Highlights

DualDiT: A Conditional Dual-Output Diffusion Transformer for Joint OCT Image and Segmentation Mask Generation

Fernando García-Torres, Rocío del Amor, Sandra Morales, Álvaro Barroso, Peter Heiduschka, Björn Kemper, Valery Naranjo

- DualDiT generates simultaneously OCT images and anatomically aligned masks.
- A shared latent space captures dependencies between images and masks.
- DualDiT achieves the lowest FID and sFID among the evaluated models.
- DualDiT-generated data improve segmentation performance on real OCT scans.
- Experts misclassified 46% of synthetic DualDiT OCT samples as real.

arXiv:2607.29337v1 [cs.CV] 31 Jul 2026

DualDiT: A Conditional Dual-Output Diffusion Transformer for Joint OCT Image and Segmentation Mask Generation

Fernando García-Torres[a,*], Rocío del Amor[a,b], Sandra Morales[a], Álvaro Barroso[c], Peter Heiduschka[d], Björn Kemper[c] and Valery Naranjo[a,b]

[a]Instituto Universitario de Investigación en Tecnología Centrada en el Ser Humano (HUMAN-tech), Universitat Politècnica de València, Camino de Vera, s/n, Valencia, 46022, Comunidad Valenciana, Spain

[b]Artikode Intelligence S.L, Camino de Vera, s/n, Valencia, 46022, Comunidad Valenciana, Spain

[c]Biomedical Technology Center of the Medical Faculty, University of Muenster, Mendelstraße 17, Münster, 48149, North Rhine-Westphalia, Germany

[d]Department of Ophthalmology, University of Muenster Medical Centre, Domagkstraße 15, Münster, 48149, North Rhine-Westphalia, Germany

ARTICLE INFO

Keywords:
Optical coherence tomography
Diffusion Transformer
Joint image-mask generation
Synthetic data augmentation
Retinal layer segmentation
Medical image synthesis

ABSTRACT

Background and Objective: Generating realistic medical images along with anatomically accurate segmentation masks is a promising approach to address the shortage of annotated data in medical imaging, particularly in optical coherence tomography (OCT) of mouse eyes, where manual retinal layer delineation is very labour-intensive due to tiny structures and demands on expert knowledge, resulting in scarce available datasets. While diffusion models have shown strong performance in medical image synthesis, joint image-mask generation has primarily relied on U-Net-based denoisers, leaving diffusion transformers largely unexplored in this context.

Methods: We propose a conditional dual-output Diffusion Transformer (DualDiT) for the simultaneous synthesis of OCT B-scans and segmentation masks of the upper retinal cell layers (URCL) of ex vivo mouse retina. DualDiT encodes both modalities into a shared latent space via a pretrained variational auto-encoder, concatenates their latent representations, and performs conditional diffusion over the resulting joint tensor. We compared DualDiT against two adapted denoising diffusion baselines: a pixel-space conditional model (DDPM) and a latent-space conditional model (LDM). Generative quality was assessed using distribution similarity metrics (Fréchet Inception Distance, FID; spatial FID, sFID), practical utility was evaluated through synthetic data augmentation for downstream U-Net segmentation, and perceptual realism was assessed by a panel of three domain experts.

Results: DualDiT achieved the best generative quality, obtaining a FID of 56.14 and sFID of 114.35, outperforming DDPM (FID 164.55, sFID 254.52) and LDM (FID 102.21, sFID 150.66). In the expert evaluation, panels incorrectly classified, on average, 46% of synthetic samples as real and 42% of real samples as synthetic. The additional use of DualDiT-generated images and masks increases Dice and IoU scores, as evaluated on a held-out test set of the segmentation model.

Conclusions: DualDiT demonstrates that transformer-based diffusion models can effectively learn the joint distribution of OCT images and segmentation masks, surpassing conventional DDPM- and LDM-based baselines in generative fidelity, downstream segmentation utility, and perceptual realism, highlighting its potential as a data augmentation strategy for annotation-scarce medical imaging applications.

1. Introduction

Generative artificial intelligence (GenAI) is profoundly transforming the field of medical imaging. Generative models can learn the underlying distribution of complex biomedical data and generate images that are not only anatomically realistic but also include clinically and diagnostically relevant image content. Methods such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Diffusion models and Transformers have been primarily employed for data augmentation, enhancing the diversity of training datasets and improving the generalisation of diagnostic algorithms when annotated data were scarce [1–3]. Generative techniques have enabled researchers to synthesise anatomically coherent images that replicate key visual patterns across different imaging modalities (X-ray, histology, magnetic resonance, computed tomography, etc.), supporting tasks such as lesion detection and segmentation [2, 4]. These early applications established the foundation for using GenAI as a data-centric strategy to overcome the limitations of small and imbalanced datasets in healthcare [2, 3]. More recent developments are pushing this paradigm even further. Advances in diffusion models and foundation architectures have enabled high-fidelity, controllable image synthesis, where models learn general-purpose latent representations that capture the essence of complex biomedical data [5].

GANs are among the earliest and most influential approaches to generative modelling. They consist of a generator–discriminator pair trained in an adversarial setting, where the generator learns to produce realistic images while the discriminator distinguishes them from real samples [4]. In medical imaging, these models have been successfully applied to data augmentation [6], modality translation [7], denoising [8], and super-resolution [8], improving image realism and diagnostic performance [4]. VAEs introduced a

*Corresponding author. Tel.: +34 664 883 363
fergart1@upv.es (F. García-Torres)
ORCID(s): 0000-0002-5337-8774 (F. García-Torres)

probabilistic framework that encodes input data into a latent distribution and reconstructs images through sampling [9]. Variants such as β-VAE (for disentangled representations), Conditional VAE (CVAE), and hybrid VAE–GAN architectures have expanded their use in medical imaging [10]. VAEs are valuable for anomaly detection, controllable synthesis, and representation learning, offering explicit uncertainty estimation—a key property for clinical interpretation [10]. Furthermore, their structured latent space and probabilistic nature make them particularly suitable for integration into modern diffusion-based frameworks.

Diffusion models represent the current state of the art in generative medical imaging. Unlike GANs and VAEs, they avoid mode collapse and training instability through a denoising score-matching objective, and produce sharper images with finer anatomical detail [11]. Their superiority has been empirically validated by Müller-Franzes et al. [12], who demonstrated substantially lower FID scores and greater diversity than GAN-based models on eye fundus, chest X-ray, and histopathology datasets. Diffusion models progressively denoise random noise into coherent images, achieving remarkable stability and visual fidelity. Notable architectures include Denoising Diffusion Probabilistic Models (DDPMs) [13], Latent Diffusion Models (LDMs) [14], and Guided Diffusion variants for conditioning on modality, anatomy, or clinical priors. These models have achieved outstanding results in denoising, reconstruction, and cross-modality translation, often surpassing GAN-based methods [15]. Transformer-based architectures are the latest addition to generative imaging. Leveraging self-attention mechanisms, they model long-range spatial and contextual dependencies that convolutional models struggle to capture. Examples include Vision Transformers (ViT), TransGAN, and Diffusion Transformers (DiT), which combine attention modules with generative backbones to improve global coherence [5]. In medical imaging, transformers are being explored for conditional and multimodal generation, integrating imaging data with metadata or textual information such as clinical reports [5, 16]. While research in this area is still emerging, transformer-based generators are expected to play a pivotal role in the next generation of foundation models, enabling large-scale, cross-modality synthesis and clinically interpretable generative pipelines.

In medical applications, optical coherence tomography (OCT) is a widely used imaging modality in ophthalmology, providing high-resolution cross-sectional images of retina. OCT enables clinicians to visualise the layered structure of the retina and is essential for diagnosing and monitoring diseases such as age-related macular degeneration, diabetic retinopathy, and glaucoma [17]. In particular, the thickness between the embedding medium-retina interface and the interface of the inner plexiform layer (IPL) and inner nuclear layer(INL), from here on denoted as the upper retinal cell layers (URCL), has been used to study glaucoma-like changes in the retina [18]. Consequently, the automated segmentation of retinal layers from OCT scans is a critical task for quantitative disease assessment, as accurate delineation of retinal cell layers helps clinicians identify structural changes associated with early pathological processes [19, 20].

While most publicly available OCT datasets consist of human retinal images, a highly research-relevant but underexplored domain is ex vivo mouse retinal OCT imaging. Mouse models are widely used in preclinical research to simulate retinal diseases such as diabetic retinopathy and glaucoma, as their retinal architecture shares key structural similarities with the human retina [21]. Ex vivo OCT imaging offers notable advantages over in vivo imaging as it enables highly reproducible acquisitions across different OCT systems and experimental conditions [22]. In recent work, methodologies have been developed for preserving mouse retinas in resin, enabling durable ex vivo OCT imaging that maintains biological properties and enhances reproducibility across different OCT systems [23]. However, unlike human OCT data, no large-scale public databases exist for ex vivo mouse retinal OCT images, and their segmentation poses additional challenges due to tiny structures and differences in layer appearance compared with human in vivo retinas [19], differences in scale across embedding media, and limited availability of expert annotations. These particularities make manual annotation of mouse retinal layers especially time-consuming and reliant on highly specialised expert knowledge [24], further restricting the development of robust deep learning-based segmentation models. These challenges make OCT an ideal domain for generative models capable of synthesising realistic images and their corresponding segmentation masks.

In this work, we introduce a conditional dual-output Diffusion Transformer (DualDiT) framework for joint image and segmentation mask generation. This framework can synthesise OCT retinal images from ex vivo mouse retina across different preservation media (*physiological* and *resin*), along with their corresponding URCL segmentation masks. To the best of the authors' knowledge, this is the first work to formulate a DiT-based framework for paired mouse ex vivo OCT retinal image and mask synthesis, enabling the simultaneous generation of anatomically aligned images and segmentation annotations conditioned on the preservation medium domain. The main contributions of this paper are summarized as follows: (i) we introduce DualDiT, a novel DiT-based generative framework that extends diffusion transformers beyond single-output synthesis towards conditional paired image–mask generation through a unified dual-output formulation; (ii) we demonstrate its ability to synthesise domain-conditioned mouse ex vivo OCT retinal images together with anatomically consistent URCL segmentation masks; (iii) we provide a comprehensive quantitative evaluation of the generated data, including image realism assessment through the Fréchet Inception Distance (FID) and comparison against state-of-the-art diffusion-based generative frameworks; and (iv) we validate the practical and anatomical relevance of the generated image–mask pairs by showing their effectiveness as a data augmentation strategy in a downstream

segmentation task, with improvements in Dice coefficient and Intersection over Union (IoU), and by conducting a systematic expert-based assessment of anatomical fidelity and perceived reliability of clinical and diagnostic relevant content.

2. Related Work

2.1. Diffusion Models for OCT image synthesis and processing

Diffusion models were introduced as Denoising Diffusion Probabilistic Models (DDPM) by Ho et al. in 2020 [13]. The application of diffusion models to OCT has gained significant momentum, addressing inherent challenges such as speckle noise, scarcity of high-quality data, and the need for accurate structural synthesis. For example, OCTDiff [25] employs a bridged diffusion model to facilitate super-resolution in portable OCT devices, effectively reducing the quality gap with clinical-grade systems. This is complemented by physics-informed diffusion models [26], which incorporate the underlying optics of OCT acquisition into the generative process to achieve high-fidelity reconstruction. Beyond posterior segment imaging, diffusion models have also been successfully applied to remove noise and generate anterior segment (AS-OCT) images [27].

The primary motivation for synthetic OCT generation is often to improve subsequent clinical tasks. RetiDiff [28] and other DDPM-based frameworks [29] have demonstrated that generating synthetic scans with corresponding layer maps can significantly improve segmentation performance. This synthesis for segmentation paradigm has been extended to specific biomarkers. For example, recent benchmark tests show that diffusion models outperform traditional architectures in detecting fluid-filled regions for retinal analysis [30].

Although DDPMs achieve high-quality synthesis, operating directly in pixel space is computationally expensive due to the high dimensionality of images. To handle the high dimensionality of medical data without prohibitive costs, several studies have adopted Latent Diffusion Models (LDMs). Cascaded amortised LDMs [31] have been proposed for efficient in-memory synthesis of high-resolution human retinal OCT volumes. The versatility of latent space also enables multimodal tasks, such as translation from standard OCT to optical coherence tomography angiography (OCT-A) [32]. These frameworks are particularly effective in diagnosing neovascularisation, as they provide high-resolution, biologically plausible synthetic OCT-A scans.

Finally, unified frameworks such as DiffusionDCI [33] represent the state of the art in dynamic OCT imaging, offering a single model for both generation and segmentation, suggesting a shift towards more integrated generative-discriminative architectures in ophthalmology. In contrast to these earlier works, which focus exclusively on human OCT data, our proposed DualDiT framework addresses the under-explored setting of mouse ex vivo OCT imaging, where no large-scale public datasets exist. Furthermore, unlike in vivo acquisition, ex vivo imaging requires tissue embedding, and domain differences across embedding media introduce additional challenges for image synthesis that DualDiT explicitly addresses.

2.2. Diffusion models for joint image–mask generation

While diffusion models were originally designed to generate a single image, they have recently been extended to structured output generation, particularly for joint image and segmentation mask synthesis. This setting is especially relevant in domains where annotated data is scarce or costly to obtain, such as medical imaging. Instead of generating images alone, these approaches aim to model the joint distribution $p_\theta(x, m)$, where x denotes the image and m its associated pixel-level label map.

Several recent works explicitly address simultaneous image-mask generation. In the field of satellite image processing, Toker et al. in [34] leverage DDPM with conditional super-resolution to augment aerial image-mask pairs for semantic segmentation. Mao et al. in [35] extend this idea by incorporating text guidance to generate paired medical images and masks, enabling more flexible, controllable data synthesis via a dual-stream diffusion with cross-attention between image and mask streams. Similarly, Frisch et al. [36] focus on uncertainty-guided diffusion-based augmentation for surgical segmentation, training an LDM with learned Vector Quatized-GAN-based latents. Other works explore pixel-level annotation synthesis in different contexts. Wu et al. generate images with an LDM and derive semantic masks from the model's cross-attention maps [37]. They exploit the cross-attention maps between text and image in the diffusion model to automatically generate high-resolution, class-specific semantic masks, enabling the training of segmentation models with synthetic data that performs comparably to real data. Li et al. [38] explore segmentation conditioned on flexible semantic prompts with LDM, bridging generative modelling and open-vocabulary recognition.

Complementary approaches investigate segmentation-aware diffusion from different perspectives. Park et al. in [39] demonstrate that off-the-shelf Stable Diffusion models [14] can be adapted for seeded mask generation, highlighting the implicit segmentation capabilities encoded within pretrained diffusion backbones. Collectively, these works demonstrate the growing interest in diffusion-based structured generation, where image realism and label consistency must be jointly preserved. Compared to unconditional or purely class-conditional synthesis, joint image-mask generation imposes stronger structural constraints, requiring models to capture fine-grained spatial correspondences between visual content and pixel-level annotations.

In summary, prior work has relied primarily on U-Net-based DDPM and LDM architectures or on attention-derived masks from pretrained LDMs, while diffusion transformers that explicitly generate images and masks simultaneously, to the best of our knowledge, have not been explored.

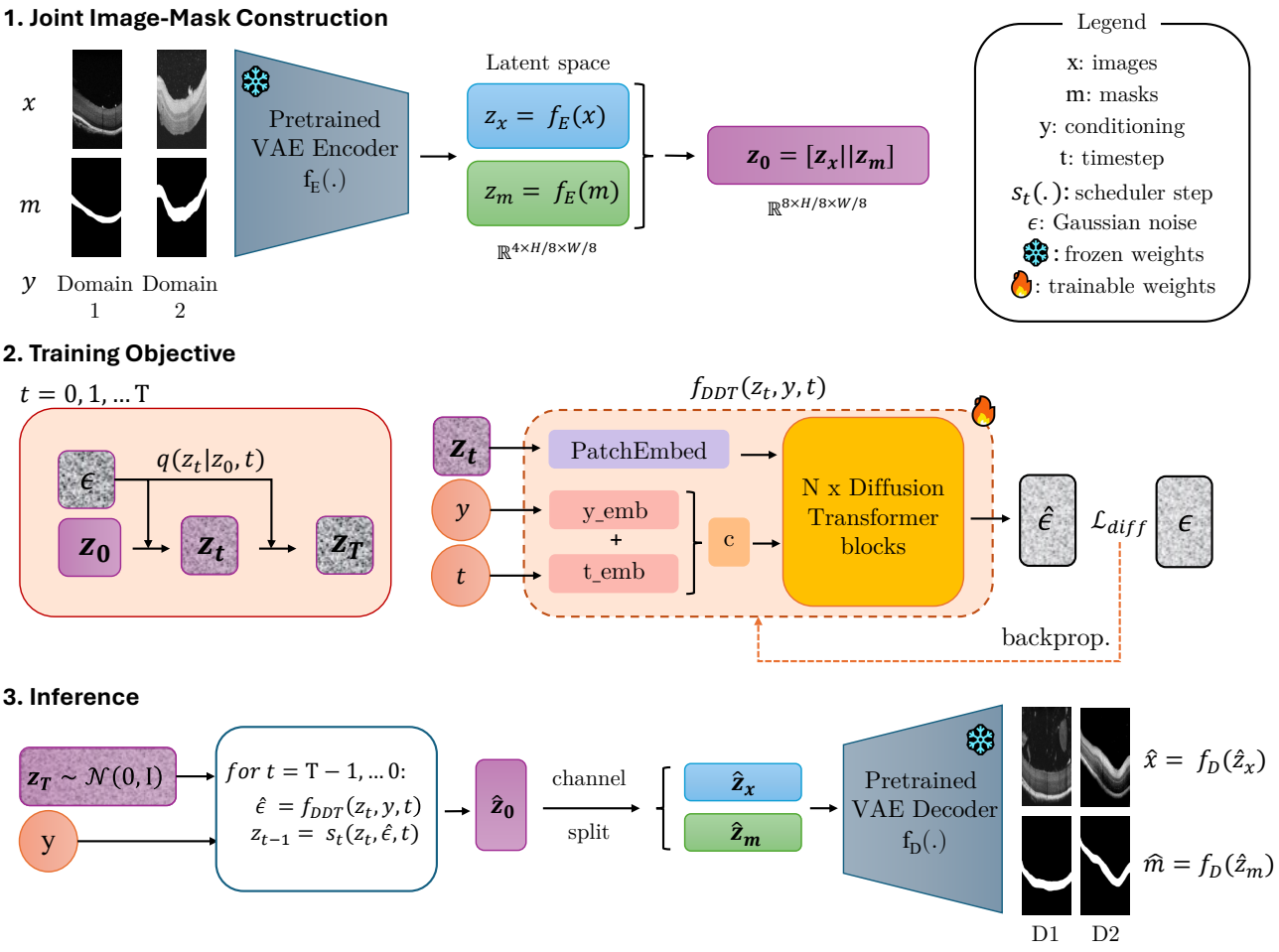


Figure 1: Overview of the DualDiT framework. Paired image and segmentation mask inputs (x, m) are encoded $(f_E(.))$ into a shared latent space and concatenated along the channel dimension to form a joint latent tensor $z_0 = [z_x \| z_m]$. During training, Gaussian noise (ϵ) is progressively added to z_0 over T timesteps following the forward process $q(z_t \mid z_0, t)$. At each timestep t, the noisy latent z_t, the timestep embedding, and the conditioning class embedding are combined into c and passed to N Diffusion Transformer blocks $(f_{DDT}(z_t, y, t))$ to model cross-modal spatial dependencies between image appearance and mask topology. During training, the network learns to predict the added noise $\hat{\epsilon}$, while during inference, this denoising is applied iteratively starting from $z_T \sim \mathcal{N}(0, I)$ until a clean joint latent $\hat{z}_0$ is recovered. The joint latent $\hat{z}_0$ is then split along the channel dimension into image and mask components and decoded $(f_D(.))$ back to pixel space (x', m').

3. Methods

Figure 1 presents an overview of the proposed **Conditional Dual-Output Diffusion Transformer (DualDiT)**. A detailed description of the different framework components is provided below.

Problem formulation: The objective of this work is to train a conditional generative model capable of jointly synthesising anatomically consistent OCT images and segmentation masks conditioned on the embedding medium of the retina. Each OCT image is paired with a segmentation mask delineating the URCL, a structure of interest for downstream analysis, and with a domain label indicating the embedding medium used during ex vivo tissue preparation. Let $\mathcal{D} = \{(x_i, m_i, y_i)\}_{i=1}^{N}$ denote an ex vivo mouse retinal OCT dataset, where $x_i \in \mathbb{R}^{1\times H\times W}$ represents a greyscale OCT image, $m_i \in \mathbb{R}^{1\times H\times W}$ its corresponding binary URCL segmentation mask, and $y_i \in \{0, \ldots, K-1\}$ the associated class label of each OCT domain.

Joint latent representation learning: To improve computational efficiency and reduce the dimensionality of the generative process, OCT images and segmentation masks are projected into a latent space using a pretrained VAE on natural images [40]. Given an OCT image x and its corresponding mask m, the encoder $f_E(.)$ maps both inputs into latent representations $z_x = f_E(x)$ and $z_m = f_E(m)$, respectively, each of spatial dimensions $\mathbb{R}^{4\times H/8\times W/8}$, where 4 corresponds to the latent channel dimensionality defined by the VAE architecture, and the spatial resolution is downsampled by a factor of 8. Since the VAE encoder and decoder are kept frozen during training, the learned latent representations are used as fixed projections. Both latent embeddings are concatenated along the channel dimension to construct a

shared latent representation:

$$z_0 = [z_x \parallel z_m], \in \mathbb{R}^{8\times H/8\times W/8}, \tag{1}$$

where $\parallel$ denotes channel-wise concatenation, yielding a joint latent tensor of dimensions $\mathbb{R}^{8\times H/8\times W/8}$. This joint latent formulation allows the model to learn the structural correspondence between retinal appearance and layer segmentation during the diffusion process while operating in a computationally efficient latent space.

Conditional dual-output Diffusion Transformer: The proposed framework introduces a conditional dual-output Diffusion Transformer architecture for the joint generation of OCT images and segmentation masks. Unlike previous DDPM- and LDM-based approaches relying on convolutional denoisers, the proposed model adopts a Transformer backbone to model long-range spatial dependencies through self-attention mechanisms.

The concatenated latent representation z_0 is progressively corrupted through a forward diffusion process over T timesteps following $q(z_t \mid z_0, t)$, which denotes the forward diffusion kernel that corrupts the clean latent z_0 by adding Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ according to a variance schedule $\{\bar{\alpha}_t\}_{t=1}^{T}$, yielding the noisy latent representation z_t at each timestep t [13]. The reverse denoising process models the conditional distribution $p_\theta(z_{t-1} \mid z_t, y)$, parameterized by a Transformer-based denoiser f_{DDT} that takes as input the noisy latent $z_t \in \mathbb{R}^{8\times H\times W}$, the domain label $y \in \{0, \dots, K-1\}$, and the diffusion timestep $t \in \{1, \dots, T\}$, and predicts the noise $\hat{\epsilon}$:

$$\hat{\epsilon} = f_{DDT}(z_t, y, t). \tag{2}$$

The timestep t is mapped to a continuous vector representation $\mathbf{t}_{\text{emb}} = \phi_t(t) \in \mathbb{R}^d$ via sinusoidal embeddings followed by a Multilayer Perceptron (MLP), and the domain label y is projected to $\mathbf{y}_{\text{emb}} = \phi_y(y) \in \mathbb{R}^d$ via a learned embedding table. Both are combined into a single conditioning vector $\mathbf{c} = \mathbf{t}_{\text{emb}} + \mathbf{y}_{\text{emb}} \in \mathbb{R}^d$. Following Peebles and Xie [41], each of the N Transformer blocks conditions on $\mathbf{c}$ via adaLN-Zero. Each block regresses six modulation parameters (scale, shift, and gate for both the self-attention and MLP sub-layers) directly from $\mathbf{c}$, and applies them residually after layer normalisation. All parameters are zero-initialised so that each block acts as an identity at the start of training. In contrast to conventional DiT frameworks [41] designed for single-image synthesis, the proposed formulation jointly predicts both OCT and segmentation latent representations within a unified denoising trajectory. This dual-output design enforces anatomical consistency between retinal structures and layer masks during generation.

Optimisation objective: Since the VAE encoder and decoder are kept frozen during training, the model parameters are optimised using the standard diffusion noise prediction objective. Given a noisy latent sample z_t and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, the optimisation minimises the mean squared error between the predicted and actual noise:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{z_0,\epsilon,t}\left[\left\|\epsilon - f_{DDT}(z_t, y, t)\right\|_2^2\right]. \tag{3}$$

This objective enables progressive reconstruction of anatomically plausible OCT image-mask latent pairs throughout the reverse diffusion process.

Joint OCT and mask synthesis: During inference, the reverse diffusion process is initialised by sampling $z_T \sim \mathcal{N}(0, I)$ and iteratively denoising according to:

$$z_{t-1} = s_t(z_t, \hat{\epsilon}, t), \quad \hat{\epsilon} = f_{DDT}(z_t, y, t), \quad \text{for } t = T-1, \dots, 0, \tag{4}$$

where $s_t(\cdot)$ denotes the scheduler step. After the reverse diffusion process completes, the joint latent $\hat{z_0}$ of dimensions $\mathbb{R}^{8\times H/8\times W/8}$ is split along the channel dimension into an image $\hat{z_x}$ and a mask component $\hat{z_m}$, each of dimensions $\mathbb{R}^{4\times H/8\times W/8}$. Both components are independently decoded through the pretrained VAE decoder $f_D(.)$ to reconstruct the synthetic OCT image and its corresponding segmentation mask:

$$\hat{x} = f_D(\hat{z_x}) \in \mathbb{R}^{1\times H\times W}, \qquad \hat{m} = f_D(\hat{z_m}) \in \mathbb{R}^{1\times H\times W}. \tag{5}$$

Since both outputs are generated from a shared latent diffusion trajectory, the synthesised image-mask pairs preserve spatial and anatomical coherence.

4. Experimental settings

4.1. Dataset

The dataset used in this study comprises OCT B-scans of ex vivo mouse retinas acquired with a high-resolution OCT system (Thorlabs Ganymede Series, Thorlabs GmbH, Luebeck, Germany). Retina samples were prepared, and OCT data were acquired as part of a previous study conducted within a project approved by the local authorities [42]. Retinas were collected from 7 male and 3 female mice aged 2 to 27.5 months (10.27 ± 9.09 months). Because several OCT volumes were derived from the same animal, sex and age are reported at the animal level, whereas volume counts in Table 1 reflect the total number of processed samples. Retina samples were prepared using two distinct embedding protocols, which serve as the primary classes for model conditioning.

The first group included 10 volumes of retina maintained in a physiological liquid medium, specifically water or agarose gel, which preserves tissue hydration close to native physiological conditions. However, in a previous study, Agarose-embedded samples showed higher background noise/scattering and degraded after several weeks, whereas resin-embedded samples remained stable for several years [23].

In contrast, the second group consisted of 12 volumes from retina embedded in resin. Resin embedding produces highly stable, durable samples, making it suitable for long-term preservation, repeated imaging, and phantom-based OCT evaluation. Barroso et al. [23] showed that resin-embedded murine retina can preserve layered retinal structures comparable to those observed in gel-based

Table 1
Distribution of the number of animals, retinal OCT volumes, and manually segmented B-scans across the two experimental groups.

Dataset	N animals	OCT volumes	Segmented B-Scans
Resin	5	12	137
Physiological	5	10	203
Total	**10**	**22**	**340**

preparations, while producing lower background scattering/noise than agarose-based media. However, resin preparation requires fixation, dehydration, and polymerisation steps, which are irreversible and may introduce subtle shrinkage or preparation-related artefacts.

The differences between the datasets define two complementary OCT imaging domains, with considerable variation in the visibility and delineation of retinal layers. This domain distinction motivates the conditional generation of both embedding types rather than modelling them as a single homogeneous distribution.

In addition to the embedding protocol, the dataset included both control retinas from untreated mice and corresponding retinas treated with N-Methyl-D-aspartate (NMDA) to simulate glaucoma-induced retina degradation. This treatment status was not used to define the conditioning domains, but it increases the morphological variability of the dataset by introducing changes in retinal layer thickness and structure.

To generate the set of ground-truth masks, an expert manually segmented the upper retinal cell layer (URCL) in 203 B-scans from 10 physiologically-embedded volumes and 137 B-scans from 12 resin-embedded volumes, ensuring a highly accurate training database. Figure 2 shows two representative examples of retinas embedded in the two different embedding media, along with the delineation of the URCL. Resin-embedded images exhibit sharper layer boundaries and higher signal contrast, albeit with occasional horizontal line artefacts arising from the air-resin interface, whereas physiologically-embedded images display a more uniform but lower-contrast signal with distributed speckle noise, resulting in more ambiguous layer transitions.

To ensure an unbiased evaluation of the downstream evaluation protocol that will be described in section 4.3.2, we divided the manually annotated B-scans into separate training, validation, and test sets, as summarised in Table 2. To prevent optimistic bias due to inter-slice correlations, splitting was performed at the volume level. For training generative models, we used only the training subset (236 B-scans: 95 from resin embedding and 141 from physiological embedding. This choice is critical because synthetic augmentation is evaluated through a downstream segmentation benchmark whose validation and test splits must remain strictly unseen during any generative training stage to maintain the independence of the augmentation assessment. Otherwise, the generative model could leak information from the held-out set into the segmentation model, compromising the fairness of the comparison.

Table 2
Distribution of volumes, segmented B-scans across training, validation, and test partitions, categorised by embedding medium.

Subset	Medium	Volumes	Segmented B-scans
Training	Resin	9	95
	Physiological	7	141
	All	**16**	**236**
Validation	Resin	1	23
	Physiological	1	20
	All	**2**	**43**
Test	Resin	2	19
	Physiological	2	42
	All	**4**	**61**

The validation set is used for the segmentation model selection and hyperparameter tuning, while the test set is kept separate for final performance evaluation. It is important to note that the downstream segmentation benchmark is evaluated only on the validation and test real images.

All images and masks were resized to 512×256 pixels to accommodate the variability in acquisition dimensions and because manual URCL delineations do not span the full retinal width. OCT B-scans are greyscale, and masks are binary (0: background, 1: URCL).

4.2. Generative Model Training

All frameworks were implemented in PyTorch and trained on an NVIDIA DGX A100 system. All experiments were conducted using PyTorch 2.5 and Python 3.10. The code is publicly available at `https://github.com/cvblab/DualDiT`.

Training inputs: For diffusion training, inputs were scaled to $[-1, 1]$ to match the scale of the Gaussian noise added during the forward process and to ensure zero-centred inputs, which improves training stability. The image and mask were then encoded into a compressed latent space using the pretrained `sd-vae-ft-ema` VAE from StabilityAI [40], originally trained on OpenImages and subsequently fine-tuned on LAION-Aesthetics and LAION-Humans, and the standard latent scaling factor provided by the VAE configuration was applied. The encoder compresses the image and mask inputs by a factor of 8, producing latent tensors of spatial resolution 64×32. For joint image-mask generation, each training input was represented as a channel-wise concatenation of the image and its corresponding mask, yielding in a 8-channel latent representation. Conditioning used two classes corresponding to the embedding medium (*resin* vs *physiological*).

DualDiT backbone: The model was instantiated as a DiT-XL/2 backbone, processing the concatenated latent tensor through a patch embedding layer (patch size 2) enriched with fixed 2D sine-cosine positional encodings, followed by 28 transformer blocks with hidden size 1152 and 16 attention heads.

Model hyper-parameters: Optimisation was performed with Adam [43] for 5000 epochs, with a learning rate of 10^{-4}

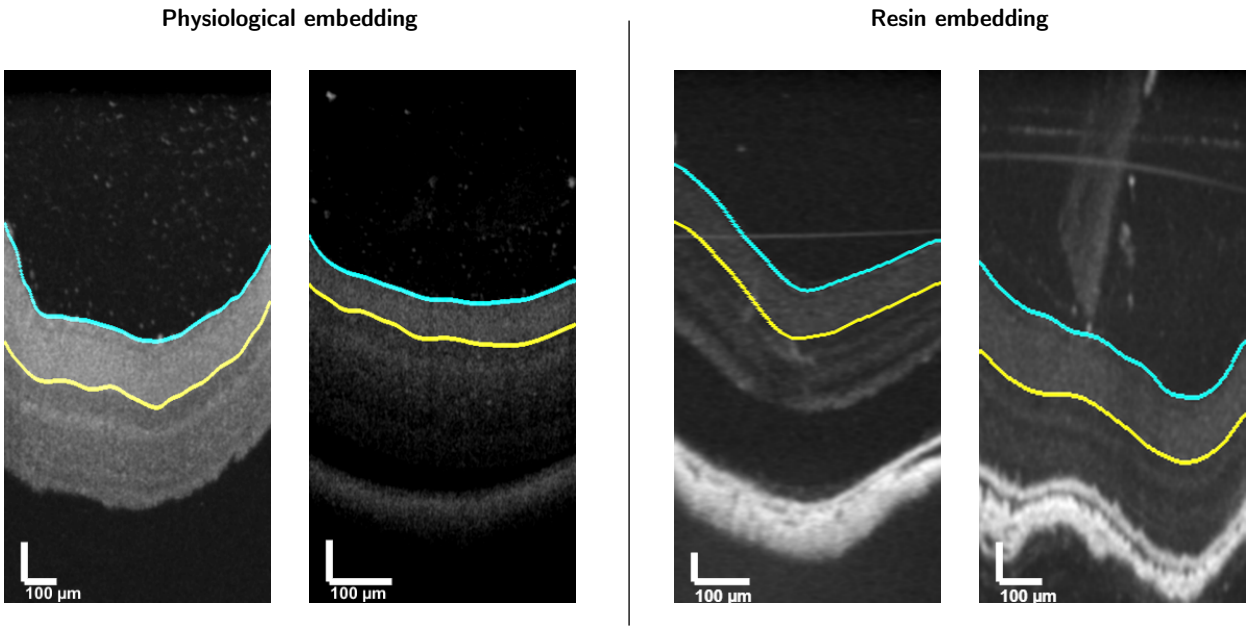


Figure 2: Examples of URCL layer segmentation for representative B-scans of fluid and resin embedded samples. The images show the upper and lower boundaries of the URCL layer across the two experimental groups (see Table 1).

and with batch size 16. The diffusion process used $T = 1000$ timesteps and a linear noise schedule ($\beta_{\text{start}} = 1.5 \times 10^{-4}$, $\beta_{\text{end}} = 1.95 \times 10^{-2}$). The model was trained with classifier-free guidance [44] with an unconditional probability of 0.1. An exponential moving average (EMA) of the diffusion parameters with decay 0.9999 was maintained during training and used for sampling.

To evaluate the efficacy of the proposed DualDiT, two state-of-the-art diffusion architectures, conditionals DDPM and LDM, were adapted as baselines, ensuring a fair comparison by configuring them for joint image-mask generation:

Conditional DDPM (Pixel Space) [13]: In this approach, the OCT image and its corresponding segmentation mask were treated as distinct channels of a single input tensor. Following normalisation, both components were concatenated along the channel dimension, resulting in a joint representation $x \in \mathbb{R}^{2\times H\times W}$. The U-Net denoiser was trained to predict the combined noise of this dual-channel input, forcing the model to learn the joint distribution of the retinal anatomy and its pixel-level annotation directly in pixel space. The DDPM baseline employed a `UNet2DModel` from the `diffusers` library with channel widths of [32, 64, 64, 128, 128], one ResNet layer per block, and attention at the two deepest resolutions. Class conditioning was implemented by concatenating a learned class embedding of size 4 with the noisy input. The optimal hyper-parameter combination was achieved by training the DDPM for 6000 epochs. All remaining hyperparameters (diffusion timesteps T, learning rate, noise schedule, batch size, and classifier-free guidance probability) are shared with DualDiT as described above.

Conditional LDM (Latent Space) [14]: This implementation utilises a pre-trained VAE [40] to perform diffusion in a compressed space. The resulting latent representations, z_x and $z_m \in \mathbb{R}^{4\times H/8\times W/8}$, were then concatenated to form an 8-channel latent tensor. The LDM uses a U-Net denoiser as DDPM to model cross-modal spatial dependencies while leveraging the computational efficiency of the latent space. The LDM baseline employed a custom U-Net with channel widths of [64, 128, 256, 512, 512], two layers per stage, a bottleneck of [512, 512], 8 attention heads, Group Normalisation, and SiLU activations. The optimal hyperparameters are shared with DualDiT as described above and were achieved training during 9500 epochs.

4.3. Evaluation protocols

4.3.1. Generative quality metrics

We evaluated the quality and diversity of the generated images using several standard metrics for generative models. First, we computed the Fréchet Inception Distance (FID) [45], which measures the distance between the feature distributions of real and generated images extracted from a pretrained Inception network. We represent the feature distributions of synthetic and real patches as $\mathcal{N}(\mu_{synth}, \Sigma_{synth})$ and $\mathcal{N}(\mu_{real}, \Sigma_{real})$, respectively. The FID expression is given by:

$$\text{FID} = \|\mu_{synth} - \mu_{real}\|^2 + \text{Tr}\left(\Sigma_{synth} + \Sigma_{real} - 2(\Sigma_{synth}\Sigma_{real})^{\frac{1}{2}}\right) \tag{6}$$

Note that FID $\in [0, +\infty)$, where lower values indicate higher similarity between real and generated distributions.

In addition, we report on the spatial FID (sFID), with sFID $\in [0, +\infty)$, which evaluates the similarity of spatial features rather than grouped activations, providing a more sensitive assessment of structural consistency in the generated images.

All metrics were computed using the evaluation pipeline provided in the public implementation of Guided Diffusion models by OpenAI [46]. The metrics were computed both globally and stratified by embedding medium to assess class-conditional fidelity.

4.3.2. Downstream segmentation model training

To evaluate the clinical utility of the synthesized image-mask pairs, we perform a downstream segmentation task. We employ a standard U-Net architecture [47] as a baseline segmentation model, initially trained solely on the real training set ($\mathcal{D}$). We then evaluate the performance gain when the training set is augmented with synthetic image-mask pairs ($\mathcal{D}'$) generated by DDPM, LDM, and DualDiT.

For real annotated images and masks, we used a standard combination of binary cross-entropy (BCE) and Dice loss [48], weighted equally ($\alpha = 0.5$):

$$\mathcal{L}_{\text{real}} = \alpha\, \mathcal{L}_{\text{BCE}} + (1 - \alpha)\, \mathcal{L}_{\text{Dice}} \tag{7}$$

When synthetic pairs were included, we adopted a confidence-guided teacher-student strategy to regulate their influence during training [49]. A teacher model trained exclusively on real data produces a pixel-wise confidence map $c_i = 2\left|p_i^{(T)} - 0.5\right|$, where $p_i^{(T)} = \sigma(f_T(x_s))_i$ denotes the teacher's predicted probability of a synthetic image (x_s) at pixel i. Thus, the confidence map down-weights uncertain synthetic regions in the BCE term while leaving the Dice loss unweighted, preserving a global structural penalty regardless of local mask quality. The total loss is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{real}} + \mathcal{L}_{\text{synth}}$, where $\mathcal{L}_{\text{synth}}$ combines confidence-weighted BCE and standard Dice over the synthetic branch.

Model architecture: The downstream segmentation model employed was a U-Net fully convolutional network adapted for single-channel OCT B-scans and binary masks. The encoder consists of four convolutional blocks with 64, 128, 256, and 512 channels, followed by a bottleneck with 1024 channels. The decoder mirrors the encoder with upsampling and skip connections, and the output layer uses a 1×1 convolution to produce a single-channel probability map for URCL segmentation.

Training protocol: All models were trained for 300 epochs (batch size 16) using Adam (lr = 10^{-4}, weight decay 10^{-4}) with a *ReduceLROnPlateau* scheduler (factor 0.5, patience 30 epochs) monitoring the validation Dice coefficient. These hyperparameters were fixed across the baseline and all augmentation strategies (DDPM, LDM, DualDiT) to ensure a fair comparison.

4.3.3. External evaluation protocol

As an additional evaluation method, a panel of experts with varying levels of expertise conducted a qualitative assessment of the synthesised images and masks. To this end, a visual evaluation test was prepared using image-mask pairs, in which the experts were required to distinguish real pairs from synthetic ones and to classify each pair according to the embedding medium (resin or physiological). In total, 100 images (50 real and 50 synthetic) were analysed. The two embedding classes were equally represented, yielding 25 images per class and per sample type (real or synthetic).

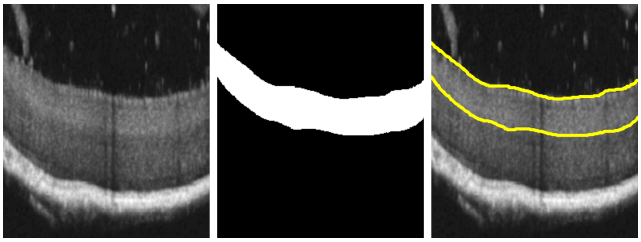


Figure 3: Example figure of a synthetic resin-embedded sample for expert evaluation: (Left) OCT B-Scan; (Middle) URCL segmentation mask; (Right) URCL delineation on the B-Scan.

To ensure the samples were assessed correctly, in each case, the B-scan, the URCL segmentation mask and the outline of the mask's edges on the B-scan were displayed. As shown in Figure 3, this representation enabled the experts to evaluate not only the quality of the generated B-scans but also the correspondence between the B-scans and their corresponding segmentations.

5. Results

5.1. Generative Quality Comparison

Quantitative results: Table 3 presents the results for FID and sFID for the diffusion models examined in this study: DDPM[13], LDM [14], and the proposed DualDiT. The *Global* column of the table shows the results for the whole real data distribution. In contrast, the *Resin* and *Physiological* columns stratify the results for each embedding medium. Our framework achieves significant performance improvements over state-of-the-art models, with FID and sFID scores of 56.14 and 114.35, respectively. When analysing the results by embedding medium, we find that DualDiT outperforms the other frameworks across both metrics.

Qualitative results and visual inspection: To qualitatively evaluate the proposed method, we present synthesised images using the DDPM, LDM, and DualDiT frameworks. Figure 4 shows comparisons of real physiological and resin-embedded images with their corresponding mask overlays and the outputs of the different methods assessed: DDPM, LDM, and our proposed model (DualDiT). The DualDiT approach demonstrates greater resolution of the retinal layers and high reliability in reproducing common OCT artefacts and retinal defects observed in the actual data collected during experimental protocols and the embedding process.

For completeness, additional synthetic samples are provided in Appendix A (Figures A.1 and A.2), showing respectively, physiological and resin-embedded retinal images alongside their corresponding masks generated by the DDPM approach. This framework exhibits hallucinations in the shape of the retina for both resin and physiological embedding mediums. While it successfully captures the grey levels and distribution of the original data, it lacks

Table 3
Quantitative comparison of generative models across datasets.

	Physiological			Resin			Global		
	DDPM [13]	LDM [14]	DualDiT	DDPM [13]	LDM [14]	DualDiT	DDPM [13]	LDM [14]	DualDiT
FID ↓	175.02	124.79	**66.96**	208.99	109.60	**59.83**	164.55	102.21	**56.14**
sFID ↓	278.87	157.93	**122.15**	329.91	153.70	**114.72**	254.52	150.66	**114.35**

the resolution required to differentiate between the URCL and the remaining retinal layers, and fails to reproduce the continuous structure of the retina and common OCT image artefacts such as speckle noise and refraction effects. Additionally, it often exaggerates intensity, resulting in overly bright white tones. With regard to mask synthesis, it can be observed how it segments artifacts outside the boundaries of the region of interest

Figures A.3 and A.4 show pairs of retinal images and masks generated by the LDM framework. This approach significantly improves upon the DDPM results, offering greater layer definition, higher overall resolution and contrast, more accurate replication of retinal shapes, and reduced output variability. However, there is still a perceived inconsistency in the resolution and detail of the retinal images. Similar to DDPM, LDM sometimes produces images that are overly saturated and bright, especially in physiological medium images (e.g., the second and fourth images in Figure A.4).

Finally, Figures A.5 and A.6 display images and masks generated by the proposed DualDiT model. The synthetic images show a closer resemblance to real data for both resin and physiological embedding mediums. Qualitatively, the layers of the retina generated by DualDiT, particularly the URCL, appear to have higher resolution and contrast compared to the results from the DDPM and LDM models.

5.2. Segmentation performance with synthetic augmentation

This section presents the validation of the proposed method as a data augmentation strategy. To evaluate its effectiveness, we assess the segmentation model described in Section 4.3.2 using the Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) [50] on a strictly held-out real test set presented in Table 2. The reference model was trained on the training subset ($\mathcal{D}$), referred to as the *Baseline* model. This model achieves a global Dice score of 0.908 ± 0.066 and an Intersection over Union (IoU) of 0.838 ± 0.106 on the test set.

Table 4 summarises the segmentation performance obtained with the Baseline and the different synthetic-data augmentation strategies across the physiological, resin, and global test subsets. By augmenting the training dataset, as $\mathcal{D} \cup \mathcal{D}'$, with $\mathcal{D}' = 50$ synthetic images and masks, comprising 25 inputs from the resin class and 25 from the physiological class, generated by the DDPM, LDM, and DualDiT frameworks, we trained three additional segmentation models. As shown in Table 4, among the three generative frameworks, DualDiT provides the most consistent results and is the only method to improve both global metrics

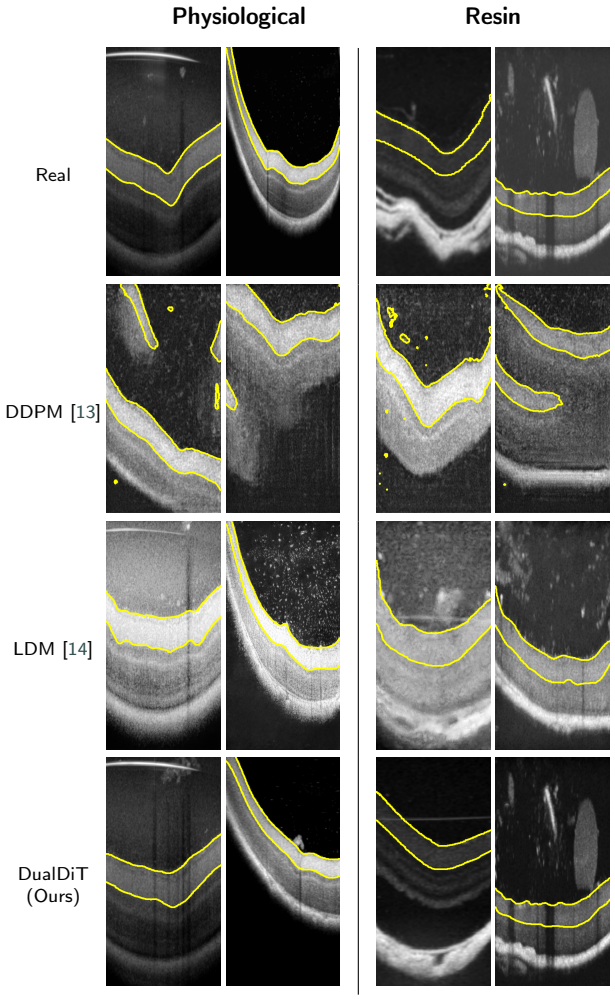


Figure 4: Comparison of real and synthetic images with their corresponding URCL mask overlays. The left two columns show physiologically embedded samples, whereas the right two columns show resin-embedded samples.

relative to the Baseline. The model trained with DualDiT-generated samples achieves a global Dice score of 0.917 ± 0.058 and an IoU of 0.852 ± 0.095, compared with 0.908 ± 0.066 and 0.838 ± 0.106 for the Baseline. This improvement is consistent across embedding media, with DualDiT achieving Dice and IoU scores comparable to or higher than those of the Baseline in both the physiological and resin subsets. In contrast, DDPM substantially degrades performance in the resin subset, while LDM yields intermediate results but does not consistently surpass the Baseline. This degradation is consistent with the generative quality metrics reported in Table 3 as DDPM is the only model for which FID and

Table 4
Comparison of segmentation performance between the baseline model and models trained with different numbers of synthetic image-mask pairs. Results are reported as mean and standard deviation for Dice and IoU. $\mathcal{D}$ denotes the original training set composed of 236 image-mask pairs.

		Dice						IoU					
		Physiological		Resin		Global		Physiological		Resin		Global	
Model	Training images	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Baseline	$\mathcal{D}$	0.912	0.051	0.906	0.072	0.908	0.066	0.843	0.079	0.836	0.116	0.838	0.106
DDPM [13]	$\mathcal{D}+50$	0.917	0.073	0.528	0.312	0.796	0.258	0.855	0.116	0.415	0.266	0.718	0.270
LDM [14]	$\mathcal{D}+50$	0.904	0.078	0.845	0.088	0.886	0.086	0.834	0.125	0.742	0.126	0.805	0.133
DualDiT	$\mathcal{D}+50$	0.919	0.064	0.913	0.041	0.917	0.058	0.856	0.105	0.842	0.067	0.852	0.095
DDPM [13]	$\mathcal{D}+100$	0.921	0.070	0.603	0.199	0.822	0.193	0.860	0.112	0.459	0.199	0.735	0.235
LDM [14]	$\mathcal{D}+100$	0.907	0.079	0.897	0.055	0.903	0.072	0.838	0.126	0.817	0.084	0.832	0.115
DualDiT	$\mathcal{D}+100$	0.914	0.069	0.915	0.036	0.914	0.060	0.849	0.112	0.846	0.061	0.848	0.099
DualDiT	$\mathcal{D}+200$	0.921	0.063	0.926	0.025	0.923	0.054	0.860	0.104	0.864	0.042	0.861	0.089
DualDiT	$\mathcal{D}+400$	0.921	0.063	0.925	0.027	0.922	0.054	0.859	0.103	0.862	0.046	0.860	0.090
DualDiT	$\mathcal{D}+600$	0.925	0.061	0.919	0.037	0.923	0.055	0.866	0.100	0.853	0.062	0.862	0.090
DualDiT	$\mathcal{D}+800$	0.920	0.062	0.929	0.023	0.923	0.053	0.858	0.102	0.868	0.039	0.861	0.088
DualDiT	$\mathcal{D}+1000$	0.927	0.054	0.919	0.033	0.925	0.049	0.869	0.090	0.852	0.056	0.864	0.082
DualDiT	$\mathcal{D}+1200$	**0.927**	**0.055**	**0.927**	**0.027**	**0.927**	**0.048**	**0.869**	**0.092**	**0.866**	**0.046**	**0.868**	**0.081**
DualDiT	$\mathcal{D}+1400$	0.926	0.057	0.917	0.034	0.923	0.051	0.867	0.094	0.848	0.056	0.861	0.085

sFID are markedly worse for the resin subset than for the physiological subset (+33.97 and +51.04 points, respectively), whereas both LDM and DualDiT achieve better resin than physiological scores despite the resin subset containing fewer training volumes (Table 2). This asymmetry suggests that the limited number of resin training samples is not, by itself, an insurmountable barrier to synthesis, but rather interacts with DDPM's more limited modelling capacity, which appears to hinder its ability to generalise from a comparatively smaller training set.

We also evaluated augmentation with 100 synthetic image-mask pairs, evenly split between the two classes. At this augmentation level, DualDiT again provides the strongest overall results among the three generative approaches, reaching a global Dice score of 0.914 ± 0.060 and an IoU of 0.848 ± 0.099. However, these values are only slightly above the Baseline and are lower than those obtained with larger DualDiT augmentation sets. DDPM continues to exhibit a marked degradation in the resin subset, whereas LDM improves substantially over DDPM but does not provide a consistent global gain over the Baseline. These results suggest that synthetic image-mask pairs generated by DualDiT can provide a consistent benefit when augmenting a limited OCT segmentation dataset, although the magnitude of the improvement depends on the number of synthetic samples added.

Overall, DualDiT is the only augmentation framework that consistently improves the Baseline across multiple augmentation sizes. DDPM performs poorly in the resin subset and exhibits substantially larger variability, which may indicate limited generalisation or a domain mismatch in the generated samples. LDM produces more competitive results than DDPM, particularly for resin-embedded images, but its global performance remains close to or below the Baseline. By contrast, DualDiT generally improves the resin results while maintaining comparable performance in the physiological subset.

Given the comparatively consistent behaviour of DualDiT, we extended the augmentation study from $\mathcal{D}$ + 200 up to $\mathcal{D}$ + 1400 synthetic image-mask pairs, following the progressive data-expansion strategy adopted in previous synthetic augmentation studies [34]. DualDiT yields global Dice and IoU values above the Baseline for nearly all evaluated augmentation sizes, although the improvement is not monotonic. The best overall performance is obtained with $\mathcal{D}$ + 1200, reaching a global Dice score of 0.927 ± 0.048 and an IoU of 0.868 ± 0.081. The gains are particularly evident in the resin subset, while the differences in the physiological subset are smaller, partly because the Baseline performance is already comparatively high. Similar improvements are observed for $\mathcal{D}$ + 600 and $\mathcal{D}$ + 1000, indicating that the benefit is not restricted to a single augmentation size in this particular dataset. This suggests that augmentation quality and diversity are more relevant than the absolute number of generated samples. Since the synthetic pairs are derived from the available training distribution, increasing their number may eventually introduce redundancy rather than additional variability. Nevertheless, this interpretation remains hypothetical and would require further experiments with larger datasets and independent test volumes.

At the global level, the $\mathcal{D}+1200$ configuration increases the mean Dice and IoU scores by 2.09% and 3.58% relative to the Baseline, respectively, indicating a modest but consistent overall benefit across the complete test set. To complement these results, a per-B-scan analysis was conducted based on relative changes in Dice and IoU relative to the Baseline. Figure A.7 presents the full per-B-scan comparison for the $\mathcal{D}$ + 1200 configuration. The Dice and IoU scatter plots in Figures A.7a and A.7b, respectively, show that most B-scans lie above the identity line, with the largest gains occurring mainly in cases where the Baseline performs poorly. Figure A.7c shows the relative percentage changes in Dice and IoU for all test B-scans, ordered by increasing IoU change.

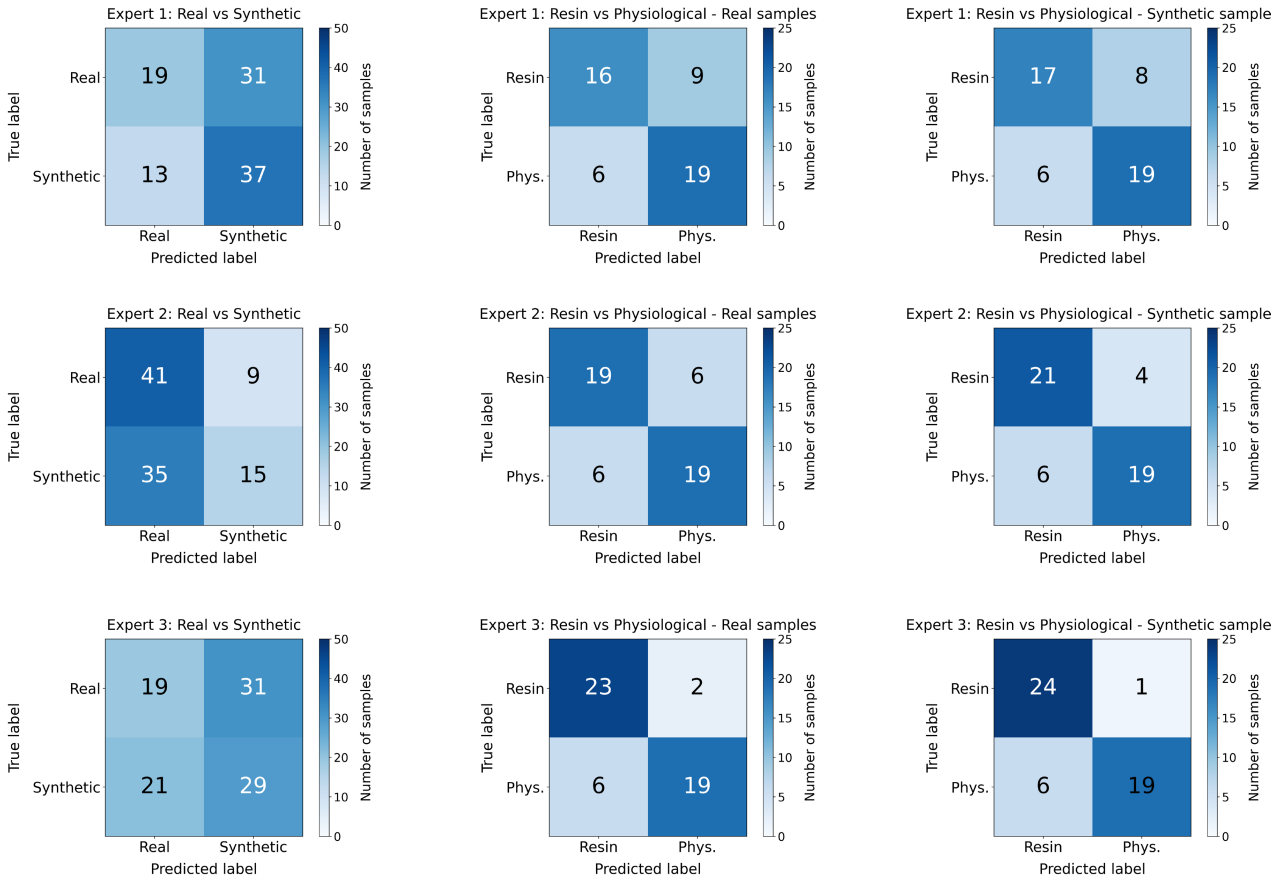


Figure 5: Confusion matrix for the classification performed by experts. (Left) Classification for real and synthetic samples; (Middle) Embedding medium classification for real samples; (Right) Embedding medium classification for synthetic pairs.

The results highlight that DualDiT outperforms the Baseline in the majority of cases, although performance decreases are observed for a small subset of B-scans.

5.3. External validation protocol

To demonstrate the usefulness and realism of the proposed model's outputs, a panel of three experts evaluated the quality of the synthetic images. Figure 5 shows the results obtained by every expert. Experts incorrectly classified, on average, 46% of synthetic samples as real. On the other hand, 42% of the real samples were deemed synthetic.

As regards the classification of embedding media, both resin and liquid embeddings were correctly identified in both real and synthetic samples, with similar success rates (76.7% and 79.3%, respectively). Additionally, we tested whether there was a statistically significant difference in classification performance between synthetic and real samples. For this purpose, the area under the receiver operating characteristic (ROC) curve (AUC) was calculated, a metric ranging from 0.5 (random performance) to 1 (perfect classification). Table 5 shows the AUC metric for identifying real (R) and synthetic (S) samples for each expert. In addition, the p-value (with $\alpha = 0.05$) is provided. Table 5 shows no statistically significant differences ($p - value > 0.5$) between the identification of media based on real or synthetic samples. This demonstrates that the developed model successfully distinguishes and represents both embedding materials.

Table 5
AUC metrics for resin vs physiological classification on real and synthetic images, including the p-value for the comparison between AUC-R and AUC-S.

Expert	AUC_{real}	AUC_{synth}	p-value
Expert 1	0.700	0.720	0.852
Expert 2	0.760	0.800	0.662
Expert 3	0.840	0.860	0.801

6. Conclusions

In this study, we introduce a conditional dual-output Diffusion Transformer (DualDiT) that jointly synthesises OCT images and their corresponding segmentation masks in a single generative process. We evaluate its performance

by generating realistic OCT B-scans of dissected ex vivo mouse retinas embedded in two distinct media, together with anatomically consistent URCL segmentation masks. The proposed model achieves an overall FID of 56.14, which is substantially lower than those of other tested diffusion models, such as DDPM and LDM, which achieved FIDs of 164.55 and 102.21, respectively. To validate the proposed method as a data augmentation strategy, we trained a URCL segmentation model and evaluated it on a separate dataset. The Baseline already achieved a relatively high global Dice score of 0.908 and an IoU of 0.838, leaving limited room for improvement. Nevertheless, DualDiT was the only evaluated augmentation framework that improved the Baseline across several augmentation sizes. The best configuration, obtained with $D + 1200$, reached a global Dice score of 0.927 and an IoU of 0.868, corresponding to relative improvements of 2.07% and 3.50%, respectively. These findings support the potential utility of augmenting limited training datasets with images and masks generated by DualDiT to aid the development of segmentation models from limited datasets. Regarding the quality inspection, the expert evaluation showed that synthetic images and masks were not always readily distinguishable from real samples. These findings suggest that DualDiT was able to reproduce both the general appearance of real OCT inputs and visual characteristics associated with the resin and physiological embedding media, with a degree of perceptual plausibility recognised by the evaluators.

Regarding the limitations of this study, the resolution and size of the generated images, as well as the amount of training data, can pose challenges to the model's applicability in clinical settings. Although the resolution used in this study (512×256) proved adequate for the experiments conducted, real-world clinical scenarios may require higher resolutions to capture greater detail for diagnostic and planning tasks. It should be noted that this limitation is primarily due to computational constraints and that the proposed approach is scalable to higher resolutions (e.g., 1024×512) at the cost of increased model complexity and hardware resource requirements.

Similarly, the limited size of the training dataset, together with the small number of annotated samples and independent test volumes, may constrain both the model's ability to generalise and the statistical strength of the downstream evaluation. The reported improvements, therefore, provide descriptive evidence of a consistent benefit, but further validation across larger datasets is required to establish statistical significance and generalisability. In particular, because B-scans within the same volume are highly correlated, the effective diversity of the dataset is smaller than the total number of annotated B-scans might suggest. Future research should evaluate the method across a wider range of acquisition conditions, specimens, and independent datasets. In this regard, it is important to note that obtaining datasets with pixel-level annotations of retinal layers is itself a significant challenge in medical imaging, as it is labour-intensive and requires expert knowledge.

Finally, the generation of synthetic data using diffusion models involves a trade-off between the diversity of the generated outputs and their fidelity to the real world. Although the proposed approach achieves a high degree of realism in the synthesised images, future research could focus on increasing the variability of the samples while maintaining this realism.

In conclusion, this work introduces a novel conditional DualDiT framework for jointly generating OCT images and segmentation masks, showing that it provides more favourable generative and downstream segmentation results than the evaluated DDPM- and LDM-based alternatives. Beyond its quantitative improvements, the proposed method highlights the potential of jointly modelling images and annotations as an effective data augmentation strategy in data-scarce medical scenarios. In general, these results suggest that transformer-based diffusion models constitute a promising direction for advancing synthetic data generation and supporting the development of robust medical image analysis systems.

7. Acknowledgments and declarations

Acknowledgments

We gratefully acknowledge Steffi Ketelhut, from the Biomedical Technology Center of the Medical Faculty of the University of Muenster, for her collaboration in the animal experimentation.

We also thank the support from the Generalitat Valenciana (GVA) with the donation of the DGX A100 used for this work, an action co-financed by the European Union through the Operational Program of the European Regional Development Fund of the Comunitat Valenciana 2014-2020 (IDIFEDER/2020/030).

Funding

This work was funded by Horizon Europe, the European Union's Framework Programme for Research and Innovation, under Grant Agreement No. 101070062 (SEQUOIA), by Horizon 2020 under Grant Agreement No. 732613 (GALAHAD), and by the Generalitat Valenciana under Grant CIPROM/2022/20 (COMTACTS2).

Ethics statement

This study was conducted in accordance with the ARVO statement on the use of animals in ophthalmic and vision research. This study was approved by the LANUV North Rhine-Westphalia, Recklinghausen, Germany, and the Animal Protection Office, University of Münster (Approval No. 84-02.04.2016.A395 and T24.036UMS).

Competing interests

The authors declare no competing interests.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used OpenAI ChatGPT (accessed in May 2026) to assist with

language polishing and consistency checks. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

CRediT authorship contribution statement

Fernando García-Torres: Conceptualization, Methodology, Data curation, Investigation, Formal analysis, Writing and Visualization. **Rocío del Amor:** Conceptualization, Methodology, Investigation, Formal analysis, Writing and Visualization. **Sandra Morales:** Conceptualization, Methodology, Investigation, Formal analysis, Writing and Visualization. **Álvaro Barroso:** Data curation, Review and editing. **Peter Heiduschka:** Data curation. **Björn Kemper:** Data curation, Review and editing. **Valery Naranjo:** Conceptualization, Review and editing, and Supervision.

References

[1] A. Kebaili, J. Lapuyade-Lahorgue, S. Ruan, Deep learning approaches for data augmentation in medical imaging: A review, Journal of Imaging 9 (2023). doi:10.3390/jimaging9040081.

[2] P. Celard, E. L. Iglesias, J. M. Sorribes-Fdez, R. Romero, A. S. Vieira, L. Borrajo, A survey on deep learning applied to medical images: from simple artificial neural networks to generative models 35 (2023) 2291–2323. doi:10.1007/s00521-022-07953-4.

[3] P. Chlap, H. Min, N. Vandenberg, J. Dowling, L. Holloway, A. Haworth, A review of medical image data augmentation techniques for deep learning applications 65 (2021) 545–563. doi:10.1111/1754-9485.13261.

[4] S. Islam, M. T. Aziz, H. R. Nabil, J. R. Jim, M. F. Mridha, M. M. Kabir, N. Asai, J. Shin, Generative adversarial networks (GANs) in medical imaging: Advancements, applications, and challenges 12 (2024) 35728–35753. doi:10.1109/ACCESS.2024.3370848.

[5] C. Oulmalme, H. Nakouri, F. Jaafar, A systematic review of generative AI approaches for medical image enhancement: Comparing GANs, transformers, and diffusion models 199 (2025) 105903. doi:10.1016/j.ijmedinf.2025.105903.

[6] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, P. R. Pinheiro, Covidgan: data augmentation using auxiliary classifier gan for improved covid-19 detection, Ieee Access 8 (2020) 91916–91923. doi:10.1109/ACCESS.2020.2994762.

[7] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, 2017, pp. 1125–1134. URL: https://openaccess.thecvf.com/content_cvpr_2017/html/Isola_Image-To-Image_Translation_With_CVPR_2017_paper.html.

[8] G. Yang, S. Yu, H. Dong, G. Slabaugh, P. L. Dragotti, X. Ye, F. Liu, S. Arridge, J. Keegan, Y. Guo, et al., Dagan: deep de-aliasing generative adversarial networks for fast compressed sensing mri reconstruction, IEEE transactions on medical imaging 37 (2017) 1310–1321. doi:https://doi.org/10.1109/TMI.2017.2785879.

[9] D. P. Kingma, M. Welling, An introduction to variational autoencoders 12 (2019) 307–392. doi:10.1561/2200000056.

[10] Z. Rguibi, A. Hajami, D. Zitouni, M. Yassine, A. Elqaraoui, Medical variational autoencoder and generative adversarial network for medical imaging 32 (2023) 1–1x. doi:10.11591/ijeecs.v32.i1.pp494-505.

[11] Y. Shi, A. Abulizi, H. Wang, K. Feng, N. Abudukelimu, Y. Su, H. Abudukelimu, Diffusion models for medical image computing: A survey 30 (2025) 357–383. doi:10.26599/TST.2024.9010047.

[12] G. Müller-Franzes, J. M. Niehues, F. Khader, S. T. Arasteh, C. Haarburger, C. Kuhl, T. Wang, T. Han, T. Nolte, S. Nebelung, et al., A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis, Scientific reports 13 (2023) 12098. doi:https://doi.org/10.1038/s41598-023-39278-0.

[13] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 6840–6851. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, 2022. doi:10.48550/arXiv.2112.10752. arXiv:2112.10752 [cs].

[15] W. Wang, J. Xia, G. Luo, S. Dong, X. Li, J. Wen, S. Li, Diffusion model for medical image denoising, reconstruction and translation 124 (2025) 102593. doi:10.1016/j.compmedimag.2025.102593.

[16] S. Chataut, S. Bhatta, B. Dahal, G. Ojha, S. Raut, B. Subedi, B. Bastakoti, Generative artificial intelligence in healthcare: A systematic review of GANs, diffusion models, large language models, and variational autoencoders for medical applications 3 (2025) 182–203. doi:10.59324/ejaset.2025.3(4).16.

[17] A. F. Fercher, W. Drexler, C. K. Hitzenberger, T. Lasser, Optical coherence tomography-principles and applications, Reports on progress in physics 66 (2003) 239–303. doi:10.1088/0034-4885/66/2/204.

[18] M. A. Mayer, J. Hornegger, C. Y. Mardin, R. P. Tornow, Retinal nerve fiber layer segmentation on fd-oct scans of normal subjects and glaucoma patients, Biomedical optics express 1 (2010) 1358–1383. doi:https://doi.org/10.1364/BOE.1.001358.

[19] S. Morales, A. Colomer, J. M. Mossi, R. del Amor, D. Woldbye, K. Klemp, M. Larsen, V. Naranjo, Retinal layer segmentation in rodent OCT images: Local intensity profiles & fully convolutional neural networks 198 (2021) 105788. doi:10.1016/j.cmpb.2020.105788.

[20] R. D. Amor, S. Morales, A. N. Colomer, J. M. Mossi, D. Woldbye, K. Klemp, M. Larsen, V. Naranjo, Towards automatic glaucoma assessment: An encoder-decoder CNN for retinal layer segmentation in rodent OCT images, in: 2019 27th European Signal Processing Conference (EUSIPCO), IEEE, 2019, pp. 1–5. doi:10.23919/EUSIPCO.2019.8902794.

[21] R. S. Allen, K. Bales, A. Feola, M. T. Pardue, In vivo structural assessments of ocular disease in rodent models using optical coherence tomography, Journal of visualized experiments: JoVE (2020) 10–3791. doi:10.3791/61588.

[22] T. Tschernig, L. Thrane, T. M. Jørgensen, J. Thommes, R. Pabst, T. Yelbuz, An elegant technique for ex vivo imaging in experimental research—optical coherence tomography (oct), Annals of Anatomy-Anatomischer Anzeiger 195 (2013) 25–27. doi:https://doi.org/10.1016/j.aanat.2012.07.005.

[23] A. Barroso, P. Heiduschka, G. Nettels-Hackert, S. Ketelhut, R. d. Amor, F. García-Torres, S. Morales-Martínez, V. Naranjo, B. Kemper, J. Schnekenburger, Durable ex vivo mouse retina 3d tissue models for optical coherence tomography, in: Label-free Biomedical Imaging and Sensing (LBIS) 2024, volume 12854, SPIE, 2024, pp. 21–23. doi:10.1117/12.3002538.

[24] F. García-Torres, R. del Amor, S. Morales-Martínez, A. Barroso, B. Kemper, J. Schnekenburger, V. Naranjo, Using diffusion models for data augmentation on limited rodent oct datasets, in: Intelligent Data Engineering and Automated Learning – IDEAL 2024: 25th International Conference, Valencia, Spain, November 20–22, 2024, Proceedings, Part I, Springer-Verlag, Berlin, Heidelberg, 2024, p. 313–324. doi:10.1007/978-3-031-77731-8_29.

[25] Y. Tian, A. McCarthy, G. Gomide, N. Liddle, J. Golebka, R. Chen, J. Liebmann, K. Thakoor, OCTDiff: Bridged diffusion model for portable OCT super-resolution and enhancement 38 (2026) 41445–41465.

[26] N. Abbasi, A. Wong, K. Bizheva, A physics-informed diffusion model for super-resolved reconstruction of optical coherence tomography data 72 (2025) 2937–2946. doi:10.1109/TBME.2025.3556794.

[27] H. Ahmed, Q. Zhang, R. Donnan, A. Alomainy, Denoising of optical coherence tomography images in ophthalmology using deep learning: A systematic review 10 (2024) 86. doi:10.3390/jimaging10040086.

[28] S. Li, M. Dan, Y. Chu, J. Yu, Y. Zhao, P. Zhao, RetiDiff: Diffusion-based synthesis of retinal OCT images for enhanced segmentation, in: J. C. Gee, D. C. Alexander, J. Hong, J. E. Iglesias, C. H. Sudre, A. Venkataraman, P. Golland, J. H. Kim, J. Park (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2025, Springer Nature Switzerland, 2026, pp. 516–525. doi:10.1007/978-3-032-04937-7_49.

[29] Y. Wu, W. He, D. Eschweiler, N. Dou, Z. Fan, S. Mi, P. Walter, J. Stegmaier, Retinal OCT synthesis with denoising diffusion probabilistic models for layer segmentation, in: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), 2024, pp. 1–5. doi:10.1109/ISBI56570.2024.10635836.

[30] K. Du, U. Doshi, B. DiCenzo, J. Jiang, E. Wu, A. Gadari, S. C. Vupparaboina, E. Sadeghi, S. C. Bollepalli, J.-A. Sahel, J. Chhablani, K. K. Vupparaboina, Benchmarking diffusion models against state-of-the-art architectures for OCT fluid biomarker segmentation 20 (2025) e0335615. doi:10.1371/journal.pone.0335615.

[31] K. Huang, X. Ma, Y. Zhang, N. Su, S. Yuan, Y. Liu, Q. Chen, H. Fu, Memory-efficient high-resolution OCT volume synthesis with cascaded amortized latent diffusion models, in: M. G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, J. A. Schnabel (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2024, Springer Nature Switzerland, 2024, pp. 478–487. doi:10.1007/978-3-031-72104-5_46.

[32] R. H. Badhon, A. C. Thompson, J. I. Lim, T. Leng, M. N. Alam, Diffusion model based OCT to OCTA translation 12 (2025). doi:10.3389/fmed.2025.1655453.

[33] B. Yang, J. Li, J. Wang, R. Li, K. Gu, B. Liu, DiffusionDCI: A novel diffusion-based unified framework for dynamic full-field OCT image generation and segmentation 12 (2024) 37702–37714. doi:10.1109/ACCESS.2024.3372863.

[34] A. Toker, M. Eisenberger, D. Cremers, L. Leal-Taixé, SatSynth: Augmenting image-mask pairs through diffusion models for aerial semantic segmentation, 2024, pp. 27695–27705. URL: https://openaccess.thecvf.com/content/CVPR2024/html/Toker_SatSynth_Augmenting_Image-Mask_Pairs_through_Diffusion_Models_for_Aerial_Semantic_CVPR_2024_paper.html.

[35] J. Mao, Y. Wang, Y. Tang, D. Xu, K. Wang, Y. Yang, Z. Zhou, Y. Zhou, MedSegFactory: Text-guided generation of medical image-mask pairs, 2025, pp. 21525–21535. URL: https://openaccess.thecvf.com/content/ICCV2025/html/Mao_MedSegFactory_Text-Guided_Generation_of_Medical_Image-Mask_Pairs_ICCV_2025_paper.html.

[36] Y. Frisch, C. Bornberg, M. Fuchs, A. Mukhopadhyay, GAUDA: Generative adaptive uncertainty-guided diffusion-based augmentation for surgical segmentation, in: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2025, pp. 3762–3771. doi:10.1109/WACV61041.2025.00370, ISSN: 2642-9381.

[37] W. Wu, Y. Zhao, M. Z. Shou, H. Zhou, C. Shen, DiffuMask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models, 2023, pp. 1206–1217. URL: https://openaccess.thecvf.com/content/ICCV2023/html/Wu_DiffuMask_Synthesizing_Images_with_Pixel-level_Annotations_for_Semantic_Segmentation_Using_ICCV_2023_paper.html.

[38] Z. Li, Q. Zhou, X. Zhang, Y. Zhang, Y. Wang, W. Xie, Open-vocabulary object segmentation with diffusion models, 2023, pp. 7667–7676. URL: https://openaccess.thecvf.com/content/ICCV2023/html/Li_Open-vocabulary_Object_Segmentation_with_Diffusion_Models_ICCV_2023_paper.html.

[39] J. H. Park, K. Jo, S. Baik, SeeDiff: Off-the-shelf seeded mask generation from diffusion models 39 (2025) 6406–6415. doi:10.1609/aaai.v39i6.32686.

[40] stabilityai/sd-vae-ft-ema · hugging face, 2022. URL: https://huggingface.co/stabilityai/sd-vae-ft-ema.

[41] W. Peebles, S. Xie, Scalable diffusion models with transformers, 2023, pp. 4195–4205. URL: https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.

[42] A. Barroso, S. Ketelhut, G. Nettels-Hackert, P. Heiduschka, R. d. Amor, V. Naranjo, B. Kemper, J. Schnekenburger, Durable 3d murine ex vivo retina glaucoma models for optical coherence tomography 14 (2023) 4421–4438. doi:10.1364/BOE.494271.

[43] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2017. URL: http://arxiv.org/abs/1412.6980. doi:10.48550/arXiv.1412.6980. arXiv:1412.6980 [cs.LG].

[44] J. Ho, T. Salimans, Classifier-free diffusion guidance, 2022. URL: https://arxiv.org/abs/2207.12598. arXiv:2207.12598.

[45] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs trained by a two time-scale update rule converge to a local nash equilibrium, in: Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

[46] P. Dhariwal, A. Nichol, Diffusion models beat GANs on image synthesis, 2021. doi:10.48550/arXiv.2105.05233. arXiv:2105.05233 [cs].

[47] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W. M. Wells, A. F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, 2015, pp. 234–241. doi:10.1007/978-3-319-24574-4_28.

[48] J. Ma, J. Chen, M. Ng, R. Huang, Y. Li, C. Li, X. Yang, A. L. Martel, Loss odyssey in medical image segmentation, Medical image analysis 71 (2021) 102035. doi:https://doi.org/10.1016/j.media.2021.102035.

[49] R. del Amor, A. Colomer, S. Morales, C. Pulgarín-Ospina, L. Terradez, J. Aneiros-Fernandez, V. Naranjo, A self-contrastive learning framework for skin cancer detection using histological images, in: 2022 IEEE International Conference on Image Processing (ICIP), 2022, pp. 2291–2295. doi:10.1109/ICIP46576.2022.9897435.

[50] D. Müller, I. Soto-Rey, F. Kramer, Towards a guideline for evaluation metrics in medical image segmentation, BMC research notes 15 (2022) 210. doi:https://doi.org/10.1186/s13104-022-06096-y.

A. Qualitative Results of Joint Synthetic Image and Mask Generation

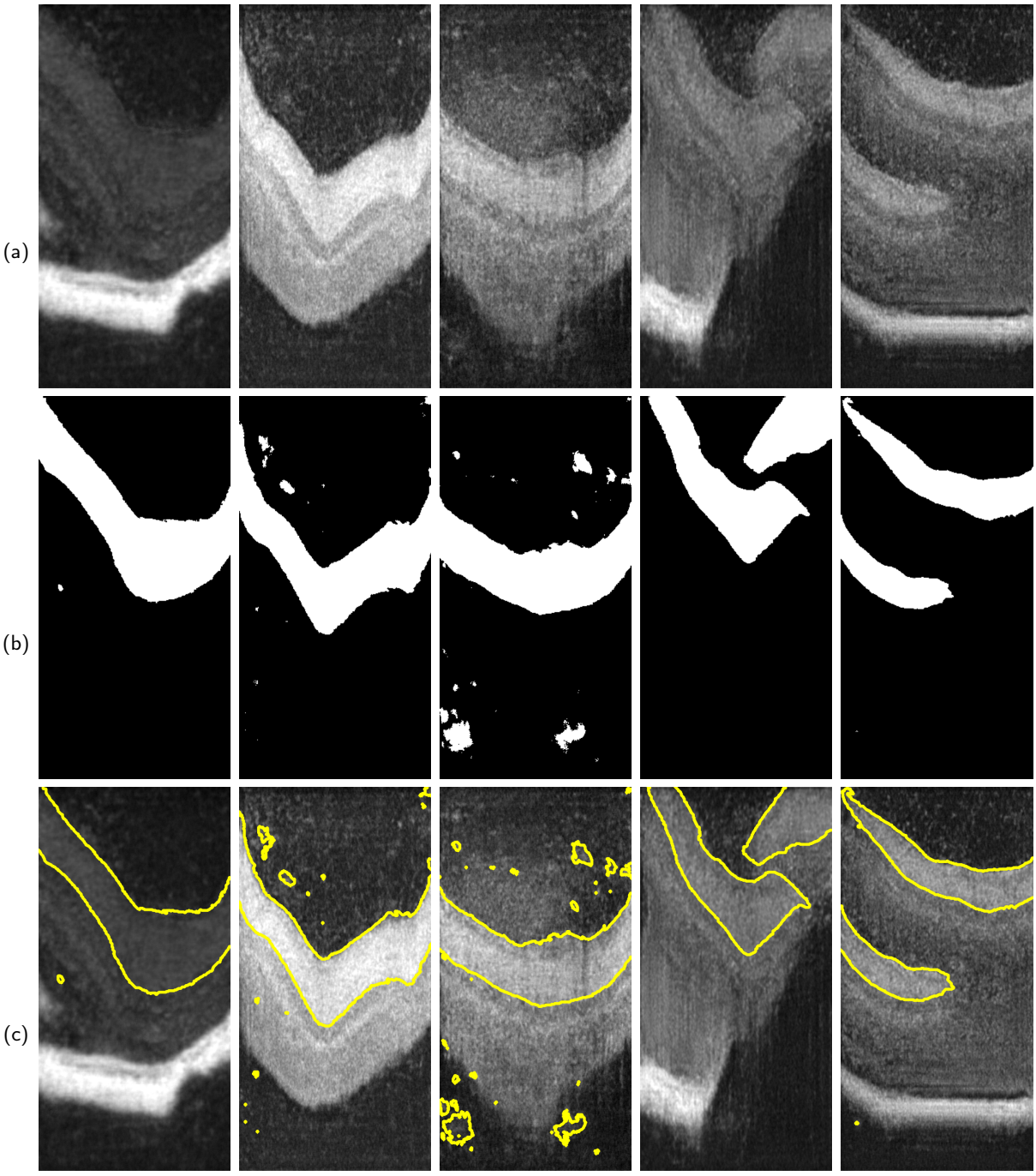


Figure A.1: Synthetic resin-embedded images generated with DDPM [13]. (a) Images; (b) masks; (c) overlaying the edges of the mask onto the image.

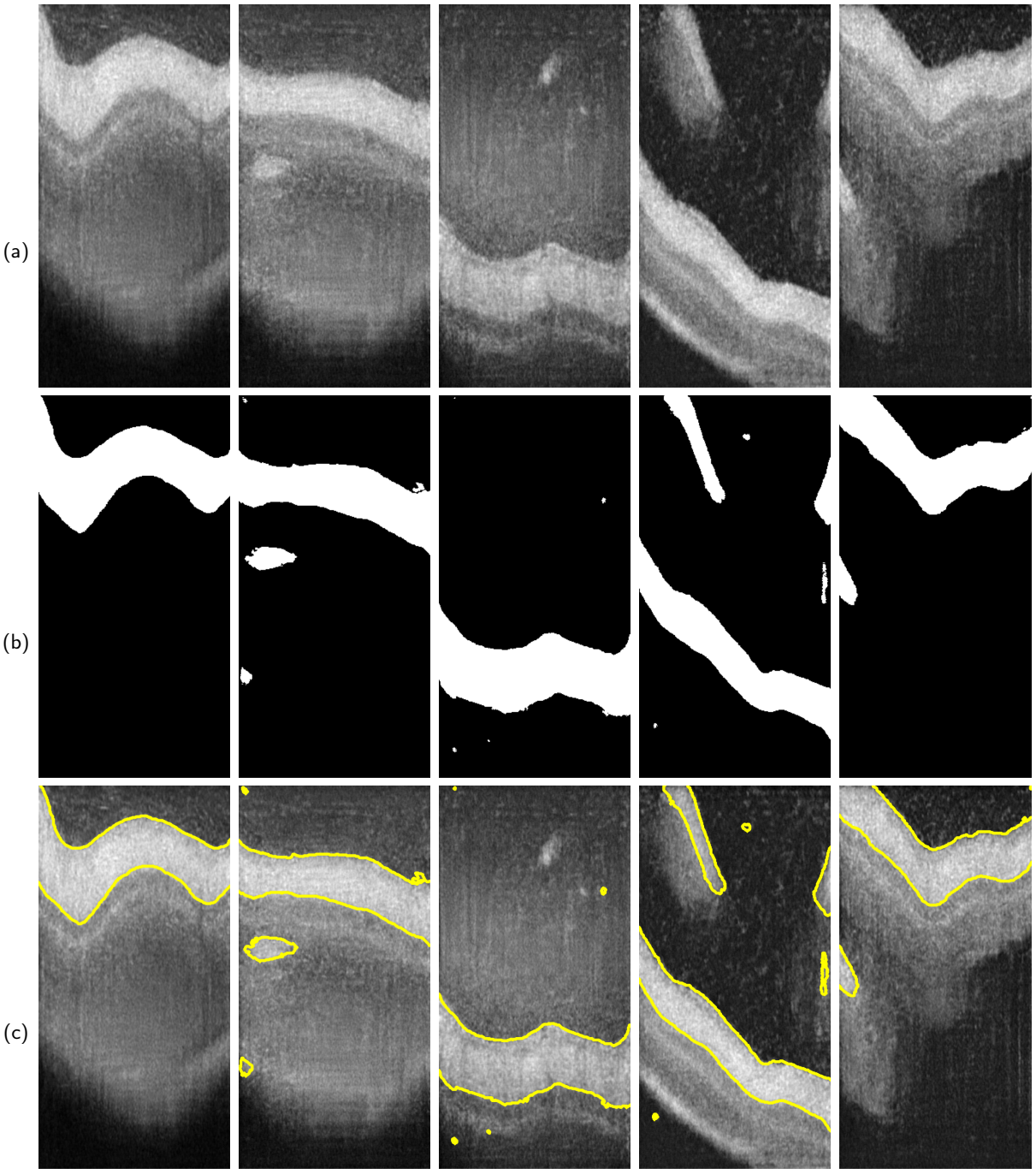


Figure A.2: Synthetic fluid-embedded images generated with DDPM [13]. (a) Images; (b) masks; (c) overlaying the edges of the mask onto the image.

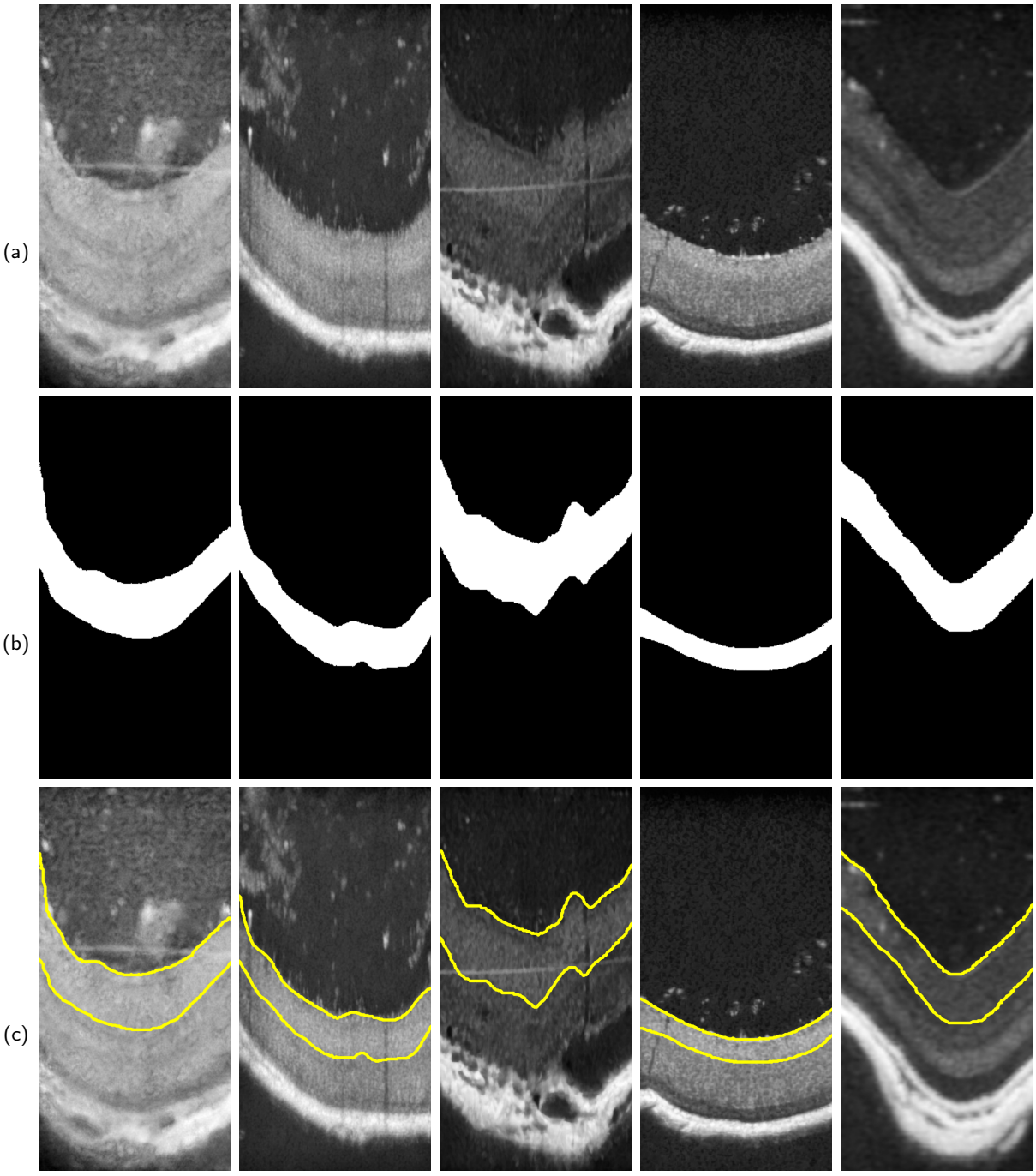


Figure A.3: Synthetic resin-embedded images generated with LDM [14]. (a) Images; (b) masks; (c) overlaying the edges of the mask onto the image.

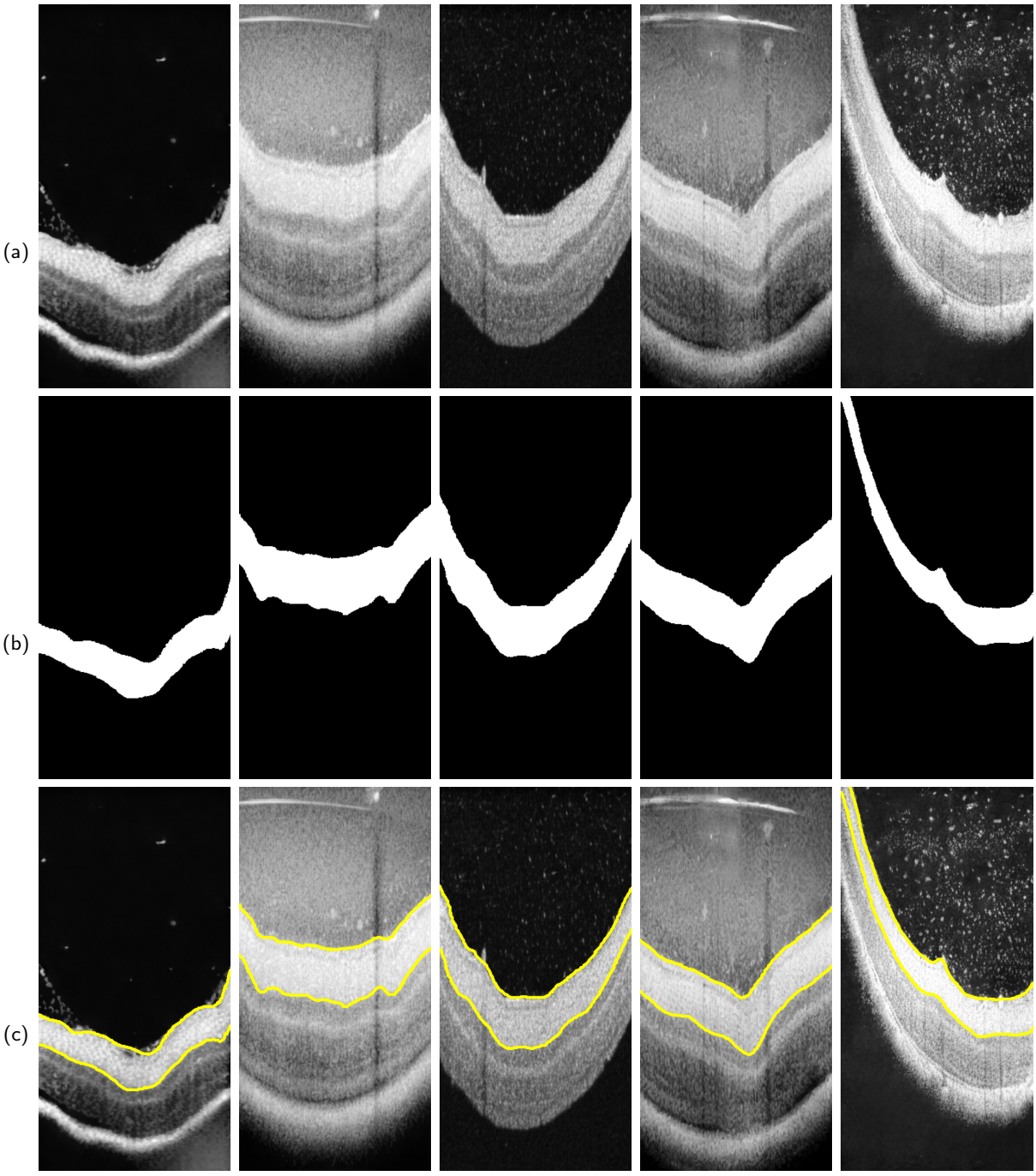


Figure A.4: Synthetic fluid-embedded images generated with LDM [14]. (a) Images; (b) masks; (c) overlaying the edges of the mask onto the image.

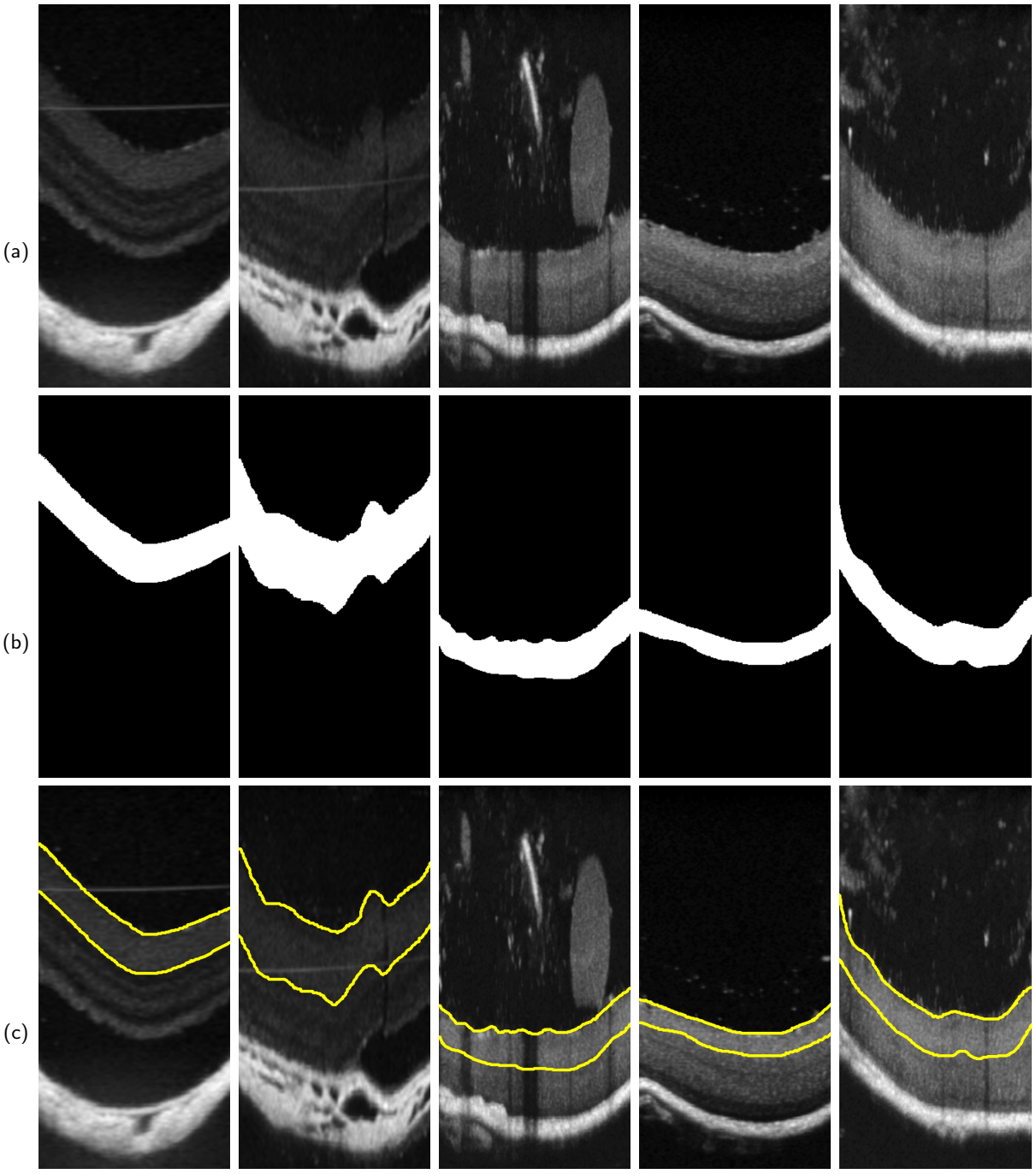


Figure A.5: Synthetic resin-embedded images generated with DualDiT. (a) Images; (b) masks; (c) overlaying the edges of the mask onto the image.

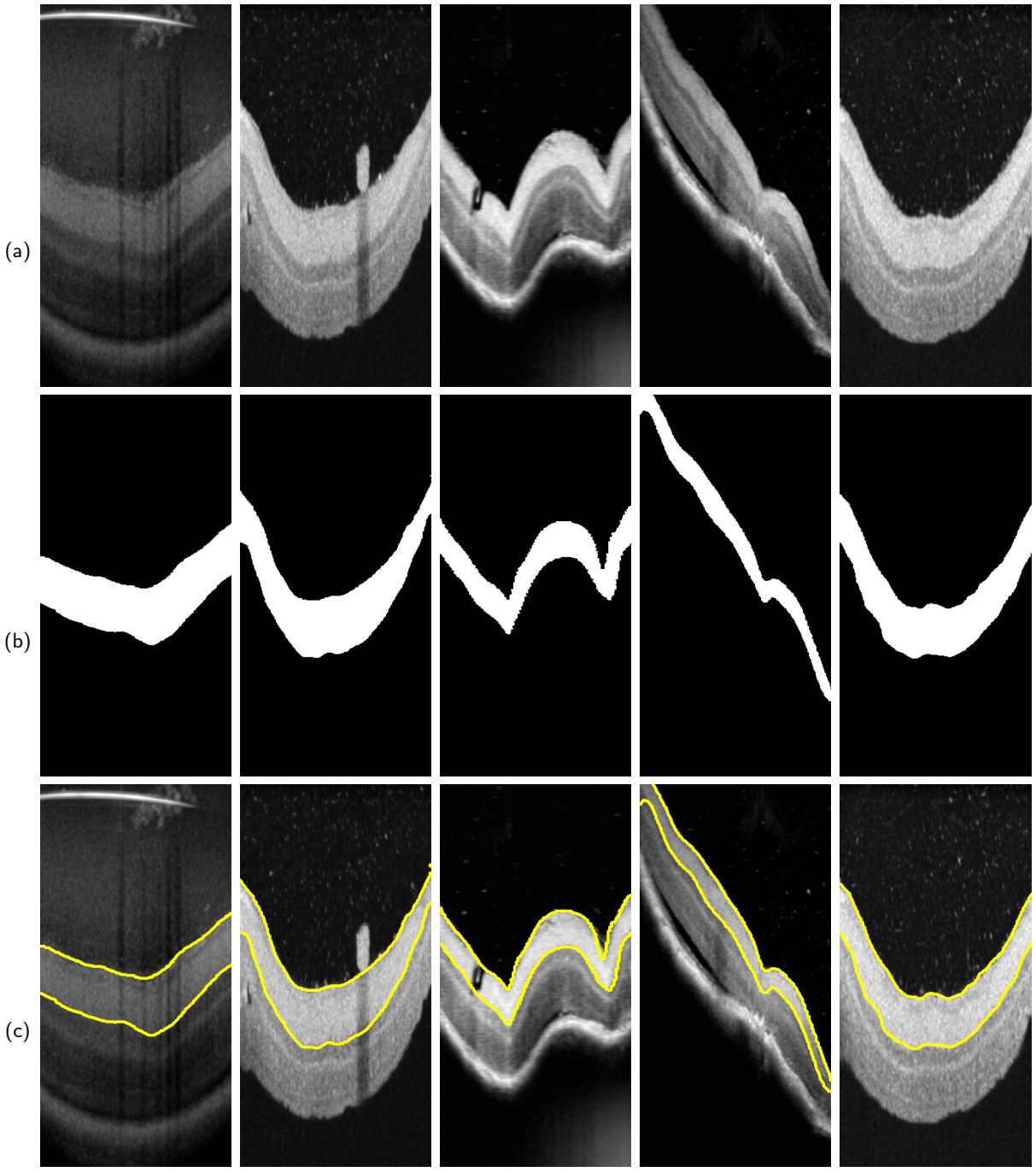
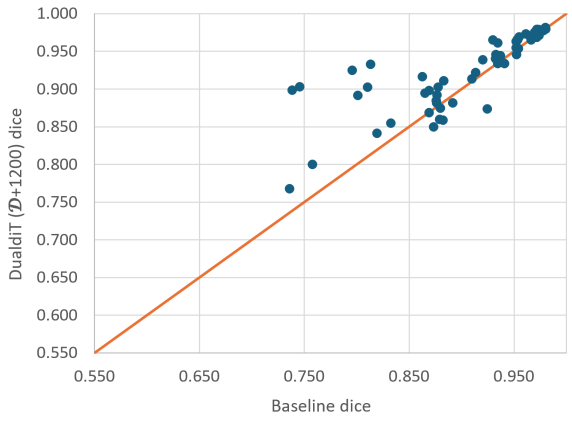
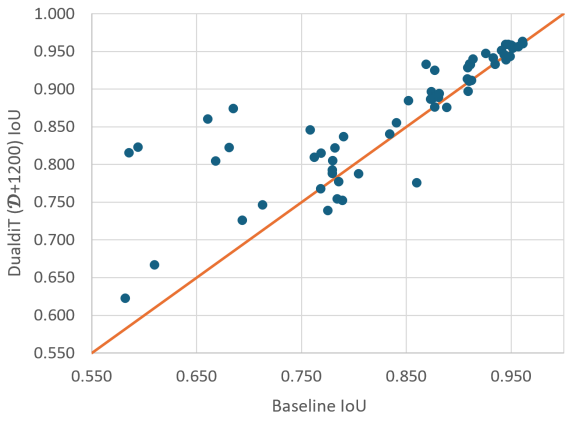


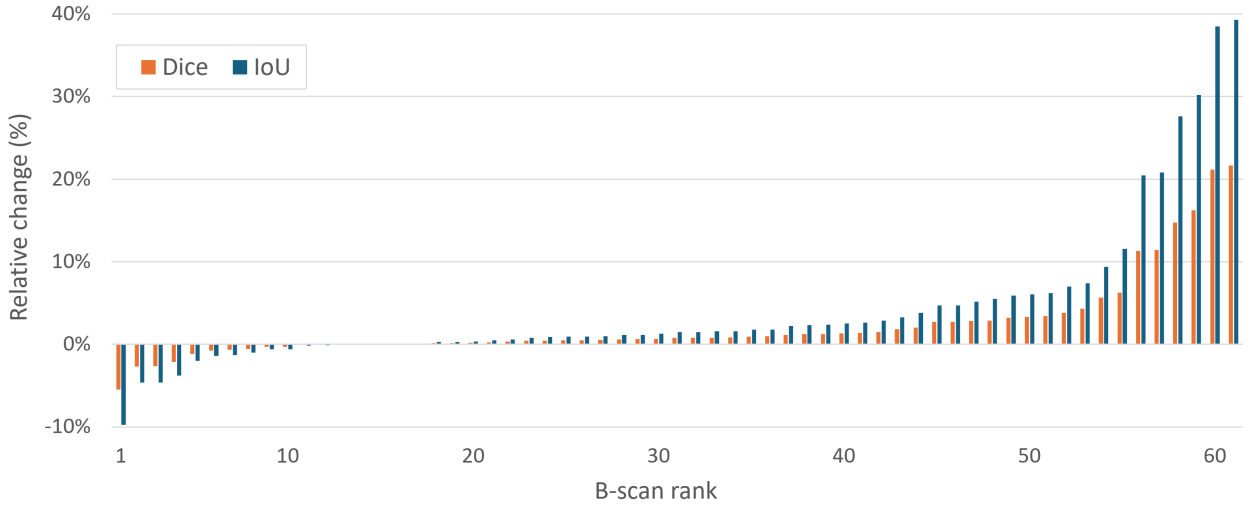
Figure A.6: Synthetic fluid-embedded images generated with DualDiT. (a) Images; (b) masks; (c) overlaying the edges of the mask onto the image.



(a) Per-B-scan Dice scores obtained with the Baseline and DualDiT trained with $\mathcal{D}+1200$. The diagonal line denotes equal performance; points above it indicate higher Dice scores for DualDiT.



(b) Per-B-scan IoU scores obtained with the Baseline and DualDiT trained with $\mathcal{D}+1200$. The diagonal line denotes equal performance; points above it indicate higher IoU scores for DualDiT.



(c) Relative change in Dice and IoU obtained with DualDiT with respect to the Baseline for each test B-scan ($n = 61$), ordered by increasing IoU change. Positive values indicate an improvement, whereas negative values indicate a decrease in performance.

Figure A.7: Per-B-scan comparison between the Baseline and DualDiT trained with $\mathcal{D}+1200$. The scatter plots show the paired Dice and IoU scores for the complete test set, with the identity line representing equal performance. The bar chart shows the relative percentage change produced by DualDiT for each B-scan, revealing that performance improves for most samples, although decreases are observed in a small number of cases.