SYNOPHSIS: UNIFIED OPHTHALMIC IMAGE AND VOLUME GENERATION ACROSS MODALITIES AND DIMENSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

An ophthalmic examination records several 2D modalities, a volumetric OCT scan at one of two slab thicknesses, the en-face image registered to it, and device and free-text clinical metadata. Assessment rests on more than one of them. Generative models of this evidence are built one modality or one dimensionality at a time, so covering an examination takes a collection of models with no shared representation, and not even the volume and the en-face image acquired with it are generated together. We present SynOphsis, a unified generative model for ophthalmic imaging. A frozen autoencoder maps every input into one latent space and per-shape patch embedders place images and volumes on a common token grid, so a volume and the en-face view acquired with it occupy one self-attention sequence, which a single stochastic-interpolant transformer denoises. Conditioning is a mask over that sequence, so any subset of the evidence in either direction is the same operation on the same weights, and one set of weights covers eleven 2D modalities, all three volume protocols, sparse-observation completion, and generation conditioned on a paired image, modality and device, with text conditioning shown qualitatively. The pair is sampled along a single trajectory and either half may be generated from the other, a mode no prior ophthalmic generator provides. One checkpoint serves every mode we evaluate, with no task-specific head and no fine-tuning, while each conditioning baseline carries a mechanism built for one task family and several modes are outputs those architectures cannot emit. On the tasks those specialists do support, the broader interface keeps their volumetric reconstruction accuracy, and joint tokenization costs under 0.01% of the parameters. Separately, our SiT-based 3D stack improves over the retrained volumetric latent-diffusion stack by 3.28 dB PSNR under matched observation conditioning, and on held-out OCT volumes SynOphsis-XL takes FVD from that baseline's 56.2 to 32.9, against a floor of 29.5 measured between two disjoint halves of the real data.

1 INTRODUCTION

A single ophthalmic examination views one eye in several ways. Color fundus photography, autofluorescence, angiography, and infrared reflectance each record a different 2D projection of the retina; OCT records a volume under a thin protocol of 60 B-scans, a thick one of 240, or as OCT angiography. Each volume has a registered en-face image recorded with it, the device and field of view that produced it, and a free-text clinical description. The clinical question and the instruments available decide which other modalities accompany them, and understanding the eye draws on several together. A generative model must span them and produce them together, to augment a scarce modality or complete an interrupted acquisition.

Generative modeling has gone the other way, one modality at a time. Fundus photographs, OCT B-scans, and OCT volumes each have their own literature, as does each translation between a pair. The conditioning those models accept is typically a mask or a grade of the image being generated. Assembling an examination takes one model per modality and one per translation direction, each

Code and trained models will be released on acceptance.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

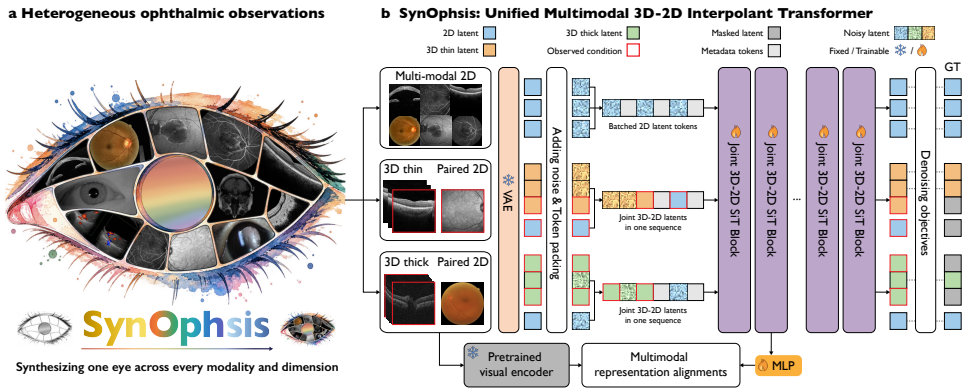

Figure 1: **One eye, every modality and dimension.** **(a)** One eye, observed through eleven 2D modalities, OCT and OCTA volumes at two depths, and free-text and device metadata; existing models are fragmented across this space, one per modality or dimensionality. **(b)** SynOphsis is one backbone over all of it. A frozen VAE encodes every input, per-shape patch embedders bring the results onto a common token grid, and a paired 2D view is the last token of that sequence, so one stack of SiT blocks denoises the joint sequence in every case.

trained on its own and sharing no representation with the rest. As a result, every additional modality is an additional model, and pairing between modalities is out of reach. An instrument records a volume and the 2D image beside it as one event, and no single-modality model reproduces that.

SynOphsis covers an examination with one model. A frozen autoencoder maps every modality into one latent space, per-shape patch embedders place 2D images and volumes at both thicknesses on one token grid, and a volume's paired 2D view is written into the *same* self-attention sequence as the volume. Thus, inputs of different raw dimensionality become sequences differing only in length and type embeddings. A single Scalable Interpolant Transformer (Ma et al., 2024) denoises whatever sequence it is given, and a volume and its en-face view are produced along one trajectory, neither derived from the other. Conditioning is a mask over the latent values in that sequence, defined on the sequence itself and not on an encoder built for a particular input, so evidence in any combination and in either direction enters the model the same way. Changing what the model is given changes the mask and nothing else, and other work can be built on the model instead of alongside it.

Method	2D	3D	Sparse 3D	2D→3D	Sparse 3D + 2D	3D→2D	Paired	One backbone
2D LDM	✓	/	/	/	/	✓	/	/
3D LDM	/	✓	✓	✓	/	/	/	/
SiT-B ControlNet	/	✓	✓	/	/	/	/	/
SiT-B Cross-attn Inpaint	/	✓	✓	✓	✓	/	/	/
SynOphsis	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Generation modes each method supports. SynOphsis is the only row with all eight.

Eight generation modes follow from that one interface. With no task-specific head, the same weights unconditionally generate eleven 2D modalities, draw a volume together with its paired en-face view at both slab thicknesses, and complete a volume from a few observed B-scans with its paired view alongside. Figure 2 shows those and a text-conditioned image. No published method covers more than three of them (Table 1), so every comparison here is against a different method.

Contributions. We present SynOphsis, a unified generative model for ophthalmic imaging. A single stochastic-interpolant transformer over modality-aware tokens covers an examination with one set of weights, and a masking interface takes whatever evidence is available. The formulation is heterogeneous multimodal generation as masked completion over a shared token sequence, where the observations differ in modality, in dimensionality and in acquisition geometry, and any subset of them may act as evidence or as target without changing the model. An ophthalmic examination makes this concrete, since one visit produces all of them and which exist varies from eye to eye.

A joint 3D–2D tokenization places a volume and the 2D view acquired with it in one self-attention sequence, so the pair is sampled along a single trajectory and either half may be generated from the other. It adds three embedding tables, 0.002% of the XL backbone's parameters. This is a general-

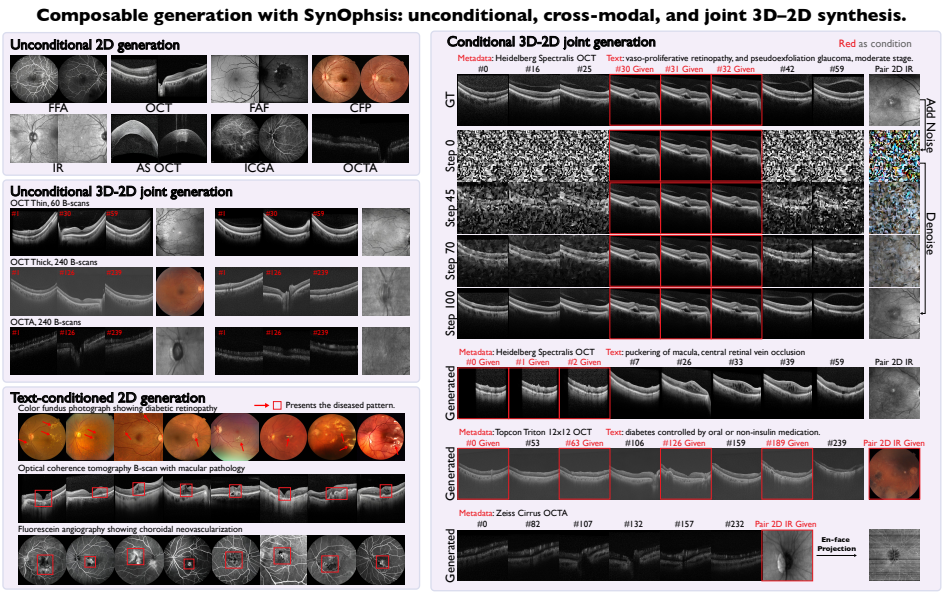

Figure 2: **The modes one model composes into.** Every panel is the same SynOphsis model under a different conditioning pattern, with no task-specific head or fine-tuning. **Unconditional 2D:** eight modalities. **Unconditional joint 3D–2D:** a volume and its paired en-face view drawn together, at 60 and 240 B-scans and for OCTA. **Text-conditioned 2D:** from a clinical description alone. **Conditional joint 3D–2D:** a few observed B-scans (red) with device and text metadata are given and the rest generated, the upper block showing the trajectory and the lower rows device and budget.

ist rather than a collection of task-specific models. The same checkpoint serves every conditioning pattern and only the observation mask changes, while the variants we compare against carry a mechanism built for one task family. On the tasks they support it keeps their volumetric reconstruction accuracy, and the same weights emit the paired view and take evidence combinations they have no output for. Underneath it we extend the interpolant backbone to 2D images and to volumes at two slab thicknesses over one token grid, training it jointly across all three. That stack improves over the retrained volumetric latent-diffusion baseline by 3.28 dB PSNR under matched conditioning, and SynOphsis-XL takes FVD on held-out OCT volumes from the baseline's 56.2 to 32.9 (Section 5).

2 Related Work

Ophthalmic image generation. The ophthalmic generation literature is organized by modality. In 2D the conditioning is derived from the image being generated, as a vessel mask, a layer map, a severity grade, a caption, or the gradient of a disease classifier (Go et al., 2024; Ning et al., 2025; Pakdelmoez et al., 2024; Wu et al., 2024c; Ilanchezian et al., 2025). Two recent models cover many 2D modalities at once. EyeDiff fine-tunes a text-conditioned natural-image latent diffusion model (Chen et al., 2026b) on 42,048 images spanning fourteen ophthalmic modalities, and a retinal generative model produces the angiographic sequence as video from one fundus photograph (Zhang et al., 2026). Both work in 2D throughout, and neither emits a volume. Our 2D training set covers the same breadth across 4.75 million images (Appendix F), and we align against an encoder pretrained on ophthalmic data. In 3D the closest work is CA-LDM (Huang et al., 2024), cascaded latent diffusion for high-resolution OCT volumes, built on Medical Diffusion's volumetric autoencoder. CA-LDM releases no weights and its published model is fitted to a different corpus, so we compare against Medical Diffusion retrained on our data (Appendix E); neither produces the 2D image acquired alongside the volume. A separate line translates between modalities (Tavakkoli et al., 2020; Shi et al., 2023; Chen et al., 2024), taking an image that already exists and producing another from it (Appendix E). Each fixes in advance which modality it starts from and which it produces.

Joint generation elsewhere. Generating several outputs at once is established outside ophthalmology, and so is breadth across modalities, where MedM2G generates CT, MRI and X-ray for other

organs within a single model (Zhan et al., 2024) and MINIM spans OCT, fundus photography, chest X-ray and chest CT from text (Wang et al., 2025). MedSyn (Xu et al., 2024) samples a CT volume together with its anatomical structure volumes, a slice-based latent diffusion model samples multi-modal brain MRI together with its multi-label masks (Kebaili et al., 2025), and DualDiT samples an OCT B-scan with its layer mask from one concatenated latent (García-Torres et al., 2026). In all three, every output lies on the same grid as the others, an annotation or another contrast of the same scan, so one shared representation is immediate. A volume and the 2D view acquired beside it differ in dimensionality and in geometry. Volumes have also been produced from a 2D backbone. Make-A-Volume (Zhu et al., 2023) inserts volumetric layers into a trained 2D slice-wise model, adapting it after the fact to one volume shape. de Vente et al. (de Vente et al., 2024) are nearest in setting, conditioning a 3D diffusion model on the en face image to *upsample* volumes. Volumes have also been generated from a color fundus photograph (Liu et al., 2025a;b), a separate acquisition from the view recorded beside the volume. The two views do inform each other there, but the en face is an input, and refining an existing volume is a different problem from producing both from noise. That work is direct prior art for our 2D-only task. No implementation is released, so we compare against our own cross-attention variant (Appendix E). On the analysis side, OCTCube-M (Liu et al., 2026) and OphMAE (Chang et al., 2026) represent volumes and their en-face images jointly and benefit from doing so; generation has no counterpart. No published model, to our knowledge, samples an ophthalmic volume and its separately acquired 2D view along a single trajectory.

3D medical generation, control, and the backbone we build on. Latent diffusion (Rombach et al., 2022) is the standard route to 3D synthesis. Medical Diffusion (Khader et al., 2023) applies it to several volumetric modalities and is our 3D baseline; MAISI (Guo et al., 2025) scales to whole-body CT with a ControlNet branch over organ masks, and MAISI-v2 (Zhao et al., 2026) swaps in rectified flow. ControlNet (Zhang et al., 2023a), or SparseCtrl (Guo et al., 2024) for sparse temporal conditioning, attaches control to a frozen backbone rather than building it in. We run both as internal baselines, holding the backbone fixed and varying only how conditioning enters. For the backbone, DiT (Peebles & Xie, 2023) replaced the convolutional U-Net with a transformer over latent patches, establishing that diffusion inherits transformer scaling; stochastic interpolants (Albergo et al., 2025) and flow matching (Lipman et al., 2023) recast diffusion as continuous-time transport, and SiT (Ma et al., 2024) combines the two. SiT is a 2D model, with a patch embedding and positional scheme that assume one image and an interpolant over a single latent tensor. We extend it to volumes and mixed-shape sequences, so every SiT row in our tables is that extension and not published SiT.

Representation alignment and domain-specific encoders. REPA (Yu et al., 2025) showed that aligning intermediate activations of a diffusion transformer to a frozen self-supervised encoder such as DINOv2 (Oquab et al., 2024) accelerates training substantially. In a specialized domain the target matters more, and ophthalmology has its own encoders. RETFound (Zhou et al., 2023) is a masked autoencoder pretrained on 1.6 million retinal images; FLAIR (Silva-Rodríguez et al., 2025) a vision–language model trained on fundus images with expert text; MIRAGE (Morano et al., 2025) trained on OCT with scanning laser ophthalmoscopy; and OCTCube-M (Liu et al., 2026) the only one whose representation is volumetric rather than per-slice. We follow REPA but align to a DINOv3 (Siméoni et al., 2025) encoder pretrained on ophthalmic data, and extend alignment to volume tokens by applying it to groups of neighboring B-scans. The same encoder supplies the domain-specific features in our metrics, so its pretraining corpus is in Appendix D.1.

3 METHOD

3.1 OVERVIEW

We model an examination in one generative framework instead of one model per modality or dimensionality. Every sample is encoded to a latent, turned into 2D, 3D or paired 3D–2D tokens, and denoised by a shared Scalable Interpolant Transformer (SiT) (Ma et al., 2024) (Figure 1b).

Per-shape patch embedders map 2D images and volumes of 60 and 240 B-scans onto one token grid, so one backbone serves all three shapes and volumetric structure is in the sequence from the first step. Because the sequence admits tokens of different dimensionality, paired 2D and volumetric tokens occupy the *same* self-attention stream with no intermediate encoding between them, and the two are drawn on one trajectory, neither derived from the other.

Interpolant formulation. Let z_0 be the latent of a sample x. For $t \in [0,1]$ with signal and noise coefficients α_t, σ_t and $\epsilon \sim \mathcal{N}(0, I)$, we take $z_t = \alpha_t z_0 + \sigma_t \epsilon$ and, following SiT, train f_θ to predict the velocity target $u_t = \dot{\alpha}_t z_0 + \dot{\sigma}_t \epsilon$; the linear path used throughout gives $\alpha_t = 1 - t$, $\sigma_t = t$ and $u_t = \epsilon - z_0$. With $\mathcal{T}(\cdot)$ our joint tokenization, $\gamma(t, y)$ an adaptive vector formed from the diffusion time and structured metadata y, and c an optional text context, the prediction is

$$\hat{u}_\theta = f_\theta(\mathcal{T}(\bar{z}_t), \gamma(t, y), c), \tag{1}$$

where $\bar{z}_t = z_t$ unconditionally. Conditioning enters as a mask on the latent values of that same sequence. For a binary mask m, one for observed elements and zero for generated ones,

$$\tilde{z}_t = m \odot z_0 + (1 - m) \odot z_t, \tag{2}$$

so that $\bar{z}_t = \tilde{z}_t$. Observed tokens stay present and visible to self-attention, and the loss is evaluated only on the complement of m, so any subset of the evidence is a choice of m over the same weights.

The 3D data differs in acquisition density by a factor of four, so the thin and thick branches take temporal patch sizes one and four for the same $G_T = 20$ temporal token groups; with 32×32 latents and spatial patch size two, every branch shares the grid $G_H = G_W = 16$ and the width d. A 2D input contributes 256 tokens, a volume 5120, and a paired 2D view another 256. Sharing the grid lets branches of different raw dimensionality share positional embeddings and the backbone. It also sets the granularity of the sparse inpainting masks, over slice groups of three B-scans.

Each branch adds positional and type embeddings to a patch embedding, and paired 2D tokens carry an embedding for the paired-view modality and another for whether that view is observed or generated. Appendix A gives the token constructions and the latent shapes.

3.2 Joint Tokens, and Conditioning as a Mask over Them

All tokens pass through a shared SiT backbone with standard adaLN-Zero blocks, under SiT's additive conditioning $\gamma(t, y) = e_t(t) + e_y(y)$. We build e_y from discrete ophthalmic labels together with continuous acquisition metadata, in place of SiT's single class, and text enters through an optional cross-attention branch when it is available. For a paired input the sequence entering the backbone is

$$X^{\text{joint}} = [X^{3D,\tau}; X^p], \qquad \tau \in \{\text{thin}, \text{thick}\}. \tag{3}$$

Both portions keep their positional information and are updated together by self-attention, so the paired 2D observation enters the backbone as a token of the same sequence as the volume, on the same footing. We split the final sequence by that ordering; a 3D head and a shared 2D head predict the two velocities, the paired portion passing through that 2D head as an ordinary 2D sample.

Every conditional task in this paper is one choice of m, applied to 2D tokens, 3D latent slice tokens, and paired 2D tokens alike. Sparse 3D inpainting selects a subset of latent slices, paired-2D-conditioned generation marks the paired tokens observed, and joint generation diffuses both. One model therefore covers the whole task family, and because the paired 2D tokens occupy the same sequence none of those three needs an encoder of its own (Appendix B.2).

3.3 Alignment and Training Objective

Cross-dimensional representation alignment. Following REPA (Yu et al., 2025) we align an intermediate SiT representation of the noisy latent to a frozen ophthalmic DINOv3-Large encoder (Siméoni et al., 2025) on the clean sample. With s_i that projection normalized at patch i and q_i the normalized teacher feature under stop-gradient, the alignment term is the patchwise negative cosine similarity $-\mathbb{E}[(1/N) \sum_i s_i^\top q_i]$. For a volume, consecutive B-scans group in threes, so the 3D teacher is that architecture applied chunkwise under its own weights.

We add two terms to that objective. A **masked normalization** weights the patch average by a generated-token weight, so clean observed slices do not become alignment targets. The **cross-dimensional term** averages the projected 3D tokens over the height axis and aligns the resulting en-face feature map against the teacher's encoding of the paired image (Appendix A).

The objective. The primary term is the interpolant loss on generated tokens. With r a validity mask for non-padded slices, $a = r \odot (1 - m)$ the generated mask, i the index and $\delta > 0$ a stabilizer,

$$\mathcal{L}_{diff} = \mathbb{E}_{t,\epsilon}\Big[\textstyle\sum_i a_i \|\hat{u}_{\theta,i} - u_{t,i}\|^2 \big/ \big(\sum_i a_i + \delta \big) \Big], \tag{4}$$

over the branches a sample generates. The objective adds the alignment terms,

$$\mathcal{L} = \mathcal{L}_{diff} + \lambda_{same}\mathcal{L}_{rep}^{same} + \lambda_p\mathcal{L}_{rep}^{p} + \lambda_{cross}\mathcal{L}_{rep}^{3D\to 2D}, \tag{5}$$

where $\mathcal{L}_{rep}^{same}$ averages the active 2D and 3D terms, each against its own teacher's encoding. All three coefficients are 0.5, and λ_{cross} is inactive without a paired target (Appendix D.3).

4 Experimental Setup

Datasets. We train on two sources. AI-READI (AI-READI Consortium, 2024; Owsley et al., 2025) contributes mainly thick OCT volumes, and an **internal ophthalmology cohort**, a retrospective collection acquired at one academic eye center, contributes mainly thin ones. The 3D collection is thin and thick OCT standardized to 60 and 240 B-scans, and some volumes carry a 2D image from the same eye and visit, on which paired 3D–2D generation is trained. Provenance, ethics and data-use terms, and the modality composition, are in Appendix F.

Latent space, teacher and training. We train every model in the latent space of a frozen Stable Diffusion VAE (Rombach et al., 2022) at 256×256 (Appendix D.2). For volumes, every three neighboring B-scans form a three-channel VAE input encoded independently, so standardized 60- and 240-frame volumes become 20 and 80 latent slice groups. The representation-alignment teachers are ophthalmic DINOv3-Large encoders (Siméoni et al., 2025), one for 2D inputs and one applying the same architecture groupwise to volumes. Both scales share this recipe. The encoder's pretraining is in Appendix D.1 and the optimization settings and hardware in Appendix D.3, with the sampler and guidance choices in Appendix C.7.

Baselines. We compare against two kinds of baseline. *External baselines* are existing methods reimplemented on our data. They are an ophthalmic LDM for 2D (Rombach et al., 2022), whose conditional variant uses cross-attention, and a 3D LDM following Medical Diffusion (Khader et al., 2023) from its public code, with a ControlNet variant for sparse observed frames and a cross-attention variant for paired 2D. The latter stands for canonical single-stage volumetric latent diffusion, not CA-LDM's cascaded refinement, which no row reproduces (Appendices E and D.4).

Internal variants share our 3D SiT backbone and differ only in how conditioning attaches. SiT (Ma et al., 2024) is a 2D model, so the per-shape extension to volumes is ours (Section 3) and every SiT row in the tables is built on it. The SiT-B and SiT-XL rows of Table 2 are that extension with the joint tokens removed, trained on the same data under the same recipe. **SiT Ctrl** adds a separate 3D control branch injected through zero-initialized residual projections following SparseCtrl (Guo et al., 2024), and **SiT Inpaint** supplies the paired 2D condition as DINOv3 cross-attention context while holding selected latent slices clean. Both train at B scale. Neither produces a volume and its paired 2D image together; joint tokens add that. All conditional runs use the same selected OCT cases and inpainting presets as our model, and every method in Table 3 receives the same clinical text and acquisition metadata. The comparison differs only in how image and volume evidence enters. Every distributional metric is also computed between two disjoint halves of the real data, the best the metric allows at that sample size, and is reported beside each benchmark.

5 Results

5.1 Unconditional Generation

Before any conditioning, a single model has to hold eleven 2D modalities and three volume protocols without giving up quality on any of them. On the third, OCTA, it reproduces the vertical flow striation and the optic nerve head in the paired view and leads both LDM baselines. Table 2 scores every model in the standard feature spaces (Inception-v3 (Szegedy et al., 2016) for images, R3D-18 (Tran et al., 2018) for volumes) and in our ophthalmic DINOv3 space. In the Inception features every SiT-based model leads the LDM baseline on all five 2D OCT benchmarks, and in the ophthalmic features it does so in 19 of the 20 model-by-benchmark cells, the exception being SynOphsis-XL on OCTID. On internal B-scans SynOphsis-XL takes FID from the LDM's 66.8 to 29.9 and the ophthalmic distance from 4.17 to 1.24, against floors of 21.3 and 0.72. On OCTID SynOphsis-B reaches 101.5 from 125.7, and real data scores 36.6. OCTDL (Kulyabin et al., 2024) and OCTID are both external collections (Appendix B.1).

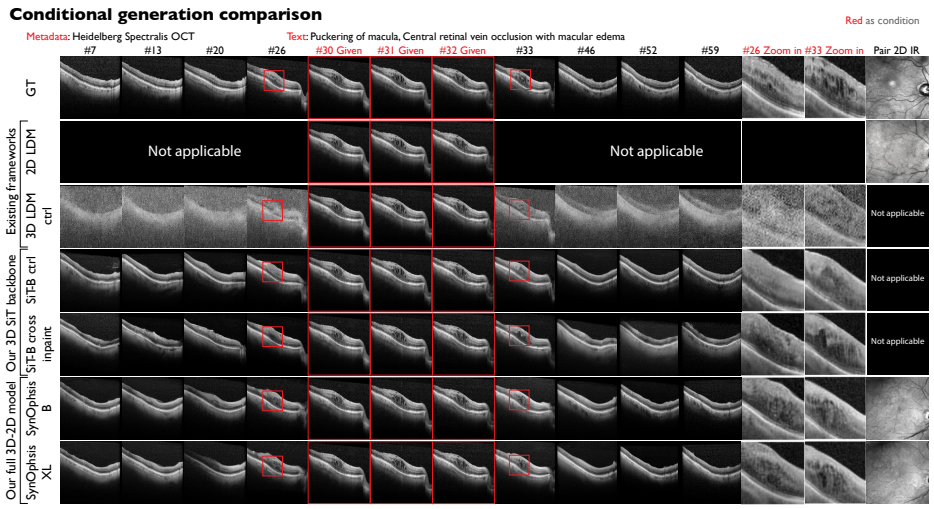

Figure 3: **Conditional generation from one observed slice.** A thin OCT volume with macular puckering and central retinal vein occlusion (Heidelberg Spectralis, 7.5×8.9 mm, 60 B-scans). Only B-scan #30 and its two neighbors are given, one latent slice, in red; the rest is generated. The next two columns magnify #26 and #33, and the last is the paired 2D infrared image from each method. Rows are banded by provenance; *Not applicable* marks what a method cannot emit.

The two volume cohorts have nearly equal floors, 29.5 on Internal and 28.5 on AI-READI. On the internal volumes SynOphsis-XL takes FVD from the LDM's 56.2 to 32.9, within 3.4 of that floor, and SynOphsis-B takes the ophthalmic distance from 7.65 to 1.07, level with the 1.078 of real data. On AI-READI none gets within 22 of that floor, and standard FVD does not separate the models. All five score between 51.0 and 59.6. R3D-18 is trained on Kinetics-400 video, not ophthalmic data, and two halves of the real volumes already score 28.5 apart at n=100, more than three times the spread across the five models (Appendix B.1). They separate under the other two. In the ophthalmic features every SiT-based model is two to four times below the LDM's 6.46. SynOphsis-B is the lowest at 1.76 and SynOphsis-XL is at 2.53, and Inception KID gives 98.6 for SynOphsis-XL against 246.1 (Table 12). OCTA, the third protocol, runs the same seven conditional tasks in Appendix C.6, and Appendix C.8 collects the failure cases. These gains are not specific to the representation we align to. Recomputed in four ophthalmic encoders from other groups, SynOphsis-XL leads the LDM baseline in 20 of the 21 benchmark-by-encoder combinations on FID and 15 on KID (Appendix C.1).

OCT takes 56.8% of the training images (Table 25); the other 2D modalities are its sparse end. Against the 2D LDM, SynOphsis-XL leads on all seven in Inception FID and six of seven in ours. Against SiT-XL, SynOphsis-XL scores within 9.6% on every modality in the ophthalmic features, the largest difference favoring the joint model. In Inception it scores within 3.4% on six of seven modalities, OCTA en-face being the exception at 13%. At B scale the two metrics disagree. Against SiT-B, SynOphsis-B is up to 29% better in the ophthalmic features but up to 34% worse in Inception. In the joint B-scale model all eleven modalities, three protocols and the paired-2D stream compete for the same capacity. At XL the gap narrows. One model therefore holds them all, leading the LDM baseline in 25 of 28 cells and staying within a 4.4% relative range of SiT-XL on average. Table 3 measures the conditional generation the joint tokens make possible.

5.2 Conditional Generation

The conditional tasks test the interface, each with the same weights under a different mask. Mid, Span, First and Last condition on latent slices by position, 2D on the paired 2D image alone, and 2D+Mid and 2D+10% on that image with the middle slice or a sparse 10% subset (Appendix B.2). We compare the LDM baselines, task-specific SiT-B baselines, and SynOphsis-B, with SynOphsis-XL as a capacity reference. Table 3 reports volumetric quality above and the paired 2D output below, by PSNR and DINOv3-Large perceptual distance. A slash marks a task a method cannot take. No published method generates the volume from observed slices and a paired 2D image together (Table 1), so the two combined tasks compare against our own cross-attention variant.

Table 2: Unconditional generation; lower is better throughout. Each model has a Standard row and a DINO row, the latter in our ophthalmic DINOv3-Large feature space: FID for 2D, FVD for 3D. LDM is the 2D latent-diffusion baseline in the 2D blocks and the 3D one in the 3D block. Each benchmark carries its floor, measured between two disjoint halves of the real data. Bold is best in column, underline second; n=1000 per non-OCT modality. Kermany-T and -H are its test and healthy splits; OCTA-enf is OCTA en-face. †These rows are our ablation, our extension of SiT to volumes with the joint tokens removed; published SiT and REPA generate neither.

		2D OCT: FID					*3D OCT: FVD*	
Model	Feature	OCTDL	OCTID	Kermany-T	Kermany-H	Internal OCT	AI-READI	Internal OCT 3D
LDM	Standard	119.6	125.7	103.2	167.4	66.8	**51.0**	56.2
	DINO	12.99	12.15	7.87	14.38	4.17	6.46	7.65
SiT-B†	Standard	104.3	116.9	78.8	126.5	57.4	59.6	36.0
	DINO	**11.04**	**9.55**	7.36	**9.59**	1.62	3.08	1.12
SynOphsis-B	Standard	94.1	**101.5**	68.2	**110.7**	40.1	59.2	35.1
	DINO	11.85	10.15	**7.24**	10.08	1.37	**1.76**	**1.07**
SiT-XL†	Standard	86.2	121.5	**67.3**	115.7	30.3	51.7	35.0
	DINO	12.16	10.76	7.39	10.75	**1.24**	2.32	1.10
SynOphsis-XL	Standard	**82.7**	116.6	73.1	112.9	**29.9**	52.2	**32.9**
	DINO	11.55	13.10	7.43	10.44	**1.24**	2.53	1.12

		2D non-OCT ophthalmic: FID						
Model	Feature	CFP	FAF	FFA	ICGA	IR/SLO	OCTA	OCTA-enf
LDM	Standard	39.2	58.9	60.8	55.9	55.5	37.2	71.7
	DINO	6.67	3.82	8.95	7.41	5.31	6.16	**7.24**
SiT-B†	Standard	43.9	68.4	61.3	66.3	68.0	58.8	87.6
	DINO	8.72	5.69	14.24	13.45	7.82	7.78	16.04
SynOphsis-B	Standard	35.2	75.2	67.7	81.8	91.2	59.9	105.9
	DINO	7.21	4.07	11.37	11.21	5.63	7.04	17.18
SiT-XL†	Standard	**22.5**	**29.1**	**30.5**	**35.9**	**30.1**	**23.6**	**48.8**
	DINO	**4.17**	2.09	**4.90**	3.82	2.39	**4.40**	7.62
SynOphsis-XL	Standard	22.9	30.1	30.9	36.6	30.3	24.2	55.2
	DINO	4.32	**2.00**	4.98	**3.69**	**2.16**	4.68	7.77

The seven conditions rank by how much of the volume's interior the evidence constrains, and every method reproduces that ordering. For SynOphsis-XL on thin volumes, three slices spread through the slab (Span) are worth 2.15 dB over the middle one alone (Mid), and a single terminal slice (First, Last) is worth 1.3 to 1.8 dB *less*, because the model must extrapolate across the slab rather than interpolate within it. Appendix C.4 sweeps the budget over 100 volumes at eleven settings. The 2D-only task supplies no B-scan and is the weakest condition for every method. SynOphsis-XL reaches 8.3 D-LPIPS on thin volumes against the LDM's 25.1, a factor of three, while PSNR spans only 12.6 to 14.2 dB, since nothing in the input anchors the output to the eye scanned (Appendix C.5).

The two internal variants isolate how conditioning enters, with backbone, data, and budget fixed (Section 4). Both improve on the 3D LDM baselines by 2.27 to 4.40 dB over the eight shared cells, and SiT-B Ctrl averages 3.28 dB over the 3D LDM under the same ControlNet conditioning. Most of that distance lies in the stack, which differs in transport, backbone, autoencoder, objective and scale (Appendix D.4). The gap between the two mechanisms is an order of magnitude smaller. Inpainting leads on four of those eight cells, trails on three, and ties on one, all within -0.18 to $+0.26$ dB. SynOphsis-XL gains 3.29 dB over SynOphsis-B on thick 2D+10% and nothing on thick 2D-only, so scale helps most where the model has the most observed context, and thick 2D+10% falls from 0.7 to 0.3 D-LPIPS as well. Over all fourteen cells the step from B to XL averages $+0.58$ dB.

Table 3 asks whether one generalist checkpoint reaches specialist quality on the tasks a specialist supports, while covering the tasks it cannot. On the overlapping tasks SynOphsis-B matches SiT-B Inpaint, which shares its backbone, data, and budget, leading in eleven of fourteen cells in D-LPIPS and nine in PSNR for an average of $+0.03$ dB. The added generality costs no reconstruction accuracy, and at equal parameter count a model carrying both has no reason to reconstruct the volume more accurately. SynOphsis-XL attains the best value in all fourteen cells under both metrics. The same checkpoint also produces what the specialists have no output for. Neither conditioning variant emits a paired 2D image at all, so that half of Table 3 exists only for the joint models (Figure 3, with three further cases in Appendix C.2). On thick volumes SynOphsis-XL under Mid ranks each generated en-face image against all 100 generated volumes' projections at a median of 18.0, against 19.5 for a real acquisition, so the two halves agree in aggregate about as closely as a co-acquired pair; on thin volumes it is uninformative for real pairs too (Appendix C.3). Against the 2D LDM, which produces a 2D image from 3D evidence without the volume beside it, both of our models lead on every task and both protocols, at XL scale by 0.3 to 1.9 dB on thin volumes and by 3.4 to

Table 3: Conditional generation: volumetric quality above, the paired 2D output below, over the same tasks. D-LPIPS is $10^3\times$DINO-LPIPS; lower is better, bold best in column and underline second at the printed precision. SynOphsis uses one checkpoint for every column and only the observation mask changes; the SiT rows are task-specific conditioning variants, and a slash marks a mode that architecture cannot emit rather than a setting left unrun. SSIM is in Table 27.

		Thin OCT							*Thick OCT*						
Model	Metric	Mid	Span	First	Last	2D	2D+Mid	2D+10%	Mid	Span	First	Last	2D	2D+Mid	2D+10%
Existing methods															
3D LDM Ctrl	PSNR	15.49	16.84	14.46	14.87	/	/	/	19.46	21.65	19.28	18.81	/	/	/
	D-LPIPS	8.8	6.1	13.1	10.8	/	/	/	7.1	4.7	8.3	9.3	/	/	/
3D LDM Cross	PSNR	/	/	/	/	12.64	/	/	/	/	/	/	16.74	/	/
	D-LPIPS	/	/	/	/	25.1	/	/	/	/	/	/	13.2	/	/
Ours: 3D SiT, alternative conditioning															
SiT-B Ctrl	PSNR	19.14	20.98	17.60	17.86	/	/	/	22.89	25.41	21.55	21.64	/	/	/
	D-LPIPS	2.2	1.3	3.8	3.0	/	/	/	2.1	1.5	4.1	4.6	/	/	/
SiT-B Inpaint	PSNR	19.18	21.24	17.81	17.86	13.87	19.20	23.88	22.71	25.24	21.68	21.62	19.67	22.71	24.59
	D-LPIPS	2.0	**1.2**	3.3	3.0	9.0	1.9	**0.5**	2.6	1.8	4.1	5.1	8.0	2.6	0.5
Ours: joint-token															
SynOphsis-B	PSNR	19.24	21.48	17.70	18.04	14.15	19.45	23.83	22.57	25.33	21.93	21.71	**19.94**	22.71	23.54
	D-LPIPS	1.9	**1.2**	3.2	2.7	8.4	1.7	**0.5**	2.5	1.6	3.7	4.2	6.8	2.3	0.7
SynOphsis-XL	PSNR	**19.59**	**21.74**	**17.83**	**18.28**	**14.17**	**19.78**	**24.29**	**23.25**	**26.36**	**22.24**	**21.96**	**19.94**	**23.46**	**26.83**
	D-LPIPS	**1.7**	**1.2**	**2.9**	**2.6**	**8.3**	**1.5**	**0.5**	**1.5**	**0.9**	**2.5**	**3.3**	**6.5**	**1.3**	**0.3**

Table 4: Paired 2D output under the same conditions. D-LPIPS is $10^3\times$DINO-LPIPS, lower better. Neither SiT variant emits a paired 2D image, so only the joint models and the 2D LDM appear.

		Thin OCT paired 2D				*Thick OCT paired 2D*			
Model	Metric	Mid	Span	First	Last	Mid	Span	First	Last
Existing methods									
2D LDM	PSNR	13.28	13.50	12.54	13.52	14.92	15.41	15.50	15.68
	D-LPIPS	19.9	18.8	19.4	20.3	21.0	22.0	22.5	23.2
Ours: joint-token									
SynOphsis-B	PSNR	14.31	14.64	14.03	13.72	18.63	18.42	17.81	18.18
	D-LPIPS	17.2	15.0	16.8	16.6	13.6	13.3	15.0	13.8
SynOphsis-XL	PSNR	**14.42**	**15.12**	**14.46**	**13.85**	**19.43**	**19.88**	**18.92**	**19.15**
	D-LPIPS	**14.8**	**12.7**	**15.0**	**14.6**	**12.3**	**11.9**	**12.8**	**13.1**

4.5 dB on thick ones. Alongside a B-scan (Mid to 2D+Mid) the paired view is worth 0.19 dB on thin volumes, and alone it is the difference between that task and no output.

6 Conclusion

We presented SynOphsis, a single stochastic-interpolant transformer covering an ophthalmic examination with one set of weights. We place a volume and the 2D view recorded with it in one self-attention sequence, sample the pair along a single trajectory, and condition by masking latent values in that sequence, so any mixture of sparse B-scans and en-face evidence, in either direction, is the same operation on the same weights. It adds under 0.01% of the parameters and keeps the volumetric reconstruction accuracy of matched task-specific variants. We align an intermediate representation to a frozen ophthalmic encoder, weighted to the generated tokens, and align the projected volume against its reading of the paired image.

We evaluated the model unconditionally on eleven 2D modalities and three volume protocols, and conditionally on seven tasks at two thicknesses. On the internal cohort the volumetric numbers are within 3.4 FVD of the floor; on AI-READI none gets within 22 and standard FVD separates none of them, so the ophthalmic and KID measures carry it. Neither conditioning specialist emits a paired 2D image, so that half of the evaluation exists only for the joint models. Separately from joint tokenization, our stack improves on the retrained latent-diffusion baseline by 3.28 dB PSNR.

One checkpoint therefore spans a task family that none of the architectures we compare against covers. Because a further combination of evidence is a further mask over the same weights, we intend the model as a base for ophthalmic generation beyond the tasks measured here.

REPRODUCIBILITY STATEMENT

The token constructions, the alignment objective and the masking scheme are given in Appendix A. The sampler, the metric definitions, the real-versus-real floors and the seven conditional task definitions are in Appendix B, and Appendix C.5 gives the volume-to-volume spread behind every conditional number together with a paired test of each margin against it. Latent-space preprocessing, the model and optimization settings and every baseline configuration are in Appendix D. The volumetric and 2D data censuses, the instrument composition and the participant-level split protocol are in Appendix F. All reported numbers use guidance scale 1, the default, for our models and the baselines alike, and Appendix C.7 reports the sweep behind that choice.

ETHICS STATEMENT

This work uses retrospective, de-identified human ophthalmic imaging. Institutional review board approval was obtained for the internal cohort, with a waiver of informed consent for the retrospective analysis of de-identified imaging, and AI-READI is used under its data use agreement. All procedures adhered to the tenets of the Declaration of Helsinki. Appendix F gives the provenance and the data-use terms for both sources. No image in this work is presented as belonging to an identifiable individual, and the clinical text and acquisition metadata used as conditioning are drawn from the de-identified record. The model produces images that resemble clinical acquisitions, so a generated volume or view should not be presented as a record of a real eye. The agreement reported in Appendix C.3 holds in aggregate and not at the level of an individual eye.

USE OF GENERATIVE AI

We used generative AI tools to draft and edit parts of this paper, including Appendix C.3, and to shorten and edit the main text for readability. We used them to design the paired-view ranking measurement that appendix reports, and to write the code that computed its ranks from the generated volumes and their en-face projections. We also used them to write the code that formats the evaluation outputs into the result tables in this paper. We used them to search for related work and to check each bibliography entry against its publisher record, and we read every entry we kept. We used AI coding assistants while implementing the model, for autocompletion and for individual modules, and we reviewed and tested what they produced. We did not use them to generate synthetic datasets, to formulate or prove mathematical claims, or to design the model, the training objective or the training procedure. We have reviewed all AI-assisted work, and we take responsibility for the final content of this paper, including every claim, number and artifact in it.

REFERENCES

AI-READI Consortium. AI-READI: rethinking AI data collection, preparation and sharing in diabetes research and beyond. *Nature Metabolism*, 6(12):2210–2212, 2024. doi: 10.1038/s42255-024-01165-x.

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Rashadul Hasan Badhon, Atalie Carina Thompson, Jennifer I. Lim, Theodore Leng, and Minhaj Nur Alam. Diffusion model based OCT to OCTA translation. *Frontiers in Medicine*, 12:1655453, 2025. doi: 10.3389/fmed.2025.1655453.

Gabriela Ben Melech Stan, Diana Wofk, Scottie Fox, Alex Redden, Will Saxton, Jean Yu, Estelle Aflalo, Shao-Yen Tseng, Fabio Nonato, Matthias Muller, and Vasudev Lal. LDM3D: Latent diffusion model for 3D, 2023. URL https://arxiv.org/abs/2305.10853.

Tienyu Chang, Zhen Chen, Renjie Liang, Jinyu Ding, Jie Xu, Sunu Mathew, Amir Reza Hajrasouliha, Andrew J. Saykin, Ruogu Fang, Yu Huang, Jiang Bian, and Qingyu Chen. OphMAE: Bridging volumetric and planar imaging with a foundation model for adaptive ophthalmological diagnosis, 2026. URL https://arxiv.org/abs/2605.02714.

Ruoyu Chen, Weiyi Zhang, Fan Song, Honghua Yu, Dan Cao, Yingfeng Zheng, Mingguang He, and Danli Shi. Translating color fundus photography to indocyanine green angiography using deep-learning for age-related macular degeneration screening. *npj Digital Medicine*, 7(1):34, 2024. doi: 10.1038/s41746-024-01018-7.

Ruoyu Chen, Kezheng Xu, Kangyan Zheng, Weiyi Zhang, Yan Lu, Mingguang He, and Danli Shi. Noninvasive synthesis of multiframe ultra-widefield fluorescein angiography from color fundus photographs. *Ophthalmology Retina*, 10(2):176–184, 2026a. doi: 10.1016/j.oret.2025.08.002.

Ruoyu Chen, Weiyi Zhang, Bowen Liu, Xinyuan Wu, Xiaolan Chen, Pusheng Xu, Shunming Liu, Mingguang He, and Danli Shi. Boosting foundation models for rare eye disease diagnosis via a multimodal text-to-image generative framework. *npj Digital Medicine*, 9(1):371, 2026b. doi: 10.1038/s41746-026-02560-2.

Hyungjin Chung and Jong Chul Ye. Score-based diffusion models for accelerated MRI. *Medical Image Analysis*, 80:102479, 2022. doi: 10.1016/j.media.2022.102479.

Coen de Vente, Mohammad Mohaiminul Islam, Philippe Valmaggia, Carel Hoyng, Adnan Tufail, and Clara I. Sánchez. Conditioning 3D diffusion models with 2D images: Towards standardized OCT volumes through en face-informed super-resolution. In *NeurIPS Workshop on GenAI for Health*, 2024.

Gozde Merve Demirci, Jiaqi Yang, Hyun Sung Song, Chao Chen, Wei-Chi Wu, and Chia-Ling Tsai. Topology-aware conditional latent diffusion for multi-view fundus image synthesis. In *ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pp. 453–457, 2025. doi: 10.1145/3721201.3725432.

Yash Deo, Yan Jia, Toni Lassila, William A. P. Smith, Tom Lawton, Siyuan Kang, Alejandro F. Frangi, and Ibrahim Habli. Metrics that matter: Evaluating image quality metrics for medical image generation, 2025. URL https://arxiv.org/abs/2505.07175.

Ziyu Gao, Xinyuan Wu, Xiaolan Chen, Zhuoran Liu, Ruoyu Chen, Bowen Liu, Bingjie Yan, Zhenhan Wang, Kai Jin, Jiancheng Yang, Yih Chung Tham, Mingguang He, and Danli Shi. EyeWorld: A generative world model of ocular state and dynamics, 2026. URL https://arxiv.org/abs/2603.14039.

Fernando García-Torres, Rocío del Amor, Sandra Morales, Álvaro Barroso, Peter Heiduschka, Björn Kemper, and Valery Naranjo. DualDiT: A conditional dual-output diffusion transformer for joint OCT image and segmentation mask generation, 2026. URL https://arxiv.org/abs/2607.29337.

Peyman Gholami, Priyanka Roy, Mohana Kuppuswamy Parthasarathy, and Vasudevan Lakshminarayanan. OCTID: Optical coherence tomography image database. *Computers & Electrical Engineering*, 81:106532, 2020. doi: 10.1016/j.compeleceng.2019.106532.

Sojung Go, Younghoon Ji, Sang Jun Park, and Soochahn Lee. Generation of structurally realistic retinal fundus images with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2335–2344, 2024.

Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, and Daguang Xu. MAISI: Medical AI for synthetic imaging. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4430–4441, 2025. doi: 10.1109/WACV61041.2025.00435.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. SparseCtrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision (ECCV)*, pp. 330–348, 2024.

Kun Huang, Xiao Ma, Yuhan Zhang, Na Su, Songtao Yuan, Yong Liu, Qiang Chen, and Huazhu Fu. Memory-efficient high-resolution OCT volume synthesis with cascaded amortized latent diffusion models. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 478–487, 2024. doi: 10.1007/978-3-031-72104-5_46.

Indu Ilanchezian, Valentyn Boreiko, Laura Kühlewein, Ziwei Huang, Murat Seçkin Ayhan, Matthias Hein, Lisa Koch, and Philipp Berens. Development and validation of an AI algorithm to generate realistic and meaningful counterfactuals for retinal imaging based on diffusion models. *PLOS Digital Health*, 4(5):e0000853, 2025. doi: 10.1371/journal.pdig.0000853.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking FID: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9307–9315, 2024. doi: 10.1109/CVPR52733.2024.00889.

Haorui Ji, Weizhe Liu, Hongdong Li, and Hengkai Guo. FLUX3D: High-fidelity 3D Gaussian generation with diffusion-aligned sparse representation, 2026. URL https://arxiv.org/abs/2606.24874.

Aghiles Kebaili, Jérôme Lapuyade-Lahorgue, Pierre Vera, and Su Ruan. Multi-modal MRI synthesis with conditional latent diffusion models for data augmentation in tumor segmentation. *Computerized Medical Imaging and Graphics*, 123:102532, 2025.

Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C. S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K. Prasadha, Jacqueline Pei, Magdalene Y. L. Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Runze Zhang, Lianghong Zheng, Rui Hou, William Shi, Xin Fu, Yaou Duan, Viet Anh N. Huu, Cindy Wen, Edward D. Zhang, Charlotte L. Zhang, Oulan Li, Xiaobo Wang, Michael A. Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M. Anthony Lewis, Huimin Xia, and Kang Zhang. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131.e9, 2018. doi: 10.1016/j.cell.2018.02.010.

Firas Khader, Gustav Müller-Franzes, Soroosh Tayebi Arasteh, Tianyu Han, Christoph Haarburger, Maximilian Schulze-Hagen, Philipp Schad, Sandy Engelhardt, Bettina Baeßler, Sebastian Foersch, Johannes Stegmaier, Christiane Kuhl, Sven Nebelung, Jakob Nikolas Kather, and Daniel Truhn. Denoising diffusion probabilistic models for 3D medical image generation. *Scientific Reports*, 13:7303, 2023. doi: 10.1038/s41598-023-34341-2.

Mikhail Kulyabin, Aleksei Zhdanov, Anastasia Nikiforova, Andrey Stepichev, Anna Kuznetsova, Mikhail Ronkin, Vasilii Borisov, Alexander Bogachev, Sergey Korotkich, Paul A. Constable, and Andreas Maier. OCTDL: Optical coherence tomography dataset for image-based deep learning methods. *Scientific Data*, 11:365, 2024. doi: 10.1038/s41597-024-03182-7.

Mingjie Li, Wenjia Cai, Rui Liu, Yuetian Weng, Xiaoyun Zhao, Cong Wang, Xin Chen, Zhong Liu, Caineng Pan, Mengke Li, Yingfeng Zheng, Yizhi Liu, Flora D. Salim, Karin Verspoor, Xiaodan Liang, and Xiaojun Chang. FFA-IR: Towards an explainable and reliable medical report generation benchmark. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Bowen Liu, Yue Wu, Ruoyu Chen, Pusheng Xu, Peng Xiao, Zhen Tian, Binwei Huang, Mingguang He, and Danli Shi. Generating 3D optical coherence tomography from 2D fundus images via diffusion models. *Research Square*, 2025a. doi: 10.21203/rs.3.rs-6160112/v1. preprint.

Bowen Liu, Weiyi Zhang, Peranut Chotcomwongse, Xiaolan Chen, Ruoyu Chen, Pawin Pakaymaskul, Niracha Arjkongharn, Nattaporn Vongsa, Xuelian Cheng, Zongyuan Ge, Kun Huang, Xiaohui Li, Yiru Duan, Zhenbang Wang, BaoYe Xie, Qiang Chen, Huazhu Fu, Michael A. Mahr, Jiaqi Qu, Wangyiyang Chen, Shiye Wang, Yubo Tan, Yongjie Li, Mingguang He, Danli Shi, and Paisan Ruamviboonsuk. APTOS-2024 challenge report: Generation of synthetic 3D OCT images from fundus photographs, 2025b. URL https://arxiv.org/abs/2506.07542.

Zixuan Liu, Hanwen Xu, Addie Woicik, Linda G. Shapiro, Marian Blazes, Yue Wu, Verena Steffen, Catherine A. Cukras, Cecilia S. Lee, Miao Zhang, Aaron Y. Lee, and Sheng Wang. A three-dimensional multi-modal foundation model for optical coherence tomography. *Nature Biomedical Engineering*, 2026. doi: 10.1038/s41551-026-01662-2. arXiv:2408.11227.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision (ECCV)*, pp. 23–40, 2024.

Tengfei Ma, Ruiqi Wu, Chenran Zhang, Ye Geng, Na Su, Xiangyuan Duanmu, Tao Zhou, Yi Zhou, and Wen Fan. Propagating structural guidance: Synthesizing fluorescein angiography from fundus images and sparse OCT scans, 2026. URL https://arxiv.org/abs/2606.16234.

José Morano, Botond Fazekas, Emese Sükei, Ronald Fecso, Taha Emre, Markus Gumpinger, Georg Faustmann, Marzieh Oghbaie, Ursula Schmidt-Erfurth, and Hrvoje Bogunović. Multimodal foundation model and benchmark for comprehensive retinal OCT image analysis. *npj Digital Medicine*, 8(1):576, 2025. doi: 10.1038/s41746-025-01852-3.

Junzhi Ning, Cheng Tang, Kaijing Zhou, Diping Song, Lihao Liu, Ming Hu, Wei Li, Huihui Xu, Yanzhou Su, Tianbin Li, Jiyao Liu, Jin Ye, Sheng Zhang, Yuanfeng Ji, and Junjun He. RetinaLogos: Fine-grained synthesis of high-resolution retinal images through captions. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Cynthia Owsley, Dawn S. Matthies, Gerald McGwin, Jeffrey C. Edberg, Sally L. Baxter, Linda M. Zangwill, Julia P. Owen, and Cecilia S. Lee. Cross-sectional design and protocol for artificial intelligence ready and equitable atlas for diabetes insights (AI-READI). *BMJ Open*, 15(2): e097449, 2025. doi: 10.1136/bmjopen-2024-097449. URL https://bmjopen.bmj.com/content/15/2/e097449.

Somayeh Pakdelmoez, Saba Omidikia, Seyyed Ali Seyyedsalehi, and Seyyede Zohreh Seyyedsalehi. Controllable retinal image synthesis using conditional StyleGAN and latent space manipulation for improved diagnosis and grading of diabetic retinopathy, 2024. URL https://arxiv.org/abs/2409.07422.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Jianing Qiu, Jian Wu, Hao Wei, Peilun Shi, Minqing Zhang, Yunyun Sun, Lin Li, Hanruo Liu, Hongyi Liu, Simeng Hou, Yuyang Zhao, Xuehui Shi, Junfang Xian, Xiaoxia Qu, Sirui Zhu, Lijie Pan, Xiaoniao Chen, Xiaojia Zhang, Shuai Jiang, Kebing Wang, Chenlong Yang, Mingqiang Chen, Sujie Fan, Jianhua Hu, Aiguo Lv, Hui Miao, Li Guo, Shujun Zhang, Cheng Pei, Xiaojuan Fan, Jianqin Lei, Ting Wei, Junguo Duan, Chun Liu, Xiaobo Xia, Siqi Xiong, Junhong Li, Kyle Lam, Benny Lo, Yih Chung Tham, Tien Yin Wong, Ningli Wang, and Wu Yuan. Development and validation of a multimodal multitask vision foundation model for generalist ophthalmic artificial intelligence. *NEJM AI*, 1(12):AIoa2300221, 2024. doi: 10.1056/AIoa2300221.

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

Danli Shi, Weiyi Zhang, Shuang He, Yanxian Chen, Fan Song, Shunming Liu, Ruobing Wang, Yingfeng Zheng, and Mingguang He. Translation of color fundus photography into fluorescein angiography using deep learning for enhanced diabetic retinopathy screening. *Ophthalmology Science*, 3(4):100401, 2023. doi: 10.1016/j.xops.2023.100401.

Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. MVDream: Multi-view diffusion for 3D generation. In *International Conference on Learning Representations*, 2024.

Julio Silva-Rodríguez, Hadi Chakor, Riadh Kobbi, Jose Dolz, and Ismail Ben Ayed. A foundation language-image model of the retina (FLAIR): Encoding expert knowledge in text supervision. *Medical Image Analysis*, 99:103357, 2025. arXiv:2308.07898.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Fan Song, Weiyi Zhang, Yingfeng Zheng, Danli Shi, and Mingguang He. A deep learning model for generating fundus autofluorescence images from color fundus photography. *Advances in Ophthalmology Practice and Research*, 3(4):192–198, 2023. doi: 10.1016/j.aopr.2023.11.001.

Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. In *International Conference on Learning Representations*, 2022.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Alireza Tavakkoli, Sharif Amit Kamran, Khondker Fariha Hossain, and Stewart Lee Zuckerbrod. A novel deep learning conditional generative adversarial network for producing angiography images from retinal fundus photographs. *Scientific Reports*, 10(1):21580, 2020.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2018.

Zuzanna A. Wakefield-Skórniewska and Bartłomiej W. Papież. Evaluation of clinically steerable retinal image generation from foundation model latent spaces, 2026. URL https://arxiv.org/abs/2608.13455.

Jinzhuo Wang, Kai Wang, Yunfang Yu, Yuxing Lu, Wenchao Xiao, Zhuo Sun, Fei Liu, Zixing Zou, Yuanxu Gao, Lei Yang, Hong-Yu Zhou, Hanpei Miao, Wenting Zhao, Lisha Huang, Lingchao Zeng, Rui Guo, Ieng Chong, Boyu Deng, Linling Cheng, Xiaoniao Chen, Jing Luo, Meng-Hua Zhu, Daniel Baptista-Hon, Olivia Monteiro, Ming Li, Yu Ke, Jiahui Li, Simiao Zeng, Taihua Guan, Jin Zeng, Kanmin Xue, Eric Oermann, Huiyan Luo, Yun Yin, Kang Zhang, and Jia Qu. Self-improving generative foundation model for synthetic medical image generation and clinical applications. *Nature Medicine*, 31(2):609–617, 2025. doi: 10.1038/s41591-024-03359-y.

Ruiqi Wu, Chenran Zhang, Jianle Zhang, Yi Zhou, Tao Zhou, and Huazhu Fu. MM-Retinal: Knowledge-enhanced foundational pretraining with fundus image-text expertise. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 722–732, 2024a. doi: 10.1007/978-3-031-72378-0_67.

Xinyuan Wu, Lili Wang, Ruoyu Chen, Bowen Liu, Weiyi Zhang, Xi Yang, Yifan Feng, Mingguang He, and Danli Shi. FFA Sora, video generation as fundus fluorescein angiography simulator, 2024b. URL https://arxiv.org/abs/2412.17346.

Yuli Wu, Weidong He, Dennis Eschweiler, Ningxin Dou, Zixin Fan, Shengli Mi, Peter Walter, and Johannes Stegmaier. Retinal OCT synthesis with denoising diffusion probabilistic models for layer segmentation. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, 2024c. doi: 10.1109/ISBI56570.2024.10635836.

Yuli Wu, Fucheng Liu, Rüveyda Yilmaz, Henning Konermann, Peter Walter, and Johannes Stegmaier. A pragmatic note on evaluating generative models with Fréchet Inception distance for retinal image synthesis. In *Medical Imaging with Deep Learning (MIDL)*, volume 315 of *Proceedings of Machine Learning Research*, pp. 2352–2368, 2026.

Xue Xia, Kun Zhan, Ying Li, Guobei Xiao, Jinhua Yan, Zhuxiang Huang, Guofu Huang, and Yuming Fang. Eye disease diagnosis and fundus synthesis: A large-scale dataset and benchmark. In *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, 2022. doi: 10.1109/MMSP55362.2022.9949547.

Pusheng Xu, Peranut Chotcomwongse, Weiyi Zhang, Xiaolan Chen, Xinyuan Wu, Florence H. T. Chung, Xueli Zhang, Mingguang He, Danli Shi, and Paisan Ruamviboonsuk. AngioReport: dataset and baseline methods for fundus angiography report generation. *British Journal of Ophthalmology*, 109(11):1283–1288, 2025. doi: 10.1136/bjo-2024-327006.

Yanwu Xu, Li Sun, Wei Peng, Shuyue Jia, Katelyn Morrison, Adam Perer, Afrooz Zandifar, Shyam Visweswaran, Motahhare Eslami, and Kayhan Batmanghelich. MedSyn: Text-guided anatomy-aware synthesis of high-fidelity 3D CT images. *IEEE Transactions on Medical Imaging*, 43(10): 3648–3660, 2024. doi: 10.1109/TMI.2024.3415032.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations (ICLR)*, 2025.

Chenlu Zhan, Yu Lin, Gaoang Wang, Hongwei Wang, and Jian Wu. MedM2G: Unifying medical multi-modal generation via cross-guided diffusion with visual invariant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11502–11512, 2024. doi: 10.1109/CVPR52733.2024.01093.

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 (ICCV)*, pp. 3836–3847, 2023a.

Weiyi Zhang, Peranut Chotcomwongse, Xiaolan Chen, Florence H. T. Chung, Fan Song, Xueli Zhang, Mingguang He, Danli Shi, and Paisan Ruamviboonsuk. Angiographic report generation for the 3rd APTOS's competition: Dataset and baseline methods. *medRxiv*, 2023b. doi: 10.1101/2023.11.26.23299021.

Weiyi Zhang, Siyu Huang, Jiancheng Yang, Ruoyu Chen, Zongyuan Ge, Yingfeng Zheng, Danli Shi, and Mingguang He. Fundus2Video: Cross-modal angiography video generation from static fundus photography with clinical knowledge guidance. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 689–699, 2024. doi: 10.1007/978-3-031-72378-0_64.

Weiyi Zhang, Jiancheng Yang, Ruoyu Chen, Siyu Huang, Pusheng Xu, Xiaolan Chen, Zhenhan Wang, Shanfu Lu, Hongyu Cao, Xianwen Shang, Danli Shi, and Mingguang He. Fundus to fluorescein angiography video generation as a retinal generative foundation model. *npj Digital Medicine*, 9(1):696, 2026. doi: 10.1038/s41746-026-02804-1.

Can Zhao, Pengfei Guo, Dong Yang, Yucheng Tang, Yufan He, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, and Daguang Xu. MAISI-v2: Accelerated 3D high-resolution medical image synthesis with rectified flow and region-specific contrastive loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. arXiv:2508.05772.

Yukun Zhou, Mark A. Chia, Siegfried K. Wagner, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023.

Lingting Zhu, Zeyue Xue, Zhenchao Jin, Xian Liu, Jingzhen He, Ziwei Liu, and Lequan Yu. Make-A-Volume: Leveraging latent diffusion models for cross-modality 3D brain MRI synthesis. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2023.

APPENDIX

A METHOD DETAILS

Section 3 states two constructions in outline. This section gives them in full, first how each input becomes tokens and then how the alignment terms are defined.

Token construction. We write C for the latent channel count, D for the depth axis, H, W for the in-plane size, and a subscript z for the VAE latent dimensions, so a volume is $C \times D \times H \times W$ with latent $C \times D_z \times H_z \times W_z$ and the 2D case omits the depth axis. Let d be the shared transformer width, ρ index spatial patches and k latent depth patches.

Before flattening into a sequence, a 2D input has shape $1 \times 16 \times 16 \times d$ and either thin or thick 3D input has shape $20 \times 16 \times 16 \times d$, and the token counts in the main text follow from these shapes.

For a 2D latent $z^{2D} \in \mathbb{R}^{C \times H_z \times W_z}$, with π^{sp}_ρ the spatial positional embedding and e^{type}_{2D} the type embedding,

$$X^{2D}_\rho = E_{2D}(z^{2D})_\rho + \pi^{sp}_\rho + e^{type}_{2D}, \tag{6}$$

giving $X^{2D} \in \mathbb{R}^{N_{2D} \times d}$. Modality identity and other structured attributes arrive separately, through the metadata embedding y.

For a volumetric latent, with π^{temp}_k the temporal positional embedding and τ the branch label,

$$X^{3D,\tau}_{k,\rho} = E^\tau_{3D}(z^{3D})_{k,\rho} + \pi^{temp}_k + \pi^{sp}_\rho + e^{type}_{3D}, \qquad \tau \in \{\text{thin}, \text{thick}\}. \tag{7}$$

The thin and thick branches use separate 3D patch embeddings because their latent temporal patch sizes differ, while both produce the common $20 \times 16 \times 16$ grid.

A paired 2D image p uses the same E_{2D}, distinguished by two further embeddings: $e^{pair}_{m_p}$ for the paired-view modality m_p and $e^{obs}_{o_p}$ for whether that view is observed or generated,

$$X^p_\rho = E_{2D}(z^p)_\rho + \pi^{sp}_\rho + e^{type}_{2D} + e^{pair}_{m_p} + e^{obs}_{o_p}. \tag{8}$$

A 2D-only sequence is scored by the 2D head and predicts the 2D target, a 3D-only sequence by the 3D head, and a paired sequence by both, its paired portion passing through the same 2D head as an ordinary 2D sample.

The alignment objective. Let ℓ_{rep} be the selected SiT layer (`encoder_depth`), i index aligned tokens, $b \in \{2D, 3D\}$ the branch, g_b its trainable projection head and ϕ_b its frozen teacher. With norm the ℓ_2 normalization and sg a stop-gradient,

$$s_i = \text{norm}\big(g_b(F_{\ell_{\text{rep}}}(\bar{z}_t)_i)\big), \qquad q_i = \text{sg}(\text{norm}(\phi_b(x_0)_i)), \tag{9}$$

and the REPA objective is the patch-wise negative cosine similarity $\mathcal{L}_{\text{REPA}} = -\mathbb{E}\big[\frac{1}{N}\sum_i s_i^\top q_i\big]$. Only the SiT representation and the projection head are optimized.

Masked normalization. For sparse inpainting we replace the uniform patch average with a weighted one. Let $w_i \in [0, 1]$ be the generated-token weight and $\delta > 0$ a stabilizer:

$$\mathcal{A}_w(s, q) = -\frac{\sum_i w_i s_i^\top q_i}{\sum_i w_i + \delta}. \tag{10}$$

The weight is one for a valid generated token, zero for an observed or padded one, and may be fractional after temporal or spatial mask resizing; for paired 2D inpainting the sample-level keep/drop indicator is broadcast over all paired-2D spatial tokens, and an all-zero weight returns zero for that sample. The uniform average would let clean observed slices become alignment targets; the weighted one does not.

One shared g_{2D} serves ordinary and paired 2D tokens and one shared g_{3D} serves thin and thick volumes. Writing $F^{2D}_{\ell_{\text{rep}}}$, $F^{3D}_{\ell_{\text{rep}}}$, $F^{p}_{\ell_{\text{rep}}}$ for the token subsequences, w^{2D}, w^{3D}, w^p for their weights,

and $\bar{\phi}_{3D}^{chunk}$ for the chunked DINOv3 features after thick-volume pooling, the same-branch losses are

$$\mathcal{L}_{rep}^{2D} = \mathcal{A}_{w^{2D}}\Big(\mathrm{norm}(g_{2D}(F_{\ell_{\mathrm{rep}}}^{2D})), \mathrm{sg}(\mathrm{norm}(\phi_{2D}(x_0^{2D})))\Big), \tag{11}$$

$$\mathcal{L}_{rep}^{3D} = \mathcal{A}_{w^{3D}}\Big(\mathrm{norm}(g_{3D}(F_{\ell_{\mathrm{rep}}}^{3D})), \mathrm{sg}(\mathrm{norm}(\bar{\phi}_{3D}^{chunk}(x_0^{3D})))\Big), \tag{12}$$

$$\mathcal{L}_{rep}^{p} = \mathcal{A}_{w^{p}}\Big(\mathrm{norm}(g_{2D}(F_{\ell_{\mathrm{rep}}}^{p})), \mathrm{sg}(\mathrm{norm}(\phi_{2D}(p)))\Big). \tag{13}$$

For thin volumes the 20 DINO depth groups match the 20 transformer temporal groups directly. For thick volumes the 80 DINO groups are average-pooled in fours to give 20. The corresponding weights expand the latent-frame validity/inpainting mask over spatial tokens.

The cross-dimensional term. Let B be the minibatch size, $K = 20$ the common temporal token count and d_e the frozen encoder width. The projected 3D tokens are reshaped,

$$S^{3D} = \mathrm{reshape}\Big(g_{3D}(F_{\ell_{\mathrm{rep}}}^{3D})\Big) \in \mathbb{R}^{B\times K\times G_H\times G_W\times d_e}, \tag{14}$$

averaged over the height-token axis and bilinearly resized to the paired 2D grid, $\bar{S}^{3D\to 2D} = \mathrm{Resize}(\mathrm{Mean}_{G_H}(S^{3D}))$, then flattened into paired-2D token order. With w^{cross} the product of the resized 3D validity mask, the paired-2D spatial mask and the paired-sample availability indicator,

$$\mathcal{L}_{rep}^{3D\to 2D} = \mathcal{A}_{w^{cross}}\big(\mathrm{norm}(\bar{S}^{3D\to 2D}), \mathrm{sg}(\mathrm{norm}(\phi_{2D}(p)))\big). \tag{15}$$

This term is not part of the original REPA objective. It is ours, and it enforces consistency between the volumetric and the paired 2D representations.

B Evaluation protocol

B.1 Sampling and evaluation

Every reported number and every figure comes from the SDE sampler at 100 integration steps, for unconditional and conditional generation alike, at a guidance scale of 1 over the full noise interval.

Guidance would do little on the conditional tasks. The conditioning there is structural, since the observed B-scan latents are given to the model directly, so there is no null conditioning to extrapolate away from. The guidance scale would act only on the class embedding, which those tasks hold fixed. Unconditional generation is a different case. The models are trained with label dropout at probability 0.1, which makes guidance available at sampling time, and we sample without it.

Sample counts. Every benchmark metric compares a generated set against a reference set of the same size, drawn from the held-out test split and never from the training data. The generator runs once per reference item, so the two sets match in count and in the conditioning metadata the reference carries, and no benchmark is subsampled after generation. Table 5 gives the counts, which one evaluation script computes from one set of features for every model in Table 2, ours and the baselines alike.

Two OCT benchmarks are smaller than their split by construction. Internal OCT 2D is a fixed 1,000-image draw from a much larger test split and Kermany-H, a split of the Kermany collection (Kermany et al., 2018), a fixed 100-image draw restricted to normal eyes, both drawn once and reused across models. The volume benchmarks are 100 volumes each, the smallest number in the paper, and FVD differences below about one unit are within its noise (Appendix B.1).

The AI-READI splits are partitioned by participant rather than by image. No participant contributes to both sides, so neither the fellow eye nor a second visit crosses the boundary. We checked this rather than assume it, because a random image-level split in ophthalmology leaks across both.

OCTDL and OCTID are both external collections, and they are assembled differently. OCTID is organized by disease category from fifty patients at one site (Gholami et al., 2020), where the internal benchmark is an unselected split of a clinical series, which is the likeliest reason the two behave differently in Table 2.

Table 5: Generated and reference items per benchmark. The two sets are always equal in size; one generated item per reference item.

Benchmark	Kind	Items
OCTDL	2D OCT	1,618
Kermany-T	2D OCT	1,000
Internal OCT 2D	2D OCT	1,000
OCTID	2D OCT	572
Kermany-H	2D OCT	100
CFP, FAF, FFA, ICGA, IR, MRI, OCTA, OCTA en-face, External	2D	1,000 each
AS-OCT	2D	746
Internal OCT 3D (thin), AI-READI 3D (thick), OCTA 3D	volume	100 each

Table 6: Real against real. Each reference set is split in half and one half scored against the other, in the same two feature spaces as Table 2. "Per side" is how many items each half holds; where it equals the benchmark's own n the floor is directly comparable to the model rows, and where it is half (OCTDL, OCTID) the floor is an over-estimate, because these distances fall as n grows.

Benchmark	Per side	Benchmark n	Standard	DINO
Internal OCT 3D	100	100	29.48	1.078
AI-READI 3D	100	100	28.51	0.905
OCTA 3D	100	100	19.22	0.871
Internal OCT 2D	999	1,000	21.28	0.721
Kermany-T	1,000	1,000	39.87	2.131
Kermany-H	100	100	83.56	5.017
OCTDL	809	1,618	22.41	0.403
OCTID	286	572	36.60	0.891

Metric definitions. FID uses InceptionV3 (Szegedy et al., 2016) pool3 features and FVD the penultimate feature of a Kinetics-400 R3D-18 (Tran et al., 2018), both taken from torchvision. The DINO-prefixed variants replace those with our ophthalmic DINOv3-Large encoder's `forward_features`, mean-pooled over patch tokens, under the same preprocessing used in training; PRDC uses the same two feature sets. D-LPIPS is DINO-LPIPS scaled by 10^3 throughout so that the column fits one decimal.

Metric floors. The distances in Table 2 have a floor that is not zero, since two disjoint halves of the *real* data are already some distance apart at these sample sizes, and no generator can do better than that. We measured it instead of estimating it, by splitting each reference set in half and scoring one half against the other. Table 6 gives the result, and two rows change meaning with it.

On Internal OCT volumes the DINO metric sits at its floor. Real against real is 1.078 and SynOphsis reports 1.07 to 1.12. That spread is smaller than this evaluation resolves, so the margin over the baseline there is not headroom. A separate unconditional run of SynOphsis-XL reads 1.48 under the same evaluation (Table 21), 0.36 above the top of that range where the range itself is 0.05 wide, so run-to-run variation here exceeds the differences between models. In the unaligned space the same benchmark still has room, with a floor of 29.5 against our 32.9 to 35.6 and the 3D LDM at 56.2; on AI-READI both spaces have room and our distance to the floor is larger than on Internal.

Inception-v3 and R3D-18 are trained on natural images and on Kinetics-400 video, and neither is trained on ophthalmic data. The floors measure the consequence, and other groups report it in other settings. Jayasumana et al. (2024) find that Inception FID misranks generators and that its features violate the normality its estimator assumes, and Deo et al. (2025) find the standard image-quality scores move little when a medical image is altered in a clinically meaningful way. On color fundus and OCT, Wu et al. (2026) find FID and its variants misaligned with the downstream classification and segmentation performance the synthetic data is meant to serve. At n=100 the floor is large in absolute terms, Kermany-H scoring 83.6 real against real in Inception space and the AI-READI volumes 28.5, so on those rows differences between models are smaller than the metric's own noise. We report the ophthalmic space beside every standard column for that reason, and Appendix C.1 repeats the unconditional evaluation in four encoders trained by other groups.

Table 7: **Reconstruction through the frozen autoencoder.** Held-out images and volumes encoded and decoded by the Stable Diffusion VAE, scored against the input. PSNR is higher-better, LPIPS lower-better; each is a mean with its standard deviation.

	PSNR ↑	LPIPS ↓
2D images		
CFP	37.02 ± 3.23	0.083 ± 0.024
OCT	27.70 ± 2.68	0.141 ± 0.029
FAF	28.69 ± 2.13	0.156 ± 0.024
FFA	30.63 ± 2.77	0.126 ± 0.026
ICGA	31.23 ± 2.50	0.148 ± 0.029
IR	32.65 ± 2.84	0.140 ± 0.028
OCTA	23.46 ± 1.96	0.136 ± 0.069
OCTA en-face	21.28 ± 3.25	0.249 ± 0.040
AS-OCT	35.52 ± 1.75	0.084 ± 0.011
External	34.83 ± 1.56	0.120 ± 0.011
MRI	31.29 ± 6.65	0.075 ± 0.034
Volumes		
Thin OCT	28.08 ± 1.77	0.190 ± 0.030
Thick OCT	31.85 ± 2.68	0.198 ± 0.045
OCTA	25.72 ± 2.04	0.136 ± 0.085

The autoencoder's ceiling. Every model here decodes through the same frozen VAE, so a round-trip through it bounds every reconstruction number the paper reports. Table 7 measures that bound. The volumes sit two to five decibels below it: SynOphsis-XL's best task reaches 24.29 dB on thin against a ceiling of 28.08, 26.83 against 31.85 on thick, and 23.81 against 25.72 on OCTA.

The two columns disagree on OCTA. By PSNR the OCTA volume is the worst of the three, 25.72 dB against 31.85 for thick structural; by LPIPS it is the best, 0.136 against 0.198. Its reconstruction error is twice as large in magnitude and smaller perceptually, which is what losing noise-like content looks like rather than losing structure. The one row poor on both metrics is the 2D OCTA en-face view.

B.2 Conditional task definitions

The seven tasks in Table 3 are named by what they observe. Mid is the middle latent slice, Span the first, middle and last, First the first alone and Last the last alone. 2D is the paired 2D image with no slice at all, 2D+Mid is that image with the middle slice, and 2D+10% is that image with a sparse 10% subset of slices.

The observation-budget figures and Table 17 vary two further things, how many B-scans are observed and where they sit. Observed B-scans come in threes, since three neighbors form one latent slice, so a budget is always a multiple of three. A block is contiguous at the center of the volume and a spread places the same number of groups evenly across it. The figures give the budget as a count out of sixty, and the table reports both placements at each count.

C Additional results

The unconditional metrics come first, then the conditional results ordered by how much the model is given, then the qualitative material. Quality tracks the amount of evidence, and wherever every method degrades together the conditioning underdetermines the output, not any architecture.

C.1 Unconditional metrics and samples

Figure 4 is the comparison against real data and the baselines; the two figures after it drop the baselines and widen the coverage instead.

Figures 5 and 6 show unconditional samples without any baseline comparison, to give a fuller picture of what the model produces than the compact comparison in Figure 4. Both use SynOphsis-XL. Samples are chosen by a contrast, structure and coverage ranking and not by index, since the first sample in each directory is often not representative.

Figure 5 covers *twelve* 2D modalities, more than the paper scores. Beyond the three held out of Table 2, which are anterior segment, an eye photographed from outside and an orbit, it reaches slit

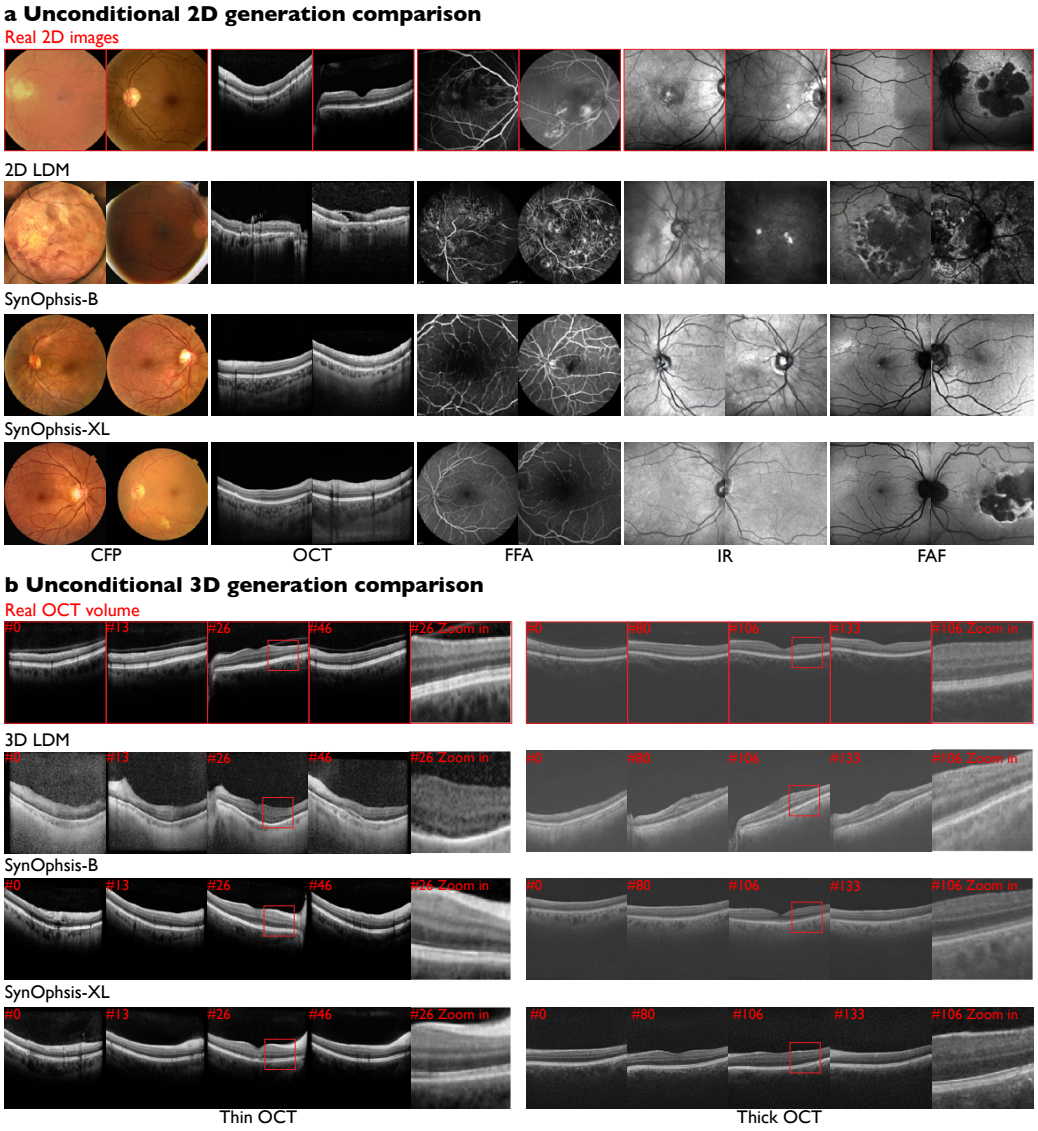

Figure 4: **Unconditional ophthalmic generation, against real data and the baselines.** Every row is unconditional. **(a)** 2D generation across five modalities (CFP, OCT, FFA, IR and FAF), with two samples each. Rows are real images, the 2D LDM baseline, SynOphsis-B and SynOphsis-XL. **(b)** 3D OCT, one thin and one thick volume, shown as four B-scans spanning the volume with a fifth column magnifying the marked region. Rows are a real volume, the 3D LDM baseline, and our two model sizes. A comparison of two generators cannot show whether either is close to an eye, which is what the real rows are for. Coverage across acquisition devices and fields of view (three vendors, 3×3 to 12×12 mm) is shown separately in Figure 6, since the 3D LDM baseline has no per-device samples to compare against.

lamp and ultrasound, which no table scores at all. Seeing them beside the retinal columns makes the point, one model with one set of weights and no per-modality head.

Text-conditioned 2D generation appears in the capability panel of Figure 2 and is not repeated here. That panel gives three prompts, one per modality, each producing an image of the modality the text names. The limitations record that it is not evaluated quantitatively.

The three modalities the main table omits. Table 8 reports anterior-segment OCT, external eye photography and orbital MRI, which Table 2 omits for the reason given in Section F, that they are

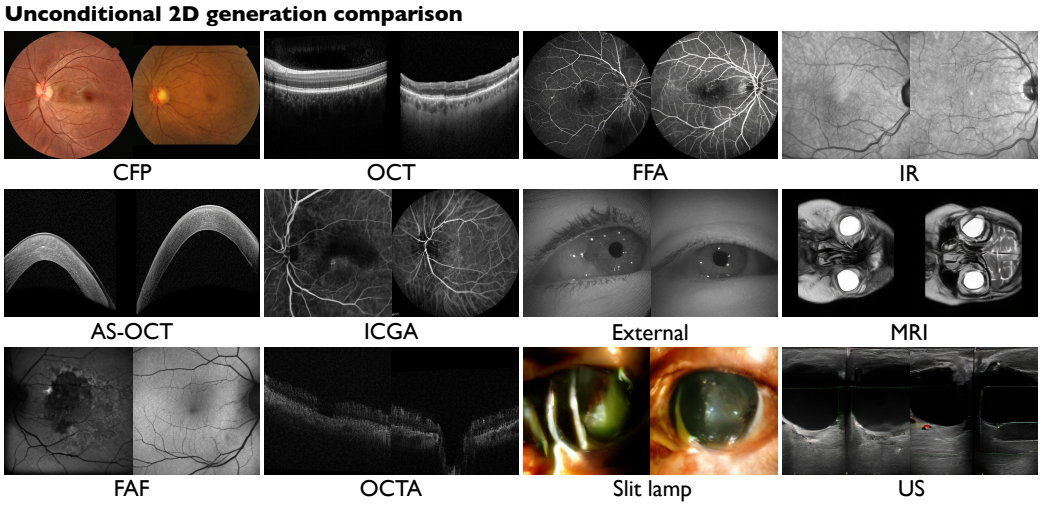

Figure 5: **Unconditional 2D samples across twelve modalities.** Two independent draws each, from one model with one set of weights and no per-modality head. Slit lamp and ultrasound carry no number anywhere in the paper, together being under $0.05\,\%$ of the training corpus and scored in no table. They are shown because the breadth is the claim, and a modality the model reaches without being specialized for it is evidence of that whether or not it has an FID. Anterior-segment OCT, the external eye photograph and the orbital MRI are scored in Table 8, the retinal columns in Table 2. SynOphsis-XL, no conditioning of any kind.

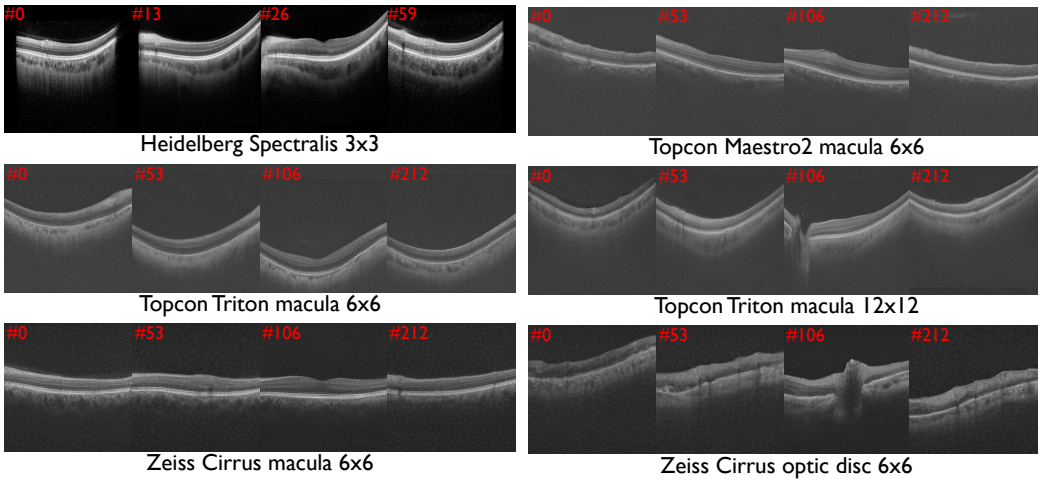

Figure 6: **Unconditional 3D OCT across acquisition protocols.** Six protocols from three vendors, both slab thicknesses and fields of view from 3×3 to 12×12 mm, with the optic disc alongside the macula. Columns are B-scans spanning each volume; the frame indices differ between the panels because the thin protocol standardizes to 60 B-scans and the thick ones to 240. One model produced all six, with the acquisition geometry an input rather than a separate head. Every panel is unconditional.

the modalities furthest from the retinal imaging the paper is about, and the main table keeps to modalities a reader compares like for like. They are evaluated exactly as the other seven (same samples, same held-out splits, same two feature spaces), and they are the three sparsest classes in the training mixture at 1.27%, 0.84% and 0.24% of the 2D images.

They are also where the scale effect is largest. SynOphsis-XL is best of the five on all three modalities in both feature spaces, and all three of its absolute FID margins over the 2D LDM baseline (external eye photography $91.1 \rightarrow 44.4$, orbital MRI $145.1 \rightarrow 79.7$, AS-OCT $184.4 \rightarrow 133.5$)

Table 8: Unconditional 2D generation on the three modalities held out of Table 2: anterior-segment OCT, external eye photography and orbital MRI. Same evaluation, same samples and same layout as the non-OCT half of that table: Inception FID on the Standard rows, ophthalmic DINOv3-Large FID on the DINO rows, lower better, bold best in column and underline second. n=1000 generated samples per modality, 746 for AS-OCT.

Model	Feature	AS-OCT	External	MRI
LDM	Standard	184.4	91.1	145.1
	DINO	10.75	4.80	6.70
SiT-B	Standard	188.7	123.3	142.5
	DINO	10.68	5.64	4.78
SynOphsis-B	Standard	185.1	105.6	144.5
	DINO	10.94	5.18	5.86
SiT-XL	Standard	<u>138.9</u>	<u>44.9</u>	<u>80.4</u>
	DINO	<u>10.30</u>	<u>2.88</u>	<u>3.38</u>
SynOphsis-XL	Standard	**133.5**	**44.4**	**79.7**
	DINO	**10.27**	**1.86**	**3.32**

Table 9: Kernel Inception Distance on the non-OCT modalities, $\times 10^3$, lower better. The same generated samples and the same two feature spaces as Table 2, which reports FID on them; KID is reported here because it is unbiased at these sample sizes and FID is not. Bold best in column, underline second. Ext. is external eye photography; OCTA-enf is OCTA en-face.

Model	Feature	CFP	FAF	FFA	ICGA	IR/SLO	OCTA	OCTA-enf	AS-OCT	Ext.	MRI
LDM	Standard	17.3	37.1	34.5	21.9	31.5	16.6	41.4	136.8	48.5	121.2
	DINO	9.5	7.5	17.6	12.5	7.0	8.7	**8.9**	15.1	7.2	10.7
SiT-B	Standard	26.5	51.8	39.9	38.1	48.3	27.5	47.8	151.2	102.7	85.0
	DINO	12.7	12.2	31.5	30.1	15.5	10.5	21.9	**13.2**	13.7	8.0
SynOphsis-B	Standard	14.7	78.4	45.7	53.3	94.7	25.4	61.9	140.7	74.6	94.5
	DINO	10.2	8.3	24.5	24.7	9.9	8.6	22.5	<u>14.2</u>	10.2	12.1
SiT-XL	Standard	**5.0**	**5.9**	**7.2**	**7.9**	**8.0**	**3.4**	**16.4**	<u>103.3</u>	<u>15.1</u>	**37.0**
	DINO	**5.7**	<u>3.5</u>	**8.4**	<u>6.4</u>	<u>3.6</u>	<u>5.5</u>	10.3	14.9	<u>2.4</u>	**6.1**
SynOphsis-XL	Standard	<u>5.3</u>	<u>8.0</u>	<u>7.7</u>	<u>8.2</u>	<u>8.8</u>	<u>4.0</u>	<u>21.8</u>	**93.0**	**14.7**	<u>38.1</u>
	DINO	<u>5.8</u>	**3.3**	<u>8.6</u>	**6.0**	**3.0**	**5.4**	<u>10.2</u>	14.4	**1.6**	<u>6.3</u>

exceed the largest in the main table, $60.8 \rightarrow 30.9$ on FFA. This is a statement about capacity and not about these modalities. The extra capacity is spent where a single-modality prior helps least, which is consistent with the joint objective buying breadth and not per-modality specialization. It does not make them solved. AS-OCT is still the hardest modality in the paper at 133.5, a 0.72 ratio to the baseline against 0.49 for external eye photography, and none of the five models comes near its retinal scores.

KID and PRDC. The paper reports FID in the main text because it is the metric the baselines report, but FID is a biased estimator at these sample sizes. Table 9 gives Kernel Inception Distance on the same generated samples in the same two feature spaces. It reproduces the FID ordering of the five models exactly on 14 of the 20 modality-by-feature-space cells, and five of the six differences are a swap between two adjacent models. The one that is not is AS-OCT in our feature space, where FID ranks SynOphsis-XL first and KID ranks SiT-B first, but all five models fall between 13.2 and 15.1 there, so neither ordering means much. In Inception features an XL model is best on all ten modalities under both metrics.

Table 10 decomposes that into precision, recall, density and coverage. The ranking largely survives. The two XL models take the first two places on all four quantities in Inception features, and on three of four in ours, the exception being recall, where the 2D LDM baseline is highest at 0.612. The fidelity–diversity split does not survive. Averaged over the ten non-OCT modalities, SynOphsis-XL scores precision 0.637 and recall 0.679 in Inception features against 0.590 and 0.490 in ours, so our encoder judges the same samples markedly less diverse. On CFP alone it is sharper, $0.79/0.79$ against $0.92/0.40$, and on OCT the sign reverses, the same model scoring $0.43/0.54$ on OCTDL in Inception features and $0.10/0.90$ in ours, precision and recall swapping which one is high.

The two spaces therefore read the fidelity–diversity split differently, and we argue only from the ranking, which they agree on. Precision and recall are properties of a set of samples *and* a feature

Table 10: Precision, recall, density and coverage on the non-OCT modalities, averaged over the ten modalities of Table 9, k=5 nearest neighbors. Higher is better. Standard rows use Inception-v3 features and DINO rows our ophthalmic DINOv3-Large features, exactly as for FID. The two feature spaces largely agree on the ranking and disagree on the fidelity–diversity split; Appendix C.1 reads that. Bold best in column, underline second.

Model	Feature	Precision	Recall	Density	Coverage
LDM	Standard	0.445	0.663	0.281	0.390
	DINO	0.466	**0.612**	0.264	0.284
SiT-B	Standard	0.392	0.348	0.226	0.270
	DINO	0.329	0.201	0.161	0.109
SynOphsis-B	Standard	0.328	0.425	0.180	0.234
	DINO	0.331	0.220	0.143	0.122
SiT-XL	Standard	**0.640**	**0.680**	<u>0.571</u>	<u>0.630</u>
	DINO	<u>0.579</u>	0.474	<u>0.382</u>	<u>0.406</u>
SynOphsis-XL	Standard	<u>0.637</u>	<u>0.679</u>	**0.577**	**0.632**
	DINO	**0.590**	<u>0.490</u>	**0.383**	**0.414**

Table 11: Precision and recall per modality, k=5, higher is better. The two XL models only, and two of the four PRDC quantities; Table 10 gives all four averaged over the ten modalities, and the full grid is 400 numbers. Each cell is precision / recall. **The two feature spaces disagree about the fidelity–diversity split, and the disagreement is not a constant offset**. On CFP the Standard rows read $0.79/0.79$ against $0.92/0.40$ in ours, so our encoder judges the same samples far less diverse, while on OCTDL, reported in Table 2 rather than here, the sign reverses. Both spaces agree on the ranking of the models, which is the part the paper argues from.

Model	Feature	CFP	FAF	FFA	ICGA	IR/SLO	OCTA	OCTA-enf	AS-OCT	Ext.	MRI
SiT-XL	Standard	0.79/0.78	0.72/0.77	0.86/0.60	0.79/0.63	0.71/0.76	0.80/0.75	0.79/0.64	0.23/0.81	0.32/0.58	0.40/0.48
	DINO	0.91/0.38	0.82/0.51	0.82/0.20	0.62/0.41	0.72/0.34	0.76/0.71	0.53/0.43	0.12/0.88	0.26/0.64	0.22/0.25
SynOphsis-XL	Standard	0.79/0.79	0.72/0.75	0.88/0.60	0.78/0.62	0.69/0.73	0.80/0.75	0.75/0.65	0.23/0.85	0.34/0.59	0.39/0.45
	DINO	0.92/0.40	0.82/0.52	0.81/0.20	0.64/0.43	0.74/0.43	0.79/0.72	0.52/0.42	0.14/0.85	0.31/0.71	0.19/0.23

space, not of the samples alone, so a fidelity-versus-diversity story told from one embedding does not transfer to another. The rank ordering of the models is what agrees across both, and it is the only part of this we argue from.

Table 11 shows the disagreement rather than quoting it. It drops to the two XL models and to precision and recall, which is what fits and where the disagreement lives, and the columns run left to right from the most retinal modality to the least. The Standard and DINO rows do not differ by a constant. On CFP our encoder reads far lower recall (0.40 against 0.79), on OCTA the two agree closely, and on AS-OCT it reads markedly lower precision at the same or higher recall. So the split cannot be read off one embedding and carried to another.

The same two metrics on OCT. Table 12 gives KID and precision/recall on the OCT benchmarks and the volumes, which Table 2 reports neither of.

KID and FID do not agree on these splits, and we do not present KID as a confirmation of FID. The full five-model ordering matches on only three of the ten benchmark-by-space cells, and the best model changes on three: Kermany-H in both spaces, and OCTID in ours. The direction is consistent. KID moves the win towards the joint-token models and never away from them, and the plain SiT rows are ours as well. Under Inception FID a SynOphsis model is best on four of the five 2D splits and under KID on four as well, but in the ophthalmic space FID gives SiT-B three of five and KID gives it one. KID is the unbiased estimator at these sample sizes and FID is not, so where they differ KID is the better-founded number; we report FID in the main text because that is what the baselines report, and we do not lean on the difference.

On the volumes KID separates the models more sharply than FVD does: SynOphsis-XL reaches 98.6 on AI-READI against the 3D LDM baseline's 246.1, a ratio of 0.40 where FVD gives 52.2 against 51.0 and separates nothing.

The volumes are also where the feature spaces disagree most, and the row that matters is not in the table. Generated thick *OCTA* reaches precision 0.11 and coverage 0.09 in our feature space against

Table 12: KID and precision/recall on the OCT benchmarks, the metrics Table 2 does not report. Same runs, same samples, same two feature spaces. KID is $\times 10^3$ and lower is better; precision and recall are $k{=}5$ and higher is better, given as precision / recall. KID is reported for the 2D splits and for both volume cohorts.

		Kernel Inception Distance, $\times 10^3$						
Model	Feature	OCTDL	OCTID	Kermany-T	Kermany-H	Internal OCT	AI-READI	Internal OCT
LDM	Standard	92.3	92.2	64.3	74.6	34.0	246.1	104.8
	DINO	20.6	15.7	8.5	13.9	4.9	4.4	2.8
SiT-B	Standard	92.7	119.2	52.5	58.1	35.3	188.8	109.8
	DINO	**17.4**	13.7	8.0	7.6	1.2	2.5	0.4
SynOphsis-B	Standard	81.5	**92.1**	39.2	42.9	16.3	181.5	77.6
	DINO	19.1	**12.4**	**7.7**	7.5	0.8	**1.0**	0.3
SiT-XL	Standard	71.6	121.0	**38.1**	36.0	6.6	105.0	89.5
	DINO	19.3	14.5	7.8	7.8	**0.5**	1.6	0.3
SynOphsis-XL	Standard	**67.3**	115.3	45.1	**31.9**	**5.7**	**98.6**	**44.9**
	DINO	18.0	21.5	7.8	**7.2**	**0.5**	1.5	**0.2**

		Precision / recall						
Model	Feature	OCTDL	OCTID	Kermany-T	Kermany-H	Internal OCT	AI-READI	Internal OCT
LDM	Standard	0.27/0.37	0.21/0.36	0.17/0.35	0.21/0.66	0.54/0.77	0.25/0.61	0.75/0.78
	DINO	0.07/0.61	0.07/0.91	0.19/0.71	0.28/0.91	0.80/0.92	0.74/0.90	0.77/0.94
SiT-B	Standard	0.50/0.11	0.56/0.08	0.39/0.13	0.47/0.39	0.75/0.45	0.45/0.32	0.87/0.58
	DINO	0.11/0.56	0.15/0.79	0.20/0.61	0.46/0.91	0.91/0.85	0.84/0.78	0.97/0.85
SynOphsis-B	Standard	0.51/0.28	0.47/0.14	0.38/0.25	0.60/0.52	0.77/0.65	0.39/0.58	0.89/0.84
	DINO	0.11/0.76	0.14/0.83	0.23/0.69	0.43/0.97	0.90/0.90	0.90/0.92	0.98/0.92
SiT-XL	Standard	0.47/0.38	0.55/0.21	0.35/0.41	0.47/0.65	0.77/0.76	0.63/0.74	0.79/0.82
	DINO	0.09/0.77	0.18/0.87	0.20/0.73	0.42/0.93	0.90/0.92	0.83/0.88	0.97/0.95
SynOphsis-XL	Standard	0.43/0.54	0.45/0.20	0.33/0.46	0.44/0.78	0.74/0.79	0.56/0.83	0.80/0.89
	DINO	0.10/0.90	0.13/0.86	0.19/0.68	0.39/0.97	0.90/0.92	0.89/0.91	0.93/0.95

0.93 and 0.98 for thin structural volumes from the same model. That is the quantitative form of the exclusion Appendix G argues from images, and it is consistent with what those images show, the volumes carrying flow structure while the paired output does not.

Unconditional metrics in independent encoders. The DINO-prefixed columns elsewhere in the paper are computed in an encoder we trained, and gains measured inside the feature space a model was aligned to can fail to survive independent evaluation (Wakefield-Skórniewska & Papież, 2026). We therefore repeated the unconditional evaluation with four ophthalmic foundation models trained by other groups on other data, using each in turn as the FID and KID feature extractor: RET-Found (Zhou et al., 2023) in its masked-autoencoder and its DINOv2-pretrained (Oquab et al., 2024) releases, VisionFM (Qiu et al., 2024) and FLAIR (Silva-Rodríguez et al., 2025). This gives 21 benchmark-by-encoder combinations, reported below by what is generated: the non-OCT modalities (Table 13), the 2D OCT benchmarks (Table 14) and the volumes (Table 15).

The result holds in every encoder. SynOphsis-XL leads the LDM baseline in 20 of the 21 combinations on FID and 15 on KID, and SynOphsis-XL leads SynOphsis-B in 19 of 21 on both. Against the plain SiT of the same size the picture is the one Appendix C.1 reports in our own feature space: SynOphsis-XL leads in 11 of 21 on FID and 10 on KID. The gain credited to the alignment objective is therefore visible to encoders never trained towards it, and the trade between the two feature spaces is not an artifact of ours.

C.2 Conditional case studies

Figure 3 in the main text shows one case with a mild focal abnormality. The three figures below vary what the model is given rather than repeating one setting. They show a single observed B-scan on a severe case, the paired 2D image alone, and the first frame of a thick volume, which asks for extrapolation across the whole slab from one end.

All four share a layout. Each column is one B-scan and its header gives the frame index, with the given frames in red and labeled *Given*; where no frame is given the label sits on the 2D panel instead. Rows are methods, bracketed by provenance as in the tables, and *Not applicable* marks what a method cannot emit at all. The last two columns are a magnification of the frame where the methods disagree most and each method's paired 2D output.

Table 13: **Unconditional FID and KID on the non-OCT modalities, computed with existing ophthalmic foundation models.** Four encoders trained by other groups on other data: FLAIR (Silva-Rodríguez et al., 2025), RETFound (Zhou et al., 2023) in both its masked-autoencoder and its DINOv2-pretrained (Oquab et al., 2024) release, and VisionFM (Qiu et al., 2024). Column groups are the encoder; the row beneath names the benchmark it was run on. Lower is better for both. Bold best in column, underline second. Magnitudes are not comparable between columns, since feature width runs from 512 to 3072 and FID from 0.06 to 2.5×10^4, so only the ordering within a column carries meaning. KID is unscaled here, unlike Table 9, because these encoders do not normalize their features, so no one multiplier suits all four.

		FLAIR	RETFound-MAE	RETFound-DINOv2	VisionFM			
Model	Metric	CFP	CFP	CFP	CFP	FFA	External	MRI
LDM	FID	**0.061**	25.6	852.4	8 896	6 221	7 278	4 964
	KID	**0.0002**	0.096	0.240	17.4	13.6	29.1	31.9
SiT-B	FID	0.172	64.4	970.3	10 766	7 448	7 929	5 080
	KID	0.0006	0.256	0.400	26.5	24.5	48.8	27.1
SynOphsis-B	FID	0.175	14.3	1 010	9 401	7 580	9 773	5 429
	KID	0.0007	0.033	0.509	18.4	24.5	62.6	30.7
SiT-XL	FID	0.104	**7.87**	**751.6**	**6 976**	**3 527**	3 593	2 676
	KID	0.0004	**0.015**	**0.094**	7.19	**4.85**	**9.94**	12.0
SynOphsis-XL	FID	0.101	9.27	760.4	7 011	3 628	**3 434**	**2 638**
	KID	0.0004	0.021	0.103	**7.11**	5.30	9.99	**11.5**

Table 14: **Unconditional FID and KID on the 2D OCT benchmarks**, in the two existing foundation models with OCT-pretrained weights, RETFound-MAE (Zhou et al., 2023) and VisionFM (Qiu et al., 2024). Each is run on the same five benchmarks, so the two halves are directly comparable. Lower is better for both; bold best in column, underline second; magnitudes are not comparable between columns. Read with Table 13, which states the conventions in full and what the three tables jointly support.

		RETFound-MAE					VisionFM				
Model	Metric	Kermany-H	Kermany-T	OCTDL	OCTID	Internal OCT	Kermany-H	Kermany-T	OCTDL	OCTID	Internal OCT
LDM	FID	110.4	95.2	228.0	164.9	63.0	25 414	12 452	15 758	33 445	4 115
	KID	0.229	**0.146**	0.545	0.251	0.163	585.7	**410.1**	603.5	729.7	56.7
SiT-B	FID	56.0	100.7	**86.4**	**39.5**	31.9	13 268	13 458	**11 383**	33 713	3 660
	KID	0.191	0.406	**0.304**	**0.112**	0.105	480.1	599.8	**448.6**	739.5	102.0
SynOphsis-B	FID	56.1	104.6	93.2	73.5	24.4	**11 830**	10 948	12 688	32 710	2 330
	KID	0.186	0.419	0.327	0.231	0.083	389.7	481.2	523.0	716.0	40.1
SiT-XL	FID	**50.4**	**88.2**	98.6	114.9	**17.9**	13 019	**10 175**	12 541	**27 729**	1 935
	KID	**0.157**	0.347	0.353	0.480	**0.058**	**382.1**	414.3	503.5	384.0	**23.7**
SynOphsis-XL	FID	52.5	89.9	88.7	129.6	21.6	12 867	10 203	12 052	28 377	**1 856**
	KID	0.162	0.348	0.310	0.523	0.072	394.1	411.0	473.2	**358.3**	23.8

The joint-token models lead the conditioning baselines on all three, but absolute quality falls with lesion size. On the severe case in Figure 7, no method reconstructs the neovascular complex faithfully, ours included.

C.3 Agreement between a generated volume and the en-face view generated with it

Figures 7 to 9 show a volume and the en-face view drawn with it along one trajectory. The figures do not establish that the two halves describe one eye, and this subsection measures it.

Averaging each B-scan of a volume along its depth axis gives its en-face projection. On a real acquisition the projection and the en-face image acquired with the volume show the same retina and correlate. The measurement below rests on that correlation.

For each of the 100 held-out cases we correlate the generated en-face image against the en-face projection of all 100 generated volumes and record the position of the case's own volume in the resulting ranking. An en-face output carrying no information about the volume beside it gives a median position of 50.5, and an output showing the same eye gives a position near the top. We repeat the measurement on real data, correlating each acquired en-face image against the en-face projection of all 100 real volumes, which gives the position a real acquisition reaches.

Table 15: **Unconditional FID and KID on generated OCT volumes**, in the same two OCT-pretrained foundation models. Each entry averages per-volume scores over the hundred volumes; a per-volume score comes from that volume's ten extracted slices and is therefore small-sample biased, but by the same amount for every model, since every model contributes the same ten slices. The comparison down a column holds; the magnitude is inflated. Lower is better for both; bold best in column, underline second; magnitudes are not comparable between columns. Read with Table 13, which states the conventions in full and what the three tables jointly support.

		RETFound-MAE		VisionFM	
Model	Metric	Thin OCT	Thick OCT	Thin OCT	Thick OCT
LDM	FID	49.8	77.4	5 397	12 240
	KID	**0.064**	**0.179**	42.8	133.5
SiT-B	FID	55.2	**68.9**	2 763	8 083
	KID	0.213	0.286	55.8	64.0
SynOphsis-B	FID	59.9	122.2	3 135	8 417
	KID	0.224	0.516	88.9	91.4
SiT-XL	FID	59.3	74.6	2 773	7 159
	KID	0.235	0.303	56.1	38.6
SynOphsis-XL	FID	**46.4**	70.6	**2 450**	**6 829**
	KID	0.172	0.270	**35.2**	**32.4**

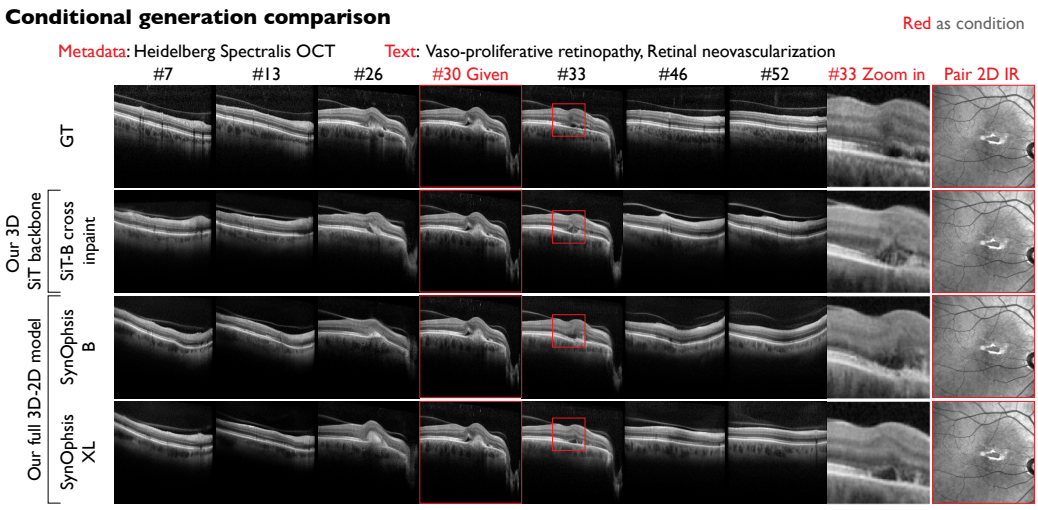

Figure 7: **Severe case, one observed B-scan.** Vaso-proliferative retinopathy with retinal neovascularization, on a Heidelberg Spectralis volume. Only B-scan #30 is given, in red; the rest of the volume and the paired 2D infrared image are generated. The zoom column magnifies #33, where the methods disagree most. Rows are bracketed by provenance as in the tables, here our 3D SiT backbone under cross-attention inpainting against the full joint-token model at both sizes. The joint-token models hold the neovascular complex; the backbone alone smooths it away.

Real acquisitions reach 19.5 of 100. SynOphsis-XL reaches 18.0 under Mid and 17.5 under Span, and SynOphsis-B 21.0 under both, so a volume and the en-face view generated with it agree about as closely as a volume and the en-face image acquired with it.

The measurement shows a consistent trend across the set rather than a strong match on an individual case, for both the generated pairs and the real ones alike. Real acquisitions place the correct volume first in 5 of 100 cases, so what the table supports is agreement in aggregate rather than identification case by case. On thin volumes the measurement carries no information at all: real pairs reach 49.5 of 100 against a chance value of 50.5. A thin acquisition has 61 B-scans against 350 to 512 in the thick cohort, and averaging 61 B-scans leaves a projection too coarse to retain the vessels the correlation depends on. Thin is therefore absent from the table.

The three conditions that receive the en-face view as an input generate no en-face image, so four of the seven appear. The 3D LDM and both conditioning variants generate none under any condition.

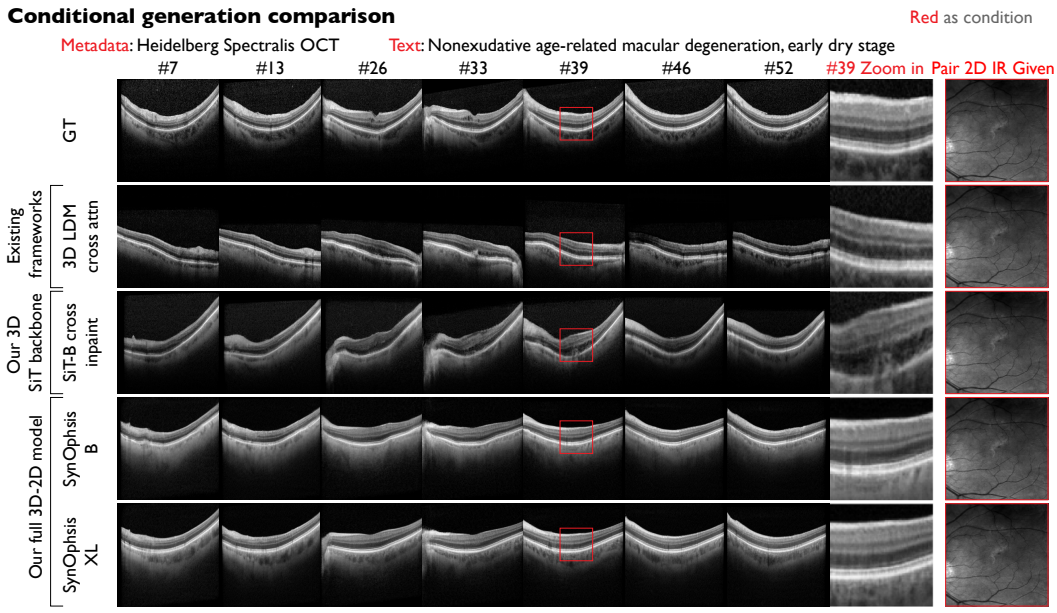

Figure 8: **Near-normal case, the paired 2D image as the only condition.** Nonexudative age-related macular degeneration, early dry stage, on a Heidelberg Spectralis volume. *No B-scan is given*. This is the 2D-only task, and the infrared en-face image at the right is the whole of the conditioning, so every column is generated. The easier pathology and the harder conditioning pull in opposite directions, so the contours stay close across methods while the 3D LDM baseline loses the layer structure altogether.

Table 16: **Agreement between the volume and the en-face view generated with it.** Averaging each B-scan of a volume along its depth axis gives its en-face projection. Each generated en-face image is correlated against the en-face projection of all 100 generated volumes, and the table gives the position of the case's own volume in the resulting ranking. Lower is better. The real row repeats the measurement on volumes and en-face images acquired together, giving the position a real acquisition reaches. Held-out thick OCT, n=100. Top 10 is the share of cases whose own volume lands in the first ten. z measures the mean rank against its chance distribution.

Model	Condition	Median rank	Top 10	z
Real co-acquired pair		19.5	34%	8.76
SynOphsis-XL	Mid	18.0	33%	9.20
SynOphsis-XL	Span	17.5	36%	9.02
SynOphsis-XL	First	22.0	29%	7.58
SynOphsis-XL	Last	20.5	36%	8.08
SynOphsis-B	Mid	21.0	25%	8.10
SynOphsis-B	Span	21.0	26%	8.00
SynOphsis-B	First	24.5	29%	6.58
SynOphsis-B	Last	18.5	27%	6.72
Chance		50.5	10%	0.00

C.4 Effect of the observation budget

Figures 10 and 11 vary how much of the volume is given while holding the case and the shown B-scans fixed. Only SynOphsis-XL is drawn; the point is what changes with the input, not a ranking. Each row carries the input it was given, so the growing red band is the variable.

Per-case PSNR is noisy. Over the 14 conditional cases SynOphsis-XL goes from 19.32 to 19.92 dB as the budget grows from 6 to 21 of 60 B-scans and SynOphsis-B from 18.97 to 19.70 dB. Individual cases, however, depart from this, one of them losing 1.8 dB at the largest budget. The numbers printed in the figures are therefore illustration. Table 17 is the evidence, the same setting swept

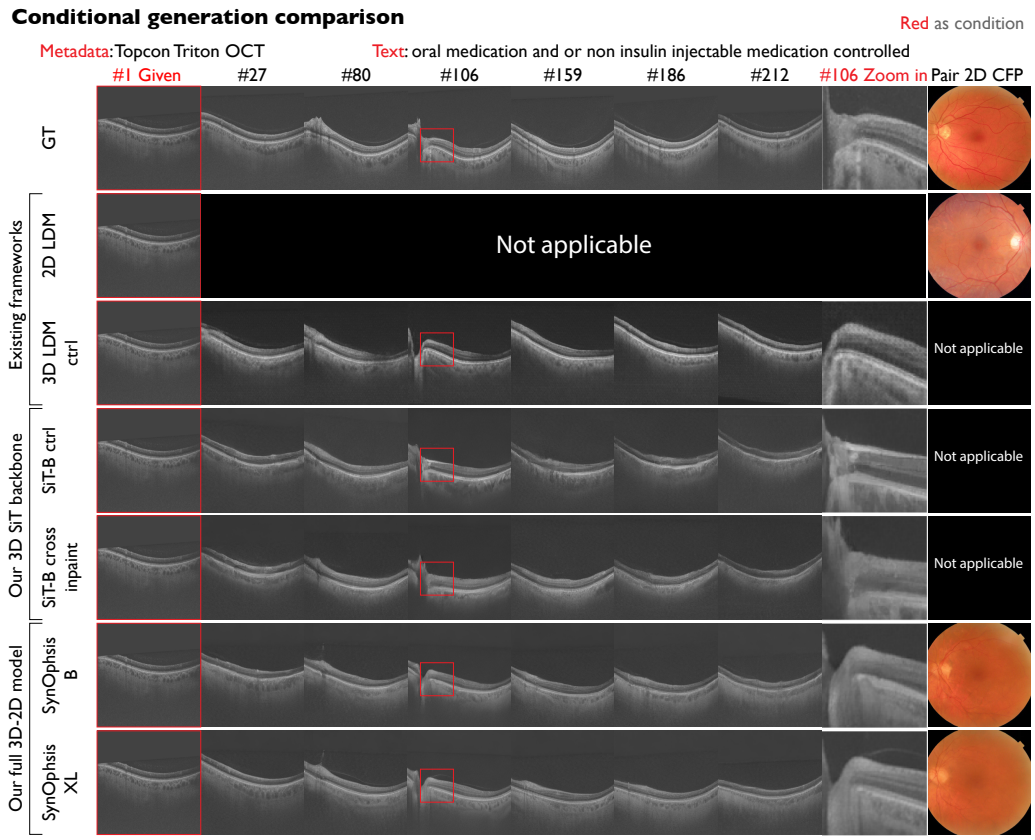

Figure 9: **Thick OCT volume, the first B-scan observed.** A Topcon Triton acquisition standardized to a thick slab, with #1 given and columns running to #212, extrapolation across the whole volume from one end, the hardest of the placements. Rows carry both existing frameworks, the 2D LDM, which emits no volume, and the 3D LDM under control conditioning, which emits no paired 2D image; *Not applicable* marks each. Only the joint-token model produces the volume and its color fundus photograph together.

over *100 volumes* with 3D conditioning only, at seven budgets blocked at the center and four spread across the volume.

Budget alone buys little. A fourteenfold increase, from 3 to 42 of 60, is worth 1.76 dB. *Placement* buys much more. With twelve observed frames, spreading them through the volume rather than giving them as a block raises PSNR from 20.14 to 22.82 dB (+2.68). With thirty frames it rises from 20.88 to 24.91 dB (+4.03). A block at the center leaves both ends to extrapolate, which is the harder half of the problem; spreading the same evidence turns extrapolation into interpolation. This is why the presets in Section B.2 vary the pattern and not only the count.

The limit of Figure 11's axis is zero B-scans with the 2D image still given, which is the 2D-only task of Section C.5. It is not drawn as a row of that figure because those outputs were sampled over a different set of cases, none of the three matching any of the fourteen used here, so a zero-B-scan row would change the eye as well as the budget. We keep the two figures separate instead of re-sampling it on these cases. Table 3 carries the quantitative statement, over a hundred volumes per protocol, and merging the figures would buy layout at the cost of a full re-run.

Giving the paired 2D image as well is worth a further +0.14 to +0.23 dB, consistently at every budget but small beside the effect of the budget itself, since the 2D image constrains the en-face appearance of the volume and not the depth structure the unobserved B-scans are mostly made of.

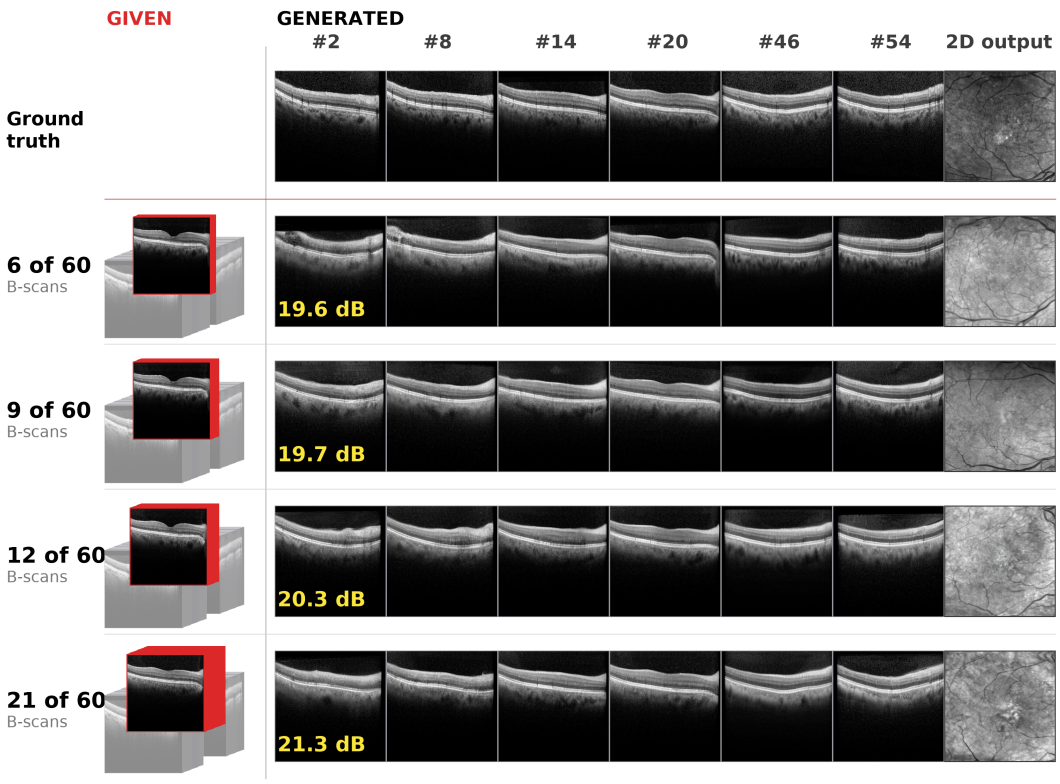

Figure 10: **Growing the observed fraction of the volume.** The vertical rule splits what the model was given from what it produced. Each row gives SynOphsis-XL a contiguous block of B-scans at the center of the volume, growing outwards from 6 to 21 of 60; the deck on the left draws that volume with the given B-scans lifted out of it in red. Everything right of the rule is generated, namely six peripheral B-scans, the same ones in every row and unobserved in all of them, and the paired 2D image, which the model emits because none was given to it. The ground-truth 2D image is in the same column of the top row. PSNR is over all unobserved B-scans of this case. The case is a thin OCT volume from the internal data on a Heidelberg Spectralis. No text condition and no paired 2D image are given, so the volume is generated from the observed B-scans alone. The case carries a clinical diagnosis of type 2 diabetes mellitus with proliferative diabetic retinopathy without macular edema, bilateral; the volume shown is the right eye.

C.5 The volume from the 2D image alone

The 2D-only task removes the B-scans entirely. The model is given the paired 2D en-face image and nothing of the volume, and must produce all 60 B-scans. No part of the output is anchored to an observed slice, so the question is whether the volume is a plausible eye consistent with that en-face image and not whether it reproduces the specific one that was scanned.

By reconstruction metrics, the task scores low. Table 3 puts SynOphsis-XL at 14.17 dB on the thin protocol and 19.94 dB on the thick one, against 17.8–24.3 dB and 22.0–26.8 dB on the tasks of the same protocol where B-scans are observed. The ordering is still ours on both protocols, by 0.30 dB over the cross-attention inpainting baseline on thin and 0.28 dB on thick over 100 volumes each, and both margins are small, so we report it as a capability the formulation makes available and not as a win.

Spread across volumes, and whether the margins survive it. Every cell of Table 3 is a mean over 100 volumes, and the volume-to-volume spread is larger than the margins: Table 18 gives a standard deviation of 1.4 to 2.3 dB against margins of a few tenths. Comparing the means alone would therefore settle nothing, which is why the margins are tested pairwise, on the same volumes, rather than read off the table.

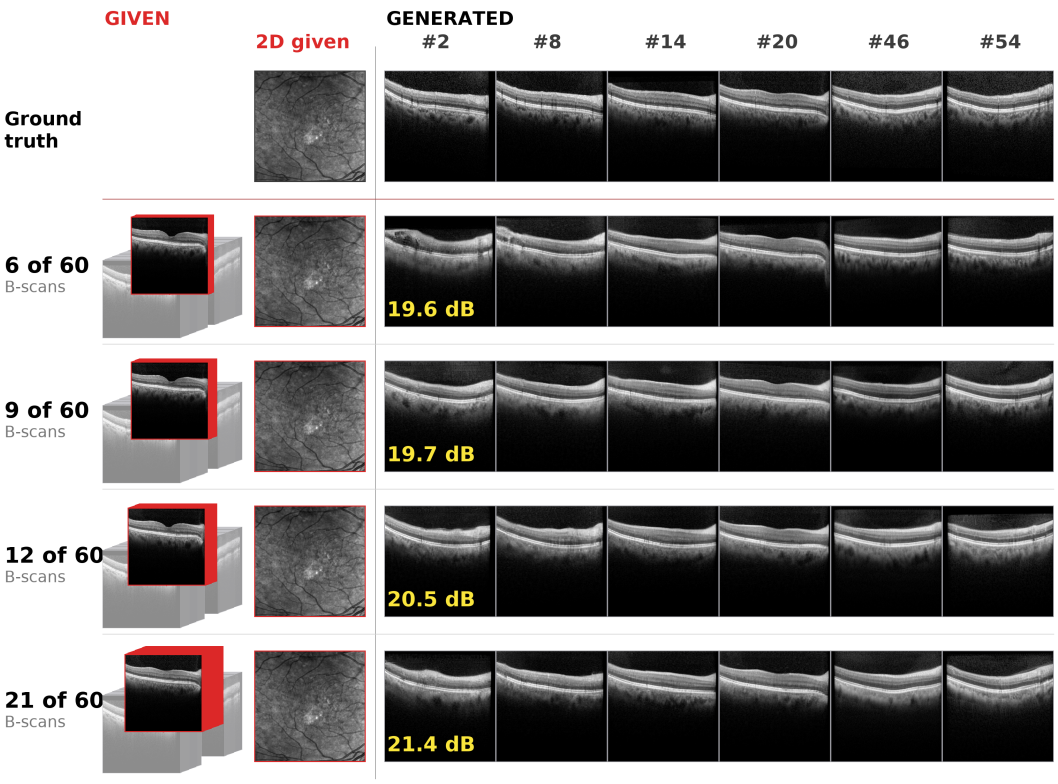

Figure 11: **The same budgets, with the paired 2D image also given.** Identical case, budgets and shown B-scans as Figure 10, with one input added, the infrared en-face photograph acquired alongside the volume, drawn red-bordered beside each deck on the given side of the rule. It is the conditioning image handed to the sampler, not an output, and nothing is generated in that column here. Conditioning on the 2D image means the model does not produce one, which is why the generated side holds only B-scans. Comparing the two figures row by row isolates what the paired 2D image contributes on top of the B-scans, which across the 14 cases is $+0.14$ to $+0.23$ dB. Same device, data source and absence of a text condition as Figure 10.

Against the strongest other method on each task, SynOphsis-XL's margin is significant on twelve of the fourteen. The two that are not are task 2A on both slabs, -0.04 and $+0.01$ dB, where no B-scan is observed and the volume is not determined by the input. That is the task where a per-volume PSNR asks the wrong question, for the reason Figure 12 shows, and the test agrees with the figure.

C.6 OCTA

OCTA is trained as one more 240-B-scan protocol and evaluated with the rest. An OCTA volume is darker than a structural one and carries a flow signal in place of the full layered reflectivity.

Generated Zeiss OCTA carries the vertical flow striation of the modality and is darker, as OCTA should be. Over forty generated volumes of each, the mean intensity is 0.058 for OCTA against 0.251 for structural OCT (Figure 2). The paired view carries an optic nerve head inside a vessel arcade.

Table 19 reports the volumes over the same seven tasks and the same evaluation as Table 3. SynOphsis-XL leads the 3D LDM baseline by 1.36 to 2.46 dB on the four slice-conditioned tasks and leads the cross-attention variant on both combined tasks. It trails the cross-attention LDM on the 2D-only task, 18.58 against 19.09 dB, the one task on this protocol where no B-scan is observed.

Table 20 reports the paired 2D output beside the thick structural case, over the four tasks that produce one. SynOphsis-XL leads the 2D LDM baseline by 2.95 dB and 0.14 SSIM and ties it on D-LPIPS. Its PSNR is 1.06 dB below the structural case and its SSIM is above it on every task.

Table 17: **Observation budget, and where the observations sit.** SynOphsis-XL on the thin protocol, *100 volumes* per row, SDE at 100 steps. Conditioning is 3D only, so the paired 2D columns score what the model emitted rather than anything it was given; volume metrics are over the unobserved B-scans alone. *Where* gives the raw frames covered for a contiguous block and the approximate raw spacing for a spread one. D-LPIPS is $10^3\times$DINO-LPIPS, as in Table 3. Placement beats budget. Growing the block fourteenfold, 3 to 42 of 60, buys 1.76 dB, while moving the same twelve frames from a block to a spread buys 2.68 ($20.14 \rightarrow 22.82$) and thirty frames 4.03 ($20.88 \rightarrow 24.91$).

		Volume			*Emitted paired 2D*		
Given	Where	PSNR ↑	SSIM ↑	D-LPIPS ↓	PSNR ↑	SSIM ↑	D-LPIPS ↓
Contiguous block at the center							
3 of 60	30–32	19.57	0.480	1.5	14.35	0.362	13.8
6 of 60	27–32	19.87	0.485	1.4	14.44	0.363	13.7
12 of 60	24–35	20.14	0.487	1.1	14.58	0.366	13.1
24 of 60	18–41	20.60	0.488	0.8	14.82	0.371	12.2
30 of 60	15–44	20.88	0.489	0.7	14.96	0.374	11.7
36 of 60	12–47	21.10	0.489	0.5	15.06	0.377	11.7
42 of 60	9–50	21.33	0.489	0.5	15.15	0.377	11.1
Spread across the volume							
12 of 60	∼15	22.82	0.555	0.7	15.19	0.381	12.3
15 of 60	∼10	23.39	0.567	0.7	15.25	0.385	12.2
21 of 60	∼7	24.27	0.585	0.5	15.34	0.386	11.6
30 of 60	∼5	24.91	0.594	0.4	15.43	0.385	11.0

Table 18: **Spread and paired tests on the conditional tasks.** Per-volume PSNR over the 100 evaluation volumes of each protocol. SynOphsis-XL is given as mean $\pm$ standard deviation; Δ is its paired mean margin over the strongest other method on that task, with a 95% bootstrap interval and a Wilcoxon signed-rank p.

Task	SynOphsis-XL	Strongest other	Δ [95% CI]	p
Thin OCT				
1A	19.64 ± 1.49	SynOphsis-B	$+0.29$ $[+0.22, +0.37]$	$< 10^{-4}$
1B	21.73 ± 1.59	SynOphsis-B	$+0.28$ $[+0.22, +0.34]$	$< 10^{-4}$
1C	17.91 ± 1.57	SynOphsis-B	$+0.15$ $[+0.04, +0.26]$	0.009
1D	18.27 ± 1.39	SynOphsis-B	$+0.16$ $[+0.06, +0.27]$	0.001
2A	13.91 ± 1.51	SynOphsis-B	-0.04 $[-0.15, +0.07]$	0.770
3A	19.73 ± 1.53	SynOphsis-B	$+0.40$ $[+0.30, +0.49]$	$< 10^{-4}$
3B	24.27 ± 1.70	SiT-B Inpaint	$+0.38$ $[+0.32, +0.43]$	$< 10^{-4}$
Thick OCT				
1A	23.25 ± 2.14	SiT-B Ctrl	$+0.36$ $[+0.12, +0.60]$	0.004
1B	26.36 ± 2.30	SiT-B Ctrl	$+0.95$ $[+0.76, +1.12]$	$< 10^{-4}$
1C	22.24 ± 2.13	SynOphsis-B	$+0.31$ $[+0.13, +0.48]$	$< 10^{-4}$
1D	21.96 ± 2.04	SynOphsis-B	$+0.25$ $[+0.07, +0.43]$	0.005
2A	19.94 ± 1.99	SynOphsis-B	$+0.01$ $[-0.13, +0.14]$	0.541
3A	23.46 ± 2.25	SiT-B Inpaint	$+0.76$ $[+0.45, +1.05]$	$< 10^{-4}$
3B	26.83 ± 1.43	SiT-B Inpaint	$+2.24$ $[+2.01, +2.47]$	$< 10^{-4}$

C.7 Sampler and guidance choices

Repeated draws from one input. Figure 12 draws the 2D-only task five times from one en-face image, varying only the seed. This is the right way to read that task. It gives no B-scan, so the volume is not determined by the input, and asking whether one draw reproduces the volume that happened to be scanned, which is what the PSNR in Section C.5 asks, answers a question the task does not pose. The draws differ in the fine structure while keeping the layer geometry the en-face image implies.

Sampler and guidance. The sampler makes little difference. Over eight paired volumes the mean absolute difference between the ODE and SDE outputs is 0.061 on $[0, 1]$ intensities, and the anatomy, curvature and position of each are the same, so the choice is not one the results depend on.

Guidance changes the images monotonically. Against the scale-1 output the mean absolute difference is 0.047, 0.070 and 0.090 at scales 1.5, 2 and 3, and the intensity range, measured from the first to the ninety-ninth percentile of a volume, widens from 0.802 to 0.938 over the same span.

Table 19: **Conditional generation on OCTA volumes.** The tasks and the model rows are those of Table 3. PSNR is higher-better and D-LPIPS is lower-better, with bold marking the best value in a column and underline the second. The paired 2D output for this protocol is reported in Table 20.

		OCTA volume						
Model	Metric	Mid	Span	First	Last	2D	2D+Mid	2D+10%
Existing methods								
3D LDM Ctrl	PSNR	19.54	21.11	18.78	19.26	/	/	/
	D-LPIPS	17.0	11.1	19.3	20.0	/	/	/
3D LDM Cross	PSNR	/	/	/	/	**19.09**	/	/
	D-LPIPS	/	/	/	/	18.8	/	/
Ours: 3D SiT, alternative conditioning								
SiT-B Ctrl	PSNR	21.21	22.63	20.61	20.26	/	/	/
	D-LPIPS	4.0	3.3	5.7	6.4	/	/	/
SiT-B Inpaint	PSNR	21.36	22.65	20.45	20.43	18.90	21.18	23.56
	D-LPIPS	3.8	3.1	6.2	5.9	**9.1**	4.1	0.9
Ours: joint-token								
SynOphsis-B	PSNR	21.44	22.81	21.06	20.59	18.68	21.31	23.61
	D-LPIPS	3.4	2.5	6.0	**5.4**	**9.1**	3.5	1.0
SynOphsis-XL	PSNR	**21.59**	**22.92**	**21.24**	**20.62**	18.58	**21.58**	**23.81**
	D-LPIPS	**2.9**	**2.2**	**5.5**	5.5	9.4	**3.2**	**0.7**

Table 20: **The paired 2D output on OCTA, beside the thick structural case.** The tasks and the model rows are those of Table 3, and the four shown are the tasks that produce a paired image. PSNR and SSIM are higher-better, and D-LPIPS is $10^3\times$DINO-LPIPS and lower-better. Bold marks the best value in a column and underline the second.

		Thick OCT paired 2D				*OCTA paired 2D*			
Model	Metric	Mid	Span	First	Last	Mid	Span	First	Last
Existing methods									
2D LDM	PSNR	14.92	15.41	15.50	15.68	15.39	15.28	15.37	15.28
	SSIM	0.602	0.605	0.599	0.619	0.550	0.548	0.555	0.549
	D-LPIPS	21.0	22.0	22.5	23.2	16.6	**15.3**	16.7	**15.3**
Ours: joint-token									
SynOphsis-B	PSNR	18.63	18.42	17.81	18.18	18.36	17.44	17.76	18.59
	SSIM	0.679	0.670	0.670	0.671	0.686	0.678	0.674	0.691
	D-LPIPS	13.6	13.3	15.0	13.8	**16.1**	15.7	15.5	15.5
SynOphsis-XL	PSNR	**19.43**	**19.88**	**18.92**	**19.15**	**18.47**	**17.81**	**18.13**	**18.72**
	SSIM	**0.691**	**0.683**	**0.679**	**0.682**	**0.696**	**0.685**	**0.687**	**0.699**
	D-LPIPS	**12.3**	**11.9**	**12.8**	**13.1**	16.4	16.8	**15.4**	**15.3**

Those changes in appearance do not carry into the metrics. Table 21 sweeps the scale on SynOphsis-XL over the two benchmarks it generates. Scale 1 is best in five of the eight columns, and in the other three the best scale improves on it by 9% or less. The 2D and volumetric halves prefer the same scale, so one sampler setting serves both outputs. We report every number in this paper at scale 1, the default, for our models and the baselines alike.

C.8 Failure cases

Extrapolation from one end. When the observed B-scans do not span the whole volume, the model has to extrapolate to the ends. Table 3 measures the cost directly. First and Last observe only the first or only the last latent slice and score 17.83 and 18.28 dB on the thin protocol against 19.59 for Mid. That loss of 1.3 to 1.8 dB comes purely from where the evidence sits. The same ordering holds on the thick protocol. Figure 7 shows that case. With six contiguous B-scans at the center every method degrades towards the ends, and on a large dome-shaped elevation with an internal cavity none reconstructs the cavity faithfully, ours included. The joint-token models keep a plausible retinal contour where the baselines break up, a failure that is graceful rather than absent.

Table 21: **Guidance scale.** SynOphsis-XL sampled unconditionally at four classifier-free guidance scales. FID and KID on 2D OCT, FVD and KID on thin volumes; the DINO-prefixed columns are the same metrics in our ophthalmic encoder rather than Inception-v3 and R3D-18. All are lower-better and KID is $\times 10^3$; bold is best in column. †The paper reports scale 1 throughout.

	2D OCT				*3D thin OCT*			
Guidance	FID	DINO-FID	KID	DINO-KID	FVD	DINO-FVD	KID	DINO-KID
1.0†	31.81	**1.584**	**7.58**	**1.07**	**37.79**	1.475	**91.4**	0.563
1.5	**31.44**	1.655	8.48	1.29	38.86	1.421	103.3	**0.515**
2.0	32.02	1.838	9.32	1.59	38.93	1.433	107.6	0.533
3.0	33.44	2.099	10.91	2.00	38.37	**1.364**	111.5	0.551

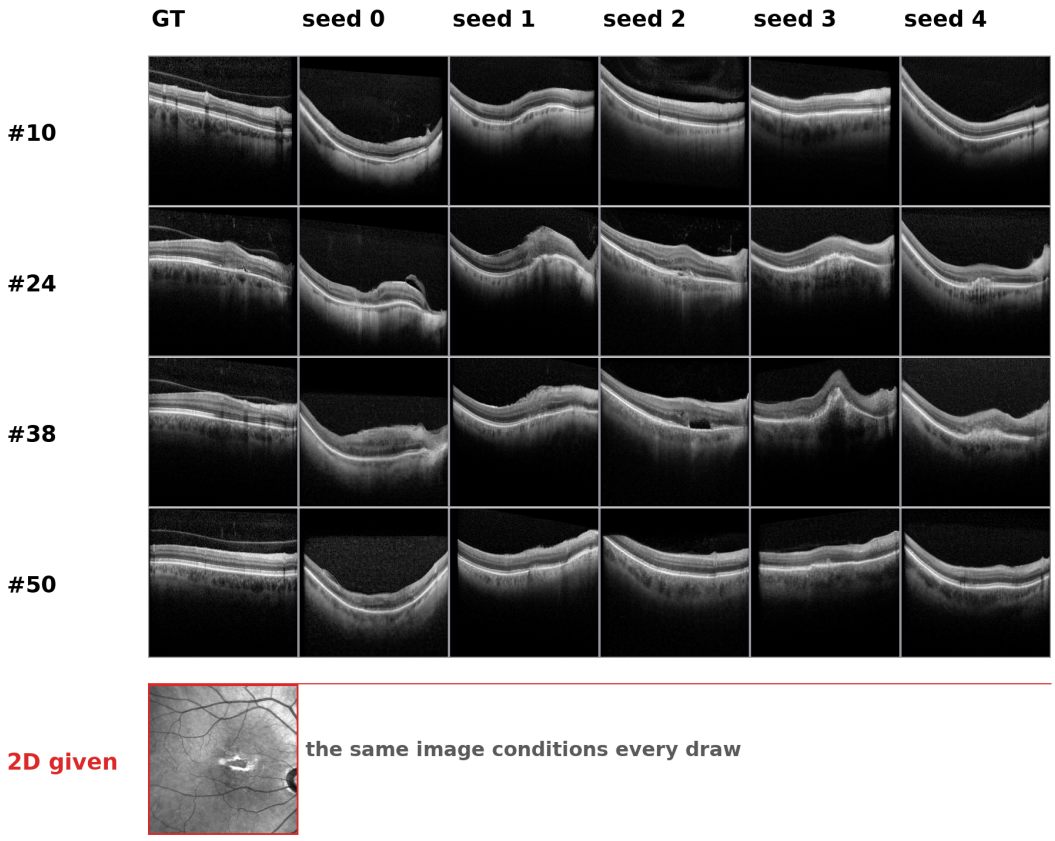

Figure 12: **Five draws from one en-face image.** The 2D-only task with the seed as the only difference between columns; ground truth first, the conditioning image at the bottom. No B-scan is given, so the volume is underdetermined and the draws are alternatives and not attempts at one answer. Thin OCT, Heidelberg Spectralis, no text condition.

D Implementation details

D.1 Ophthalmic DINOv3 encoder

The representation-alignment teacher is a DINOv3-Large (Siméoni et al., 2025) encoder pretrained on ophthalmic images and kept frozen during diffusion training.

Pretraining corpora. The two teachers are pretrained on different subsets of one ophthalmic image collection, neither of which includes any evaluation split used in this paper.

The 2D teacher sees 2,900,649 images, of which 2,148,182 are OCT B-scans, 308,322 color fundus photographs, 221,828 infrared en-face photographs, 88,572 fluorescein angiograms, 65,280 autofluorescence images, 30,794 indocyanine-green angiograms and 37,671 en-face OCTA projections. It

contains no OCTA volumes, no anterior-segment OCT, no MRI and no external-eye or slit-lamp photography.

The 3D teacher continues from it on a larger subset of $15{,}614{,}117$ images, which adds the volumetric material, 11.32M OCT B-scans, 2.41M OCTA frames, 0.76M infrared photographs, 0.41M color fundus photographs, 0.12M fluorescein angiograms, 0.098M corneal images, 0.072M autofluorescence images, 0.049M indocyanine-green angiograms, 0.057M external-eye photographs, 0.042M en-face OCTA projections, 0.011M orbital MRI slices, 0.0067M anterior-segment OCT B-scans, and smaller sets of slit-lamp and iris images. Both subsets are built from the same two cohorts the diffusion models train on, the internal OCT archive and AI-READI, together with public per-modality datasets for the modalities neither cohort covers. The OCT and OCTA majority is frame-level, so one volume contributes many rows.

Pretraining recipe. The 2D ophthalmic run continues the public DINOv3 ViT-L/16 checkpoint pretrained on LVD-1689M, and the 3D run continues the 2D teacher, both with all weights trainable. The objective is the standard DINOv3 pairing of a DINO head and a separate iBOT head, each with $131{,}072$ prototypes, a 384-dimensional bottleneck and three layers of hidden width $2{,}048$, with Sinkhorn–Knopp centering, a Koleo regularizer at weight 0.1, and iBOT masking of half the samples at a mask ratio drawn in $[0.1, 0.5]$. Each image gives two global crops at 256×256 over scale $[0.32, 1.0]$ and eight local crops at 112×112 over $[0.05, 0.32]$. Horizontal flips are disabled, because laterality is anatomy in ophthalmology and a mirrored macula is not another sample of the same distribution. AdamW runs at a peak learning rate of 5×10^{-5}, weight decay 0.04, layerwise decay 0.9, patch-embedding learning-rate multiplier 0.2, gradient clipping at 10, EMA momentum 0.994 and teacher temperature warming from 0.04 to 0.07, in bf16.

The 3D teacher. The volumetric teacher applies the same ViT-L/16 architecture as the 2D one, groupwise. A volume $B \times 1 \times D \times H \times W$ is reshaped to $BD/3 \times 3 \times H \times W$, so each group of three consecutive B-scans occupies the three input channels of one ordinary 2D forward pass, and the per-group patch tokens are flattened back along depth. It is therefore a chunked 2D encoder and not a native volumetric one. Its weights come from a separate DINOv3 pretraining run, so the two teachers share an architecture and not a checkpoint. That run initializes from the 2D teacher's weights and continues on a larger table of the same collection, with the same objective and schedule.

The 3D teacher wrapper. The wrapper gives the alignment loss one token axis to work along. If the 2D encoder returns N patch tokens per group of three B-scans, the features are reshaped as $B \times (D/3) \times N \times C_e$ and the group and patch axes flattened into $B \times ((D/3)N) \times C_e$.

D.2 Latent space and preprocessing

Latent space. All models operate in the latent space of a frozen Stable Diffusion VAE, the MSE-finetuned variant, which takes a 256×256 image to 32×32 over four channels under its standard 0.18215 latent scaling. Grouping happens in pixel space, before the encoder. Three neighboring B-scans are stacked as the three input channels of one VAE call, so a 60-B-scan volume is encoded as 20 three-channel inputs and becomes 20 latent slice groups, and a 240-B-scan volume becomes 80. The VAE's 2D convolutions therefore mix the three B-scans of a group directly.

Volume standardization. Volumes are resampled along the B-scan axis to a fixed depth, 60 for the thin protocol and 240 for the thick one, by linear interpolation between neighboring B-scans; the first and last acquired B-scans are preserved and interior ones are blended. Thin-slab volumes with fewer B-scans than the target are excluded rather than interpolated up, while thick-slab volumes are resampled in either direction. B-scans are resized in-plane to 256×256.

Intensity normalization. Intensities are min–max normalized to $[0, 1]$ against the data's own extremes, with no dataset-level statistics. The scope over which the extremes are taken follows the acquisition rather than a single global rule, with the thick-slab cohort normalized over the whole volume and the thin-slab cohort per B-scan. The two come from different manufacturers and devices, whose detectors differ in gain and in how much of the dynamic range a single frame occupies, so a scope that is right for one is not right for the other. Each cohort is internally consistent, which is what the model sees, and the per-shape branches let both scopes be handled at once.

Table 22: Model configurations. Parameter counts are in Table 24, measured from the evaluated checkpoints.

Model	Depth	Hidden	Heads	Patch
SiT-B/2	12	768	12	2
SynOphsis-B/2	12	768	12	2
SiT-XL/2	28	1152	16	2
SynOphsis-XL/2	28	1152	16	2

Table 23: Optimization settings, as passed in the training scripts. Both scales share every value in this table; batch composition and hardware are in the text above.

Setting	SynOphsis-B SynOphsis-XL
Learning rate	5×10^{-5}
Precision	fp16
Optimizer	AdamW, weight decay 0
Latent resolution / patch	256 / 2
Alignment loss weight	0.5
Alignment encoder depth	8

Paired 2D images. The data manifest fixes which 2D image belongs to a volume, so the same-eye and same-visit constraints and any ambiguity between candidates is settled there, not at training.

The paired image is not used as acquired. OCT DICOMs record where each B-scan falls on the en-face image, and the first and last B-scans therefore delimit the region the volume swept. We warp the 2D image so that this region fills the frame, which leaves the 2D input spatially registered to the volume, so that a pixel of the 2D token and the corresponding column of the corresponding B-scan refer to the same retinal location. A mask records which pixels fall inside the source image, since the warp can reach beyond it.

If no pair exists, or the location metadata is missing, the 2D input is zeroed and flagged absent and the volume is kept. This is how the volume-only conditional tasks run against the same checkpoints as the paired ones.

D.3 Model and optimization settings

Optimization and hardware. Both scales use AdamW at 5×10^{-5} on a cosine schedule after a $1{,}000$-step warmup, fp16, and no weight decay. SynOphsis-XL trains on a global batch of 32 volumes and 256 2D images with two gradient accumulation steps; SynOphsis-B on 16 volumes and 128 2D images with no accumulation. Alignment coefficients are $\lambda_{same} = \lambda_p = \lambda_{cross} = 0.5$ and the paired-2D denoising coefficient is 1.0. The alignment layer is `encoder_depth` $= 8$ at both scales. Training runs on 32 A100-40GB GPUs under a distributed launcher for up to $300{,}000$ steps, with the volume branch at 256×256 resolution and patch size two. We estimate the compute budget at up to $8{,}000$ GPU-hours; elapsed time on a shared cluster might be several times that.

Sampler. Every reported number and every figure uses the SDE sampler at 100 integration steps with a classifier-free guidance scale of 1.0, that is, unguided, for all models including the baselines.

D.4 Baseline configurations

2D LDM. A latent diffusion model (Rombach et al., 2022) in the original authors' codebase, with a frozen VQ-$f4$ first stage at 256×256 and a UNet denoiser of 128 base channels, multipliers $(1, 2, 4, 4)$, two residual blocks per resolution and spatial-transformer cross-attention at strides 2, 4 and 8. Conditioning is a jointly trained class-and-text embedder over the modality labels. It is trained on *the same* 2D corpus as our models, the same table of images and the same split, and differs from them in seeing no volumes at all.

3D LDM. Medical Diffusion (Khader et al., 2023), run as a two-stage pipeline, a 3D VQ-GAN over the raw volume, then a DDPM over its latent grid with a 3D UNet denoiser at 128 base channels,

Table 24: Parameters in the sampling path, counted from the evaluated checkpoints. Frozen autoencoders are listed separately and are not part of either total; alignment heads are excluded from ours because they are not used at sampling time.

Model	Sampled parameters	Frozen autoencoder
3D LDM (thin)	82,381,064	1,530,345
3D LDM (thick)	82,381,064	6,105,561
2D LDM	169,608,861	55,322,782
SiT-B	197,231,712	SD-VAE
SynOphsis-B	197,246,304	SD-VAE
SiT-XL	917,759,328	SD-VAE
SynOphsis-XL	917,781,216	SD-VAE

multipliers $(1, 2, 4, 4)$, eight attention heads, 200 diffusion steps and an L_1 objective. Two separate models are trained, one per protocol, because the two protocols differ in depth. The thin volumes enter at 60 B-scans and thick at 240, both of which the first stage reduces to a $15 \times 64 \times 64$ latent grid. The thin model trains on the same table of the internal volumes as ours, and both are evaluated against the same reference volumes as every other model in Table 2.

Plain SiT, the joint-token ablation. The plain SiT rows are our own model with the joint tokens removed. They use the same code, the same three data tables, and the same optimizer, learning rate, schedule, precision, and batch composition as the joint-token models at the matching scale. They differ in the entry point, since sequences are built per branch rather than jointly, and therefore in the loss terms that only exist when a volume and its paired 2D view share a sequence.

Model sizes. Table 24 counts the parameters that are actually sampled from, excluding the frozen autoencoder, the frozen text encoder, and the alignment heads through which $\mathcal{L}_{\text{REPA}}$ is computed but which are discarded at sampling time. The row that matters for Table 2 is the 2D LDM's, which at 169.6M is within fifteen percent of SynOphsis-B, so the comparison there is not between a small baseline and a large model. It spends that budget entirely on 2D images, where SynOphsis-B spends the same on eleven 2D modalities and three volume protocols at once.

The other row that matters is the last pair. Joint tokenization adds three embedding tables and nothing else, one for token type, one for the paired-view modality and one for whether that view is observed: 21,888 parameters at XL scale and 14,592 at B, against backbones of 918M and 197M. That is 0.002% of the XL backbone and 0.007% of the B one, so the two models in each pair are the same network to within a hundredth of a percent, so the comparison in Table 3 is one of sequence construction and not of capacity.

Differences between the two stacks. The baseline and our model differ in more than one place at once. The baseline follows Medical Diffusion (Khader et al., 2023), a discrete-time convolutional U-Net over a native 3D VQ-GAN latent, without representation alignment. Ours is a continuous-time interpolant transformer over slice-wise SD-VAE latents with $\mathcal{L}_{\text{REPA}}$. Transport, architecture, autoencoder, training objective *and scale* differ together. The baseline's denoiser is 82M parameters against the 197M of the SiT-B over which the 3.28 dB is measured, and 918M at XL (Table 24). The first contribution is that replacement, and the number measures what it buys under identical conditioning. We do not attribute it to any one component. We follow REPA (Yu et al., 2025) in aligning an intermediate representation to a pretrained encoder, which accelerates convergence, and we align to an ophthalmic encoder so that the acceleration carries the domain. We observe much faster convergence and no drop in final quality. REPA reports speed rather than a better optimum, and we claim no more than that for the term here.

The two internal conditioning variants. **SiT Ctrl** copies the backbone into a parallel control branch of the same depth and injects it through linear projections whose weights and biases are zero-initialized, so the branch contributes nothing at initialization. The observed-frame mask enters as a per-slice condition channel, following SparseCtrl (Guo et al., 2024). **SiT Inpaint** adds cross-attention blocks whose keys and values are a linear projection of the ophthalmic DINOv3 encoding of the paired 2D image, and performs sparse 3D inpainting by holding the observed latent slices clean throughout denoising. Neither changes the backbone, so both isolate how conditioning enters.

E EXTENDED RELATED WORK

2D ophthalmic generation. Most 2D ophthalmic generation targets data augmentation for a downstream segmentation or grading task, and its conditioning is accordingly an annotation of the image being generated. Our conditional tasks supply no annotation, so these methods cannot be run as baselines here.

Cascaded latent diffusion for OCT volumes. CA-LDM (Huang et al., 2024) synthesizes 512^3 OCT volumes within 24 GB using cascaded latent diffusion and a non-holistic autoencoder. Its released implementation takes the perceptual loss, the codebook and the VQ-GAN of Medical Diffusion (Khader et al., 2023), extending the last, so its volumetric autoencoder is Medical Diffusion's. No weights are released and the published model was trained on OCTA-500, so running it here would compare our models against a checkpoint fitted to a different corpus. Medical Diffusion trained on our data is the same formulation under the same corpus, preprocessing and budget, and Section D.4 gives its configuration.

Modality translation. Translation between ophthalmic modalities is a developed line of work, and most of the pairs an examination contains have been attempted. A color fundus photograph has been mapped to fluorescein angiography (Tavakkoli et al., 2020; Shi et al., 2023), to indocyanine green angiography (Chen et al., 2024), to autofluorescence (Song et al., 2023), to an ultra-widefield angiogram (Chen et al., 2026a) and to an angiographic video (Zhang et al., 2024), and OCT has been mapped to OCTA (Badhon et al., 2025). The angiographic sequence has also been generated from a report alone, without an image to start from, by a diffusion transformer over a video latent (Wu et al., 2024b). A world model has been proposed to hold several of these directions in one latent state (Gao et al., 2026). One of these methods takes two acquisitions at once: Ma et al. (2026) guide fundus-to-angiography synthesis with sparse OCT scans, projecting depth-resolved OCT features onto the fundus plane. The output there is a 2D angiogram, so the volume is evidence rather than a target. In each case the conditioning image must already exist, one output modality is produced, and the anatomy is fixed by the source. In our setting both views are sampled from noise, and none of the three holds.

Multi-output generation in general vision. Sampling more than one output together is established in general vision. A text-to-RGBD model generates an image and its depth map from one prompt, concatenating the two as extra channels of a single autoencoder input (Ben Melech Stan et al., 2023), and MVDream (Shi et al., 2024) draws several views of one object by letting them attend to each other inside one denoiser, which is the mechanism we use for a volume and its paired view. The outputs in both cases share a geometry. A depth map is registered to its image pixel for pixel, and the views of an object are renderings of one scene at known poses. An OCT volume and the en-face photograph taken beside it are two acquisitions of the same eye on different grids, so the pairing has to be learned from data rather than imposed by construction. Within ophthalmology the nearest case is several fundus fields of view of one eye synthesized together from a single image (Demirci et al., 2025), where the outputs are one modality and the input is an acquisition that already exists.

Mixing 2D and 3D tokens in one attention sequence is established in general vision as well. FLUX3D concatenates image tokens and sparse volumetric latents into a single stream under joint attention, with a modality-aware positional encoding that places the image on a virtual plane outside the volume (Ji et al., 2026). There the 2D side is the condition and the 3D asset is the only output, so the sequence carries evidence into one generated object. Ours carries two generated acquisitions of the same eye.

Generation from sparse measurements. Recovering a volume from a few observed slices is an inverse problem, and outside ophthalmology those are solved with a diffusion prior over the image and the measurement model applied at sampling time. Song et al. (2022) and Chung & Ye (2022) reconstruct sparse-view CT and undersampled MRI that way. Our seven conditional tasks have that shape. The measurement operator there is physical, a projection or a Fourier subsampling, where ours selects latent slices, and the target there is one modality, where a sample here carries a volume and the 2D view acquired with it in one sequence. Those methods are built for their own operators, so we take the framing from them and not the machinery.

Table 25: Modality composition of the 2D training set, 4,752,726 images in total. One row of the training CSV is one 2D image.

Modality	Images	Share	Where it appears
OCT	2,701,450	56.84 %	Table 2, OCT half
IR	996,328	20.96 %	main text
CFP	406,639	8.56 %	main text
OCTA	237,402	5.00 %	main text
FFA	120,937	2.55 %	main text
FAF	80,895	1.70 %	main text
External eye	60,191	1.27 %	Table 8
ICGA	53,072	1.12 %	main text
OCTA en-face	41,984	0.88 %	main text
AS-OCT	40,055	0.84 %	Table 8
MRI	11,356	0.24 %	Table 8
Slit lamp	1,874	0.04 %	not evaluated
Ultrasound	489	0.01 %	not evaluated
Other[†]	54	< 0.01 %	not evaluated
OCT	2,701,450	56.84 %	
All non-OCT	2,051,276	43.16 %	

[†]One residual modality label in the source metadata, 54 images. It is carried as its own class by the dataset code but is not expanded here, since at under 0.01 % of the corpus it supports nothing the paper claims.

Volume generation conditioned on a 2D view. de Vente et al. (de Vente et al., 2024) super-resolve an acquired volume from its en-face image; the volume is an input there and an output here. The comparison for our 2D-only task is therefore internal. Our cross-attention variant holds the backbone, the data and the budget fixed and differs only in how the 2D view enters the sequence, which isolates the formulation and leaves the task's difficulty uncalibrated against other work.

A parallel line generates the volume from a color fundus photograph. Fundus2OCT produces volumes of 32 B-scans with a two-stage latent diffusion model trained on paired UK Biobank data (Liu et al., 2025a), and the APTOS-2024 challenge benchmarked that task across 342 teams with nine finalists (Liu et al., 2025b). The conditioning image there comes from a separate instrument, and the volume is the only output, so those numbers calibrate fundus-to-volume generation. The paired view in our setting is recorded with the volume and is generated alongside it.

F TRAINING DATA

Table 25 gives the modality composition of the 2D training set. OCT is the majority modality at 56.8%, which is the context for the non-OCT results discussed in Section 5.1, where the remaining modalities are individually much sparser, and the model is not trained to a balanced mixture.

The 2D collection. The training set, the non-OCT evaluation and the main table cover different numbers of modalities, and the last column of Table 25 gives each modality's role.

Training uses all fourteen: color fundus photography (CFP), fundus autofluorescence (FAF), fluorescein angiography (FFA), indocyanine green angiography (ICGA), infrared reflectance and scanning laser ophthalmoscopy (IR/SLO), OCT angiography as volumes (OCTA) and as en-face projections (OCTA en-face), anterior-segment OCT (AS-OCT, including corneal and iris sections), external eye photography, orbital magnetic resonance imaging (MRI), slit-lamp photography, ultrasound, one residual label, and OCT itself at 56.8% of the images. Nothing is rebalanced.

The non-OCT evaluation covers ten of them, dropping slit lamp, ultrasound and the residual label, which together come to under 0.05% of the corpus and are too few to estimate a distributional metric from. Those ten together with OCT are the eleven 2D modalities the paper reports. Each of the ten is scored on 1,000 generated samples, except AS-OCT at 746, and every model is scored on the same set. OCT has benchmarks of its own, seven columns in the OCT half of Table 2. Four are external collections no model trains on, two are the internal splits, and the seventh is AI-READI, whose thick volumes the models do train on and whose test split is partitioned by participant.

Table 26: The volumetric training and evaluation data. Thin is the internal cohort cohort, thick is AI-READI. Counts are acquisitions; a thick acquisition carries both the structural volume and, where present, the OCTA flow cube and its en-face projections.

		Train	Test
Thin (internal)	Volumes	236,188	59,649
	Patients	29,733	7,434
	With a paired 2D image	236,041	59,613
	Device	Heidelberg Spectralis, 61 B-scans	
Thick (AI-READI)	Acquisitions	20,777	3,777
	Participants	1,917	347
	Topcon Triton	7,548	1,331
	Zeiss Cirrus	7,454	1,376
	Topcon Maestro2	3,792	713
	Heidelberg Spectralis	1,983	357
	Macula 6×6 / 12×12 / 20×20 / disc 6×6	11,263 / 3,798 / 1,983 / 3,733 (train)	

The main table prints seven of the ten. AS-OCT, external eye photography and an orbital MRI are the three furthest from the retinal imaging this paper is about, so the non-OCT half of Table 2 keeps to modalities a reader compares like for like. All ten are reported here: Table 8 carries the other three in the same layout and from the same evaluation.

Volume census. Table 26 counts the volumes these models were actually trained and evaluated on, which is a subset of each cohort and not the cohort's published total.

The thin protocol is one device and one geometry throughout, and every volume is a Heidelberg Spectralis acquisition of 61 B-scans, left and right eyes in even proportion, and 99.9% of them carry the paired infrared photograph the joint tokens need. Fields of view cluster tightly around 7.5×8.9 mm, varying by a few tenths because the Spectralis records the scanned extent per acquisition instead of snapping to a nominal size. The thick protocol is the opposite, with four devices from three manufacturers, and four scan geometries from 6×6 mm at the macula out to 20×20 mm, which is why the metadata conditioning exists at all.

Both splits are partitioned by patient. No patient in the thin training set appears in its test set, and no participant in the thick training set appears in its test set, both checked against the tables themselves. A random volume-level split would have leaked across the fellow eye and across repeat visits, and for the thin cohort, where one patient contributes eight volumes on average, the leak would have been severe.

The non-OCT evaluation splits. Each non-OCT modality is scored on a held-out split of the same aggregated corpus the models are trained on, sampled to n=1000. OCTA and OCTA en-face are drawn from AI-READI (AI-READI Consortium, 2024); CFP from the FLAIR assembly (Silva-Rodríguez et al., 2025), EDDFS (Xia et al., 2022) and other public fundus collections; FFA and ICGA chiefly from the third APTOS angiographic competition dataset (Zhang et al., 2023b; Xu et al., 2025), with FFA also drawing on FFA-IR (Li et al., 2021) and MM-Retinal (Wu et al., 2024a); and FAF and IR/SLO predominantly from a clinical trial cohort, with the remainder from AI-READI retinal photography. The public collections for these modalities are aggregated into the corpus the models train on, and each non-OCT benchmark is a held-out split of that aggregate. No image in an evaluation split appears in the diffusion training or in the DINOv3 teacher's pretraining. The setting is inductive throughout. OCT is the exception, with four collections in Table 2 that no model trains on, and two the internal splits.

Splits. Training, validation and evaluation splits are partitioned at the patient level, so no eye or visit from an evaluation split appears in training. The ophthalmic DINOv3 teacher's pretraining corpus is likewise disjoint from every evaluation split used here.

Preprocessing. Volumes are read from DICOM and resampled along depth to a fixed number of B-scans, 60 for the thin protocol and 240 for the thick, by linear interpolation between neighboring source B-scans at evenly spaced positions, so no slice is duplicated or dropped outright. Each B-scan

Table 27: SSIM for the conditional tasks, volumetric above and paired 2D below, for the same models and tasks as Tables 3 and 4. Higher is better. SSIM orders the methods as PSNR does.

		Thin OCT							*Thick OCT*						
Model	Metric	Mid	Span	First	Last	2D	2D+Mid	2D+10%	Mid	Span	First	Last	2D	2D+Mid	2D+10%
Existing methods															
3D LDM Ctrl	PSNR	15.49	16.84	14.46	14.87	/	/	/	19.46	21.65	19.28	18.81	/	/	/
	SSIM	0.251	0.279	0.212	0.227	/	/	/	0.513	0.559	0.499	0.492	/	/	/
3D LDM Cross	PSNR	/	/	/	/	12.64	/	/	/	/	/	/	16.74	/	/
	SSIM	/	/	/	/	0.175	/	/	/	/	/	/	0.453	/	/
Ours: 3D SiT, alternative conditioning															
SiT-B Ctrl	PSNR	19.14	20.98	17.60	17.86	/	/	/	22.89	25.41	21.55	21.64	/	/	/
	SSIM	0.468	0.514	0.423	0.433	/	/	/	0.643	0.676	0.628	0.630	/	/	/
SiT-B Inpaint	PSNR	19.18	21.24	17.81	17.86	13.87	19.20	23.88	22.71	25.24	21.68	21.62	19.67	22.71	24.59
	SSIM	0.470	0.525	**0.435**	0.440	0.302	0.474	0.575	0.642	0.677	0.630	0.630	0.604	0.642	0.695
Ours: joint-token															
SynOphsis-B	PSNR	19.24	21.48	17.70	18.04	14.15	19.45	23.83	22.57	25.33	21.93	21.71	**19.94**	22.71	23.54
	SSIM	0.473	0.528	0.432	0.442	**0.311**	0.480	0.574	0.640	0.676	0.632	0.631	**0.607**	0.641	0.681
SynOphsis-XL	PSNR	**19.59**	**21.74**	**17.83**	**18.28**	**14.17**	**19.78**	**24.29**	**23.25**	**26.36**	**22.24**	**21.96**	**19.94**	**23.46**	**26.83**
	SSIM	**0.481**	**0.535**	0.434	**0.446**	**0.311**	**0.487**	**0.585**	**0.647**	**0.685**	**0.636**	**0.632**	**0.607**	**0.648**	**0.718**

		Thin OCT paired 2D				*Thick OCT paired 2D*			
Model	Metric	Mid	Span	First	Last	Mid	Span	First	Last
Existing methods									
2D LDM	PSNR	13.28	13.50	12.54	13.52	14.92	15.41	15.50	15.68
	SSIM	0.315	0.322	0.312	0.320	0.602	0.605	0.599	0.619
Ours: joint-token									
SynOphsis-B	PSNR	14.31	14.64	14.03	13.72	18.63	18.42	17.81	18.18
	SSIM	0.355	0.362	0.356	0.351	0.679	0.670	0.670	0.671
SynOphsis-XL	PSNR	**14.42**	**15.12**	**14.46**	**13.85**	**19.43**	**19.88**	**18.92**	**19.15**
	SSIM	**0.362**	**0.378**	**0.367**	**0.359**	**0.691**	**0.683**	**0.679**	**0.682**

is then resized to 256×256 bicubically and center-cropped. Intensity is min–max rescaled at the scope set out above, per B-scan for the internal volumes and over the whole volume for AI-READI.

The paired 2D image is brought into the volume's own frame. For the internal volumes the DICOM records, for the first and last B-scan, the coordinates of that scan line in the fundus image's own coordinate system; those four corners define a perspective transform, and the 2D image is warped through it so that its extent is the volume's scan footprint and a validity mask marks anything that falls outside the source image. AI-READI does not carry those coordinates, so the same effect is approximated from geometry, dividing the field of view in millimetres by the DICOM pixel spacing to give a box in pixels, centerd on the image and warped to the same 256×256 frame. If a volume has no paired 2D image at all, the 2D slot is zero and its validity flag is off, and the loss terms that involve it are skipped for that sample.

Data sources, provenance and use. The thin OCT volumes and the 2D collection come from an **internal ophthalmology cohort**, a retrospective collection acquired at one academic eye center, which we name on acceptance, with the thin protocol acquired on Heidelberg Spectralis instruments at approximately 60 B-scans of 256×256 per volume. The thick volumes come from AI-READI (AI-READI Consortium, 2024; Owsley et al., 2025).

Institutional review board approval was obtained, with a waiver of informed consent for the retrospective analysis of de-identified imaging; AI-READI is used under its data use agreement. All procedures adhered to the tenets of the Declaration of Helsinki. No image in this work is presented as belonging to an identifiable individual, and the clinical text and acquisition metadata used as conditioning are drawn from the de-identified record.

G LIMITATIONS

Unscored results. Text conditioning is an input to every number in Table 3, since each model is run with the clinical description recorded for the case, but generation from a description alone ap-

pears only in Figure 2 and carries no number. Quantifying it would need a text-conditioned baseline on these modalities and volumes, and the nearest published work (Ning et al., 2025) measures fundus photographs. It is a capability we demonstrate and not a result we measure. The 2D-only task is a demonstration rather than a measured result. Given an en-face image and no B-scans, nothing anchors the output to the eye that was scanned, and PSNR is low for every method including ours (Appendix C.5). No matched specialist exists for the paired 2D output, since neither conditioning variant emits one; the volumetric half of that task is therefore compared against a matched variant and the paired half against the 2D LDM, which is the closest method that emits a 2D image from 3D evidence. The comparison against matched task-specific variants runs at B scale, and it measures whether the broader interface costs volumetric reconstruction accuracy rather than whether joint tokenization improves any one task. The generalist holds parity there, 0.03 dB over fourteen cells, inside the per-case variability reported below. Scale adds 0.58 dB from SynOphsis-B to SynOphsis-XL, and the 3.28 dB is a stack-level comparison rather than a gain attributable to joint tokenization.

Metrics and scope. Half the columns in every table are computed in the ophthalmic encoder our alignment loss is taken against, and the external baselines were never trained towards it. We report both feature spaces for every row, and Appendix C.1 repeats the unconditional evaluation in four ophthalmic encoders no model was aligned to. Reconstruction margins are small against the volume-to-volume spread, so they are read from paired tests on the same volumes and not from the means (Table 18). On the 2D-only task at both slabs the margin does not reach significance, which is what a task with no observed B-scan should give. Two evaluations this work does not carry are the ones that would show the model used rather than measured. A reader study would ask whether generated images are taken for acquired ones and whether the pathology on them is credible, and a downstream experiment would train a segmentation or grading model on generated data and report what it gains. We leave both to future work. All models are trained and evaluated on the retinal imaging described in Section 4. We make no claim about other ocular structures, about acquisition devices absent from the training distribution, or about any clinical use. Nothing here has been assessed by readers for diagnostic adequacy, and synthetic ophthalmic images should not be treated as substitutes for acquired ones.