbb{R}^{S\times d_h}$, where S is the spatial token count and $d_h \in \{320, 640, 1280\}$ is the block's hidden dimension for resolutions 32^2, 16^2, 8^2 respectively, provide the shared query $\mathbf{Q} = W_Q\,\mathbf{h}$. The conditioning tokens, $\boldsymbol{e}_{\text{clean}}$, $\boldsymbol{e}^{\text{t}}_{\text{aoe}}$, $\boldsymbol{\Delta} \in \mathbb{R}^{N\times D}$ with $N{=}16$, $D{=}768$, are projected into d_h dimensions by independent key-value projections. Each pathway attends separately:

$$\mathbf{z}_{\text{a}} = \text{softmax}\left(\frac{\mathbf{Q}\,(W_K\,\mathbf{e}_{\text{clean}})^\top}{\sqrt{d_h}}\right) W_V\,\mathbf{e}_{\text{clean}} \tag{6}$$

$$\mathbf{z}_{\text{d}} = \text{softmax}\left(\frac{\mathbf{Q}\,(W^{\text{d}}_K\,\mathbf{e}^{\text{t}}_{\text{aoe}})^\top}{\sqrt{d_h}}\right) W^{\text{d}}_V\,\mathbf{e}^{\text{t}}_{\text{aoe}} \tag{7}$$

The anatomy pathway applies Stable Diffusion's [28] pretrained W_K, W_V (with bias) to $\boldsymbol{e}_{\text{clean}}$, preserving structural reconstruction capacity. The disease pathway applies separate, bias-free projections $W^{\text{d}}_K, W^{\text{d}}_V$ (warm-started from pretrained weights) to $\boldsymbol{e}_{\text{aoe}}$. Crucially, bias removal ensures zero-valued delta tokens yield exactly zero output during training. At inference, the delta pathway processes $\boldsymbol{\Delta}$ through these identical bias-free projections.

Frequency-aware routing gates. The relative importance of anatomy versus disease information varies across U-Net resolution levels: low-resolution feature maps capture global structure while high-resolution maps encode fine texture. We exploit this by assigning each block a resolution-dependent gate pair $(g_{\text{a}}, g_{\text{d}})$ that weights the two pathways.

Output combination. The final attention map is obtained after combining all pathways. Fixed routing gates $(g_{\text{a}}, g_{\text{d}})$ weight the anatomy and disease contributions per block; λ_{steer} scales the delta signal as in Eq. (8).

$$\mathbf{z}^{(\ell)} = g^{(\ell)}_{\text{a}}\mathbf{z}^{(\ell)}_{\text{a}} + g^{(\ell)}_{\text{d}}\mathbf{z}^{(\ell)}_{\text{d}} + \lambda_{\text{steer}}\mathbf{z}^{(\ell)}_{\delta}, \tag{8}$$

where $\lambda_{\text{steer}}{=}0$ during training and the delta pathway activates only at inference to inject z_δ (formally defined in Sec. 3.7) and $(g^{(\ell)}_{\text{a}}, g^{(\ell)}_{\text{d}})$ denotes the fixed routing pair assigned to U-Net block ℓ.

3.7. Delta Steering

CFG is commonly used to strengthen conditioning at inference, but it requires running both conditional and unconditional passes, effectively doubling computational cost. It does not provide an inherent notion of directionality (increasing disease severity vs. decreasing it). We introduce delta steering, which operates directly at the cross-attention level to provide signed, magnitude-proportional severity control in a single forward pass.

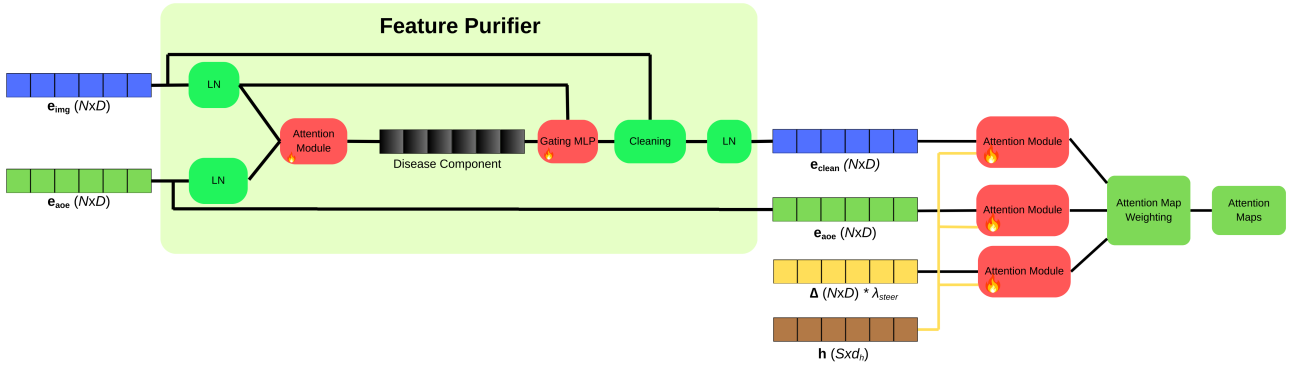


Figure 3. **Triple Pathway Cross-Attention:** All pathways share query $\mathbf{Q} = W_Q\,\mathbf{h}$ from U-Net hidden states $\mathbf{h}$. The anatomy pathway applies pretrained W_K, W_V to purified tokens $\mathbf{e}_{\text{clean}}$. The disease pathway applies separate bias-free projections $W_K^{\text{d}}, W_V^{\text{d}}$ (warm-started from W_K, W_V) to AOE tokens $\mathbf{e}_{\text{aoe}}^{\text{t}}$. The delta pathway (inference only) processes the ordinal difference Δ using these same bias-free projections, ensuring $\Delta{=}\mathbf{0}$ yields zero output during training. Outputs are combined via resolution-aware routing gates. **Feature Purifier:** Cross-attention identifies disease $\boldsymbol{d}$; a gating MLP generates a sigmoid mask $\boldsymbol{g}$; gated residual subtraction isolates anatomy tokens $\boldsymbol{e}_{\text{clean}}$.

The delta embedding is computed as a post-projection difference, where $\text{proj}(\cdot)$ denotes the two-layer MLP, so that shared biases cancel exactly (Eq. (9)).

$$\mathbf{\Delta} = \text{proj}(E[y_t]) - \text{proj}(E[y_s]) \in \mathbb{R}^{N\times D} \tag{9}$$

When $y_t > y_s$ the delta points toward more severe disease; when $y_t < y_s$ it reverses direction. The magnitude naturally scales with $|y_t{-}y_s|$, providing ordinal proportionality without manual tuning.

At inference, $\mathbf{\Delta}$ is processed through the same bias-free disease projections to produce an independent attention contribution:

$$\mathbf{z}_\delta = \text{softmax}\left(\frac{\mathbf{Q}\,(W_K^{\text{d}}\,\mathbf{\Delta})^\top}{\sqrt{d_h}}\right) W_V^{\text{d}}\,\mathbf{\Delta} \tag{10}$$

Three design choices guarantee that this pathway is exactly inactive during training: (i) post-projection subtraction makes $\mathbf{\Delta}{=}\mathbf{0}$ when $y_s{=}y_t$; (ii) bias-free projections ensure $W_K^{\text{d}}\,\mathbf{0} = \mathbf{0}$; and (iii) $\lambda_{\text{steer}}{=}0$ provides an explicit gate.

3.8. Training Objective

We train with the standard ϵ-prediction objective [14], weighted by Min-SNR-γ [9] ($\gamma{=}1$) to balance loss contributions across timesteps.

In the routing-gate variants, neither the image tokens nor the AOE tokens are dropped or zeroed: delta steering provides inference-time control without requiring an unconditional pass. In ablation configurations in which delta steering and routing gates are removed (IP-Text and IP-AOE), we fall back on CFG, which requires random image-token dropout ($p{=}0.1$) during training. The AOE negative embedding for CFG is not a null vector but a contrastive severity: healthy images (MES 0) are guided away from mild disease features, while diseased images (MES $\geq$1) are guided away from healthy mucosa, with smooth interpolation in between.

3.9. Downstream Data Augmentation

For each training image with source label y_s, we generate synthetic images for all other MES levels, effectively transforming the naturally imbalanced training set into a balanced corpus. To measure the downstream clinical utility of this augmentation, a ResNet-18 [10] classifier is trained on each augmented configuration and evaluated on a held-out test set using accuracy and Quadratic Weighted Kappa (QWK). To explicitly assess target severity fidelity, an independent ResNet-18 regression judge, trained exclusively on real LIMUC data, evaluates the synthetic outputs in isolation. This ensures the generative process genuinely transforms pathology rather than merely replicating the source.

4. Experimental Setup

Dataset. We use the LIMUC dataset, comprising 11,276 MES-labeled endoscopic images from 564 patients. Following [3], 992 images with artifacts are removed; the remaining 10,284 images are cropped to 224×224 and resized to 256×256. The patient-level split yields 7,180 train, 1,661 val, and 1,443 test images, with severe imbalance: MES 0-3 are 54.4%, 27.0%, 11.0%, and 7.6%, respectively.

Training Configuration. We fine-tune Stable Diffusion v1.4 [28] with frozen VAE and CLIP ViT-L/14; the U-Net, Perceiver Resampler, AOE, and Feature Purifier are trained jointly. We use AdamW [20] (lr=10^{-4}, 2× multiplier for Resampler and Purifier) with linear warmup (2 epochs) followed by cosine annealing to 10^{-6}. Training runs for 150 epochs with a batch size of 64 on NVIDIA RTX 6000 Ada GPUs using FP16, gradient checkpointing, and Exponential Moving Average (EMA) ($\tau{=}0.999$).

The diffusion process uses 1,000 timesteps with linear schedule ($\beta_1{=}8.5{\times}10^{-4}$, $\beta_T{=}0.012$) and Min-SNR-γ [9] ($\gamma{=}1.0$). Denoising Diffusion Implicit Models (DDIM) [32]

with 50 steps is used for inference. Class-balanced sampling mitigates dataset imbalance. The AOE uses $K{=}4$ classes, $D{=}768$, $N{=}16$ tokens, and delta init scale 0.05. Augmentation: random flips, rotation ($\pm 5°$), center crop $224{\times}224$, random perspective ($p{=}0.2$).

4.1. Experimental Design

We evaluate against two baselines:

- **IP-Adapter with CLIP text encoder (IP-Text):** IP-Adapter is used with the CLIP text encoder to generate images using severity descriptions as text prompts (e.g., *"MES k"*, $k \in \{0, 1, 2, 3\}$). For a fair comparison, the U-Net is also fine-tuned for this baseline; we observed that utilizing a frozen U-Net was insufficient to capture the specific anatomical and pathological nuances of the endoscopic domain.
- **IP-Adapter with AOE (IP-AOE)**: This baseline combines IP-Adapter with the AOE for disease conditioning. IP-Adapter injects CLIP ViT-L/14 image features into the U-Net via decoupled cross-attention, providing anatomical conditioning from a reference endoscopic image. The AOE encodes the target MES label as a cumulative ordinal embedding, which is injected alongside the image features and steered at inference via CFG.

To isolate the contribution of routing gate assignment, we compare three configurations that share identical training hyperparameters. All variants include the Feature Purifier and delta steering; they differ only in the gate weights $(g_{\mathrm{a}}, g_{\mathrm{d}})$ per U-Net resolution level.

- **Hierarchical Routing (DADD-H)**: High-resolution blocks (32^2, 16^2) emphasize disease ($g_{\mathrm{d}}{=}0.9$, $g_{\mathrm{a}}{=}0.1$); low-resolution blocks (8^2) emphasize anatomy ($g_{\mathrm{a}}{=}0.9$, $g_{\mathrm{d}}{=}0.1$). This reflects the U-Net resolution hierarchy: fine-grained layers encode inflammatory texture while coarse layers capture global mucosal structure.
- **Inverted Routing (DADD-I)**: Gate assignment reversed relative to Hierarchical Routing, where disease emphasis is on low-resolution blocks, and anatomy emphasis on high-resolution blocks.
- **Uniform Routing (DADD-U)**: All blocks use $g_{\mathrm{a}}{=}g_{\mathrm{d}}{=}0.5$.

Fig. 5 shows generated images at steering scales $\lambda_{\text{steer}} \in \{1, 3, 5, 7\}$. At low scales, disease modulation is subtle and severity levels remain difficult to distinguish; at high scales, artifacts dominate, and anatomical consistency degrades. We therefore select $\lambda_{\text{steer}}{=}3$ as the operating point for all subsequent evaluations.

4.2. Evaluation Protocol

Generation quality. Generation quality is assessed via Fréchet Inception Distance (FID) [12], CLIP Maximum Mean Discrepancy (CMMD) [16], and improved Precision and Recall [19], which measure how closely the synthetic distribution matches real data. Images are generated across all four MES classes using both LIMUC validation and testing images as anatomy conditions. This combined conditioning set is utilized to ensure a sufficiently large sample size for robust statistical estimation of the distribution-based metrics. The IP-Text and IP-AOE are steered via CFG guidance scale $w{=}3$, while the other variants use attention-level delta steering scale $\lambda_{\text{steer}}{=}3$.

Downstream classification. For each variant, we produce a balanced augmented training set (Sec. 3.9) by combining real training images with synthetic samples. A ResNet-18 classifier is trained on this augmented set for 30 epochs and evaluated on the held-out test set to report accuracy and QWK.

Regression Judge. To directly assess whether synthetic images reflect their target severity, we employ a Regression Judge: a ResNet-18 trained exclusively on real LIMUC training data to predict MES as a continuous value. The judge is evaluated solely on synthetic images generated using the held-out test set as anatomical conditions, with target severity interpolated from MES 0 to 3 in increments of 0.2. Performance is reported using QWK and Root Mean Square Error (RMSE) against these target labels, with QWK computed after clipping predictions to $[0, 3]$ and discretizing both predictions and targets to the nearest MES class. A model that genuinely transforms the appearance of disease will score highly; a model producing near-copies of the source will score near zero, regardless of FID.

Feature-Space Analysis. To verify overlap between real and synthetic distributions, we compute the global Silhouette Score [29] (using cosine distance) directly on the 512-d ResNet-18 penultimate features. UMAP [21] is utilized strictly for 2-D qualitative visualization.

5. Results

Quantitative Evaluation. As shown in Table 1, the DADD-I and DADD-U configurations achieve lower (better) FID and CMMD scores than DADD-H. However, these metrics primarily measure distributional similarity rather than the faithfulness of disease transformation. A model that produces near-copies of the conditioning image achieves low FID by construction, as its outputs inherently match the real data distribution regardless of disease fidelity. Consequently, lower FID scores for DADD-I indicate under-modulation rather than higher quality synthesis. This outcome validates our architectural premise: fine-grained layers are responsible for encoding localized inflammatory textures, whereas coarse layers capture global mucosal structure. By reversing this injection strategy, DADD-I fails to adequately alter pathological features, resulting in severely under-modulated disease representations. Crucially, as established in our scale selection (Fig. 5), attempting to compensate for this under-

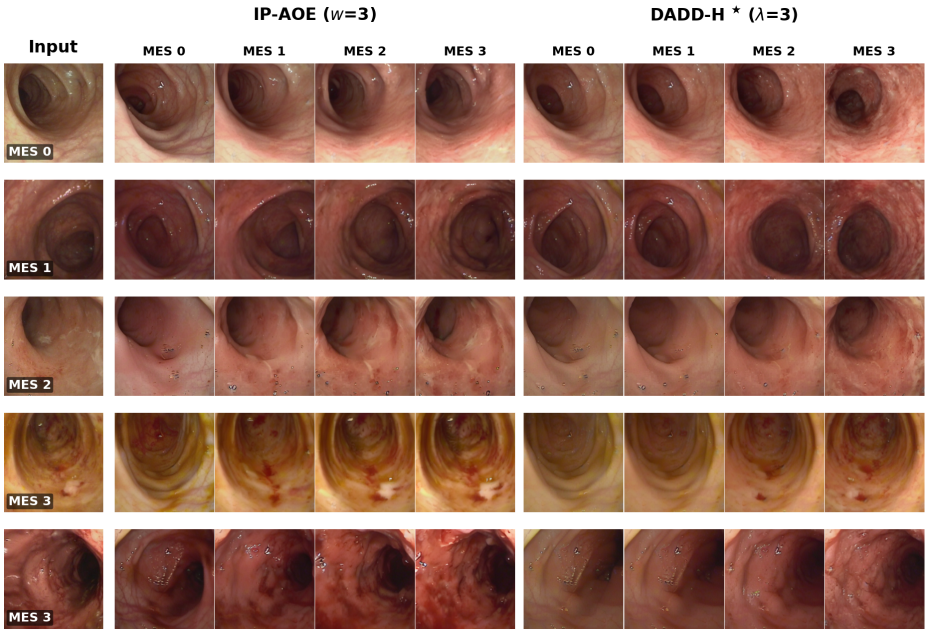


Figure 4. **Qualitative comparison: IP-AOE vs. DADD-H.** Each row shows a source image (MES 0-3, input at left) generated at four target MES levels by the IP-AOE and DADD-H.

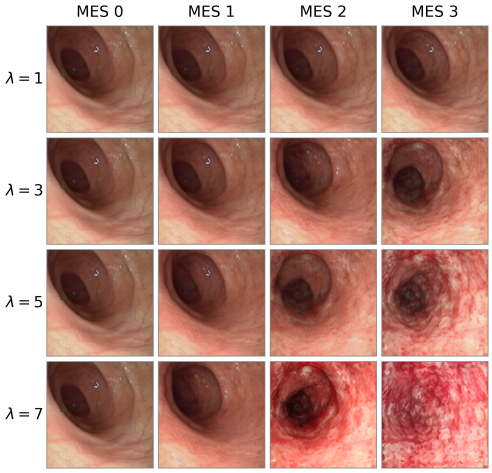


Figure 5. **Steering scale selection.** A MES 0 source image generated by DADD-H at various steering scales across target MES levels. At λ_{steer}=3, the model produces clearly distinguishable severity levels while preserving anatomical consistency. Lower scales under-modulate and higher scales introduce artifacts.

modulation by increasing the steering scale ($\lambda_{steer} > 3$) does not recover the missing pathological textures, but rather induces severe anatomical collapse, confirming the necessity of correct frequency routing. We therefore jointly interpret FID and CMMD with the downstream classification results (see Tab. 2), which directly assess whether the synthesized disease features are discriminative.

Tab. 2 details downstream classification (via class-balanced *Mixed Training* or *Synthetic Only* data) and severity fidelity (*Regression Judge*). The shaded reference row denotes baselines trained strictly on imbalanced real data using

Table 1. **Generation quality of baseline methods and DADD across configurations.** Improved Precision and Recall [19] metrics were calculated using a neighborhood size of $k = 3$ on random subsets of up to 10,000 samples per distribution.

Variant	FID↓	CMMD↓	Prec↑	Rec↑
IP-Text	35.60	.033	.736	.417
IP-AOE	25.36	.033	.606	**.584**
DADD-H	27.20	.037	.692	.463
DADD-I	**23.46**	.032	.760	.518
DADD-U	23.69	**.032**	**.768**	.521

Table 2. **Downstream Evaluation.** Classification performance (mixed vs. synthetic-only data) and regression-based severity fidelity. For the synthetic-only task, training sets are balanced by generating all MES levels per image and downsampled to the minority class size. Values for synthetic-only data denote mean ± standard deviation across 5 runs.

	Mixed Training		Synthetic Only		Regression Judge	
Variant	Acc↑	QWK↑	Acc↑	QWK↑	QWK↑	RMSE↓
Real data	74.84	83.14	—	—	83.28	0.47
IP-Text	76.10	84.46	36.42 ± 2.50	13.89 ± 8.90	15.44	1.26
IP-AOE	76.44	**84.53**	54.54 ± 1.23	42.20 ± 1.51	54.53	0.80
DADD-H	76.30	84.50	**61.68 ± 2.38**	**70.74 ± 1.80**	**81.53**	**0.46**
DADD-I	**76.51**	83.34	25.69 ± 0.44	24.55 ± 5.53	22.14	1.10
DADD-U	73.66	83.00	19.25 ± 0.70	16.77 ± 5.87	17.09	1.19

Table 3. **Inference speed under a fair protocol** on NVIDIA RTX 6000 Ada (DDIM-50, 256×256). All models are evaluated with the same precision (FP16), and timing includes *generation only* (model loading excluded).

Variant	ms/img↓ (BS=1)	ms/img↓ (BS=8)	Img/s↑ (BS=8)	Mem (GB)↓ (BS=8)
IP-Text	2766.5	351.2	2.85	**7.16**
IP-AOE	2442.0	314.6	3.18	8.02
DADD-H	**1432.6**	**187.3**	**5.34**	8.17
DADD-I	1437.9	189.5	5.28	8.17
DADD-U	1476.5	196.2	5.10	8.17

cross-entropy (CE) loss, which outperformed Focal Loss (74.15 Acc, 82.08 QWK). Most variants improve classification accuracy over the imbalanced baseline, confirming the benefit of synthetic balancing. However, the 'Synthetic Only' results provide a more rigorous test. While DADD-H enables a classifier to reach 61.68% accuracy without seeing real data, outperforming the IP-AOE baseline (54.54%), the near-random performance of DADD-I and DADD-U (<26.0%) confirms they fail to synthesize discriminative pathological features. Similarly, the failure of the IP-Text baseline (13.89 QWK) underscores that generic text embeddings lack the granularity required for monotonic progression. The Regression Judge provides the final validation of severity fidelity. DADD-H achieves 81.53 QWK / 0.46 RMSE, approaching the real-data reference (83.28 / 0.47) and confirming that synthesized images strongly align with intended MES levels. This dissociation between FID and severity fidelity confirms that distributional similarity alone is insufficient to validate conditional progression synthesis. By contrast, the IP-AOE baseline and inverted/uniform variants achieve significantly lower fidelity (54.53, 22.14, and 17.09 QWK, respectively) despite achieving lower (better) FID than DADD-H.

Inference Speed. Tab. 3 demonstrates that all DADD variants significantly outperform the baselines in generation speed. By utilizing single-pass delta steering instead of the two-branch CFG computation required by IP-Text and IP-AOE, DADD achieves roughly a 2× speedup (187.3 ms/img for DADD-H at batch size (BS)=8), trading a marginal increase in memory footprint for substantially faster inference.

Feature-Space Analysis. To quantify distribution separability, we compute the global Silhouette Score (cosine distance) on the native 512-d ResNet-18 features. Fig. 6 confirms real-synthetic overlap, supported by a low score of 0.1296, alongside a 2-D UMAP [21] visualization.

Qualitative Model Comparison. As illustrated in Fig. 4, our model yields clearly distinguishable transitions in severity while largely preserving instance-specific anatomy. Crucially, when synthesizing lower severity levels from a diseased source (e.g., MES 3 → 0), our approach successfully erases existing pathological textures. In contrast, IP-AOE fails to remove the source disease embedded in the IP-Adapter tokens, resulting in entangled generations in which healthy mucosa is incorrectly superimposed on existing inflamed structures.

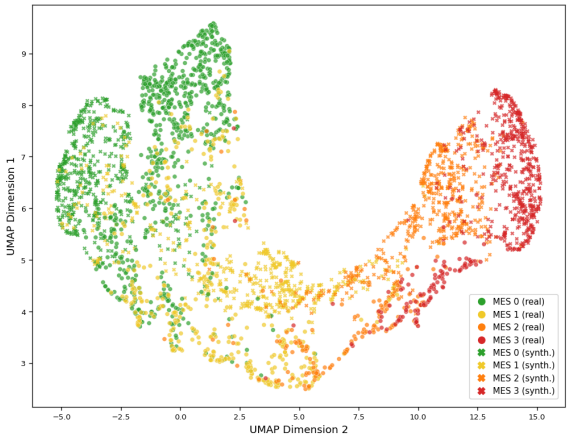


Figure 6. **Feature-space alignment.** UMAP of real (•) vs. synthetic (×) images (DADD-H, λ_{steer}=3). The global Silhouette Score of 0.1296 confirms overlap.

6. Conclusion

We presented DADD, a disentangled conditioning framework for synthesizing endoscopic disease progression. By utilizing a Feature Purifier to erase source disease from CLIP tokens and frequency-aware routing gates to direct anatomical and pathological signals to appropriate U-Net layers, our model enables a plausible visualization of that specific individual's bidirectional disease trajectory, accurately simulating both progressive inflammation and mucosal healing. We mitigated the semantic conflicts inherent in standard adapters and suppressed existing pathological textures from the condition image. Our delta steering mechanism further enables precise, single-pass severity control. Validated on the LIMUC dataset, our hierarchical model configuration (DADD-H) achieved a QWK of 81.53, closely matching the real-data ceiling of 83.28 on the regression judge task. We demonstrated that synthetic balancing consistently improves classification accuracy over imbalanced baselines, and that low FID scores can be misleading as they fail to assess faithful pathological transformation. Despite these gains, model performance remains sensitive to the steering scale λ_{steer}, where values $\lambda_{steer} \geq 5$ can introduce structural artifacts. Future research will explore per-class steering schedules to optimize transitions across severity levels and extend the DADD architecture to other longitudinal medical modalities.

Acknowledgements

Participation of Alptekin Temizel was supported by the EPAM Türkiye AI Research Fund, administered by the Graduate School of Informatics

References

[1] Marco Aversa, Gabriel Nobis, Miriam Hägele, Kai Standvoss, Mihaela Chirica, Roderick Murray-Smith, Ahmed Alaa, Lukas Ruff, Daniela Ivanova, Wojciech Samek, Frederick Klauschen, Bruno Sanguinetti, and Luis Oala. Diffinfinite: large mask-image synthesis via parallel random patch diffusion in histopathology. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2023. Curran Associates Inc. 2

[2] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *ArXiv*, abs/2306.03819, 2023. 3

[3] Umit Mert Caglar, Alperen Inci, Oguz Hanoglu, Gorkem Polat, and Alptekin Temizel. Ulcerative Colitis Mayo Endoscopic Scoring Classification with Active Learning and Generative Data Augmentation . In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 462–467, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 5

[4] Pierre Chambon, Christian Bluethgen, Jean-Benoit Delbrouck, Rogier Van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P Langlotz, and Akshay Chaudhari. Roentgen: vision-language foundation model for chest x-ray generation. *arXiv preprint arXiv:2211.12737*, 2022. 2

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[6] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018. 1

[7] Tatiana Gaintseva, Andreea-Maria Oncescu, Chengcheng Ma, Ziquan Liu, Martin Benning, Gregory Slabaugh, Jiankang Deng, and Ismail Elezi. Casteer: Cross-attention steering for controllable concept erasure. *arXiv preprint arXiv:2503.09630*, 2025. 3

[8] Ning Han, Zhenyu Ge, Feng Han, Yuhua Sun, Chengqing Li, and Jingjing Chen. Groce: Graph-guided online concept erasure for text-to-image diffusion models. *arXiv preprint arXiv:2511.12968*, 2025. 3

[9] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7441–7451, 2023. 5

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[11] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 3

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Neural Information Processing Systems*, 2017. 6

[13] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *34th International Conference on Neural Information Processing Systems*, 2020. 5

[15] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 2

[16] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9307–9315, 2024. 6

[17] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88:102846, 2023. 1

[18] Meryem Mine Kurt, Ümit Mert Çağlar, and Alptekin Temizel. Progressive disease image generation with ordinal-aware diffusion models. *Diagnostics*, 15(20), 2025. 1, 2, 3

[19] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Neural Information Processing Systems*, 2019. 6, 7

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 5

[21] Leland McInnes and John Healy. Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv*, abs/1802.03426, 2018. 6, 8

[22] Chong Mou, Xintao Wang, Liangbin Xie, Jing Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI Conference on Artificial Intelligence*, 2023. 2

[23] Kien Nguyen, Anh Tran, and Cuong Pham. Suma: A subspace mapping approach for robust and effective concept erasure in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19587–19596, 2025. 3

[24] Viet Nguyen, Anh Nguyen, Trung Dao, Khoi Nguyen, Cuong Pham, Toan Tran, and Anh Tran. Supercharged one-step text-to-image diffusion models with negative prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18004–18013, 2025. 3

[25] Gorkem Polat, Haluk Tarik Kani, Ilkay Ergenc, Yesim Ozen Alahdab, Alptekin Temizel, and Ozlen Atug. Improving the computer-aided estimation of ulcerative colitis severity according to mayo endoscopic score by using regression-

based deep learning. *Inflammatory Bowel Diseases*, 29(9): 1431–1439, 2022. 1

[26] Tianhao Qi, Shancheng Fang, Yanze Wu, Hongtao Xie, Jiawei Liu, Lang Chen, Qian He, and Yongdong Zhang. Deadiff: An efficient stylization diffusion model with disentangled representations. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8693–8702, 2024. 2

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 2, 3

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2022. 1, 2, 3, 4, 5

[29] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. 6

[30] Kenneth W. Schroeder, William J. Tremaine, and Duane M. Ilstrup. Coated oral 5-aminosalicylic acid therapy for mildly to moderately active ulcerative colitis. *New England Journal of Medicine*, 317(26):1625–1629, 1987. 1

[31] Hoo-Chang Shin, Neil A. Tenenholtz, Jameson K. Rogers, Christopher G. Schwarz, Matthew L. Senjem, Jeffrey L. Gunter, Katherine P. Andriole, and Mark Michalski. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *Simulation and Synthesis in Medical Imaging*, 2018. 3

[32] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 5

[33] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. *ArXiv*, abs/2302.07944, 2023. 3

[34] Peng Xing, Ning Wang, Jianbo Ouyang, and Zechao Li. Inv-adapter: Id customization generation via image inversion and lightweight parameter adapter. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47:9938–9952, 2025. 2

[35] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1, 2, 3

[36] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, 58:101552, 2019. 1

[37] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2