Anatomy-Preserving Unpaired Medical Image Translation via Shared Latent Anchoring

Zhilin Zou[1] **Hieu Le**[2] **Saumya Gupta**[1]
Jingyi Xu[1] **Prateek Prasanna**[1] **Chao Chen**[1]
[1]Stony Brook University, NY, USA
[2]University of North Carolina at Charlotte, NC, USA
zhizou@cs.stonybrook.edu chao.chen.1@stonybrook.edu

Abstract

Unpaired medical image-to-image translation requires synthesizing target-modality images from source-modality images without aligned supervision. Existing methods learn a direct source-to-target mapping that entangles structure preservation and appearance generation within a single model, often distorting clinically meaningful anatomy in the translated output. We argue that these two objectives should be decoupled entirely: anatomical structure should be captured independently of generation, and the generator should operate only on a frozen structural representation. To this end, we propose a two-stage framework for unpaired medical image translation. In the first stage, we learn a shared structure space by training a structure encoder-decoder on source images, target images, and structural masks jointly, using mask reconstruction to anchor the latent space to modality-agnostic anatomy. For source images without mask annotations, a target-domain memory bank provides soft structural supervision via prototype retrieval. In the second stage, a conditional flow model learns to render target-domain appearance under fixed structural conditions. This ensures the generator inherits target appearance statistics and can focus entirely on structural conditioning rather than learning appearance from scratch. We validate our framework on OCT-to-OCTA synthesis and brain MR-to-CT translation. Our method outperforms all unpaired baselines on both tasks and, despite requiring no paired supervision, also surpasses paired baselines, improving PSNR by up to 11.2% and SSIM by up to 26.4% on OCT-to-OCTA, and achieving consistent gains across whole-image, soft-tissue, and bone regions on MR-to-CT. The project repository is available on GitHub.

1 Introduction

Multi-modal medical imaging is important in clinical diagnosis because different imaging modalities provide complementary anatomical, functional, and physiological information for disease assessment [1]. However, acquiring multi-modal images remains challenging due to factors such as cost, acquisition time, scanner availability, patient motion, and protocol inconsistency [1]. For example, optical coherence tomography (OCT) captures high-resolution retinal structure but not blood flow, whereas OCT angiography (OCTA) provides vascular information but requires additional hardware and software support [2], which limits its availability. Such trade-offs between modality utility and accessibility are common across clinical imaging settings. Medical image-to-image translation, therefore, offers a practical solution by synthesizing missing target modalities from available source modalities while preserving source anatomical structures. A key challenge, however, is that most existing translation methods require *paired* source-target data during training [3–5], where images from different modalities are spatially aligned and correspond to the same subject, anatomy, and acquisition setting. Such paired multi-modal data are difficult to obtain in practice, especially when

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

modalities differ in acquisition protocols, availability, scan timing, field of view, or patient compliance. This motivates *unpaired* translation, where source- and target-domain images are available as separate collections without one-to-one correspondence.

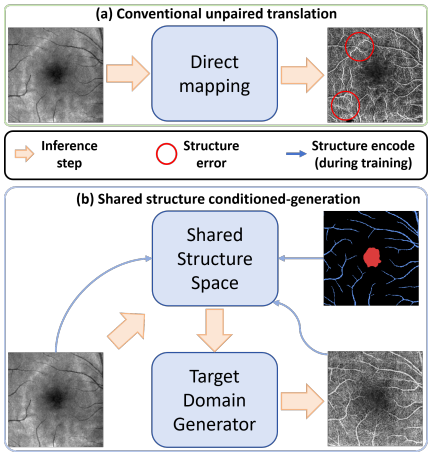


Figure 1: Conventional unpaired translation vs. our shared structure-conditioned generation.

Medical image translation is not simply an appearance-transfer problem. Although modalities differ in contrast, texture, intensity distribution, and imaging response, the translated image should preserve the anatomy of the input source image. These modality-shared anatomical structures provide a natural basis for cross-domain translation: OCT and OCTA are tied by vessel organization and the foveal avascular zone (FAZ), while magnetic resonance imaging (MR) and computed tomography (CT) share brain tissue, ventricles, and skull-related structures. A translated image is therefore clinically meaningful only if it changes modality-specific appearance while preserving the underlying anatomy. Existing unpaired translation methods usually learn a direct source-to-target mapping through adversarial matching, cycle consistency, contrastive patch learning, or diffusion-based conditional generation [1, 6, 7]. These objectives encourage target-domain realism, but they do not explicitly specify which anatomical information should remain invariant. As a result, structure preservation and appearance synthesis are entangled, and without source-target correspondence, the model may produce target-like images with missing vessels, distorted FAZ boundaries, or hallucinated tissue patterns, as illustrated in Fig. 1. This motivates an explicit formulation that defines the cross-domain structure to be preserved before target-domain appearance is generated.

To this end, we ground the translation process in the underlying anatomical structure. Instead of learning a direct source-to-target mapping, we formulate unpaired medical image translation as structure-conditioned target-domain generation. We first construct a cross-domain shared structure latent space that specifies the anatomy to be preserved across modalities. A structure encoder-decoder maps source images, target images, and available structural masks into this shared space and reconstructs structural maps, encouraging the latent representation to capture modality-agnostic anatomy rather than modality-specific appearance. We then decouple structure extraction from target-domain appearance synthesis by using the learned structure space as fixed anatomical guidance for generation.

Learning such a shared structure space under unpaired supervision is challenging. On the one hand, accurately registered cross-domain image pairs are difficult to acquire. On the other hand, obtaining reliable structural annotations for both modalities is also unrealistic in many medical settings, because different modalities reveal different structures with different reliability. For example, bone-related structures are easier to define in CT than in MR, while vascular structures are more directly visible in OCTA than in OCT. Therefore, we avoid assuming accurate source-domain structural annotations. Instead, we propose a Structural Memory Bank (SMB), which stores target-domain structural masks and their latent representations as prototypes. For each source image, SMB retrieves structurally similar target-domain prototypes and aggregates them into a soft structural target, enabling supervision of source-derived structure latents using only target-domain structural annotations. We then freeze the learned structure encoder-decoder and use its structural representations to condition a target-modality flow model [8]. Initialized from an unconditional target-domain prior, the flow model focuses on rendering target appearance under fixed anatomical guidance, thereby separating structure preservation from appearance generation.

We evaluate the proposed framework on two representative medical image translation tasks: OCT to OCTA and Brain MR to CT translations. These tasks involve different anatomies, imaging mechanisms, and appearance transformations while sharing the same goal: preserving the source-domain structure while generating target-domain appearance. Experiments show that explicitly modeling the shared structure to be preserved improves both image fidelity and structural consistency compared with competing baselines, supporting the effectiveness of structure-conditioned generation for unpaired medical image translation.

In summary, our contributions are as follows:

1. **Structure-conditioned formulation.** We formulate unpaired medical image translation as structure-conditioned target-domain generation. This explicitly defines modality-shared anatomical structure as the information to be preserved and target-domain appearance as the information to be generated.
2. **Shared structure latent modeling with SMB.** We construct a cross-domain shared structure latent space using a structure encoder-decoder, and propose a Structural Memory Bank (SMB). SMB stores target-domain structural prototypes and provides soft supervision for source-derived structure latents, supporting shared structure learning through target-domain structural supervision.
3. **Comprehensive experimental validation.** We evaluate our framework on OCT-to-OCTA and brain MR-to-CT translation, demonstrating improved image fidelity and structural consistency over competing paired and unpaired baselines. Held-out-center testing, independent vessel evaluation, downstream classification, and a blinded reader study further assess generalization and preservation of clinically relevant information.

2 Related work

Image-to-image translation. Image-to-image (I2I) translation has been widely studied for cross-modality medical image synthesis, aiming to infer a target modality from an acquired source modality. Supervised methods require paired data, such as Pix2pix [3], diffusion-bridge models [4], MuTri for paired 3D OCT-to-OCTA synthesis [2], and MISA-LDM for missing-modality MRI synthesis [5]. Despite their effectiveness, these methods depend on accurately registered multimodal data, which are difficult to collect at scale. Unpaired I2I translation reduces this requirement by learning cross-domain mappings from unaligned image collections [6, 7, 9–14]. Representative methods include CycleGAN [6], which uses cycle consistency, CUT [7], which adopts patch-wise contrastive learning, and MUNIT [9], which disentangles content and style. Recent approaches further explore probabilistic formulations, such as UNSB [15] with Neural Schrödinger Bridges, and SynDiff [1] with cycle-consistent adversarial diffusion for medical image translation. Although these methods reduce the need for paired data, they still largely learn translation as a direct domain mapping, where structure preservation and appearance generation are entangled. In the related setting of unpaired image restoration, DDSB [16] incorporates degradation-aware optimal transport and dynamic transport consistency into a Schrödinger bridge. Our method instead uses a frozen shared structure representation to decouple anatomical structure modeling from target-domain appearance rendering.

Conditional generative models for medical image synthesis. We focus on conditional generative models for medical image synthesis. Beyond unconditional diffusion models such as DDPMs [17] and transport-based models such as flow matching [8], recent methods improve controllability using anatomical, pathological, or physical conditions. LeFusion [18] and spine MRI inpainting [19] use lesion or pathology masks for structure-aware synthesis and editing, while adaptive diffusion priors [20], Dose-Aware Diffusion [21], and Poisson-flow generative models [22] incorporate acquisition or physics constraints for MRI, PET, and CT reconstruction. Topology-aware methods further introduce persistent-homology or multi-scale topological constraints for synthetic masks and cellular layouts [23, 24]. More recently, SynFlow [25] uses mask-conditioned flow matching to generate topology-preserving crack and vessel images from binary masks. Overall, these studies demonstrate the value of ROI/mask conditioning, physics guidance, topology-aware constraints, and flow-based generation, but generally require image-condition correspondence during training, which can be difficult to obtain in medical settings.

3 Method

We consider unpaired medical image translation from a source domain $\mathcal{X}_A$ to a target domain $\mathcal{X}_B$, where no image-level correspondence is available across domains. As both domains depict the same anatomical organ or region, they contain comparable anatomical structures, although their appearance (such as intensity and noise) can differ substantially due to modality- or protocol-specific characteristics. We assume access to target-domain image–mask pairs (x_B, m_B), where m_B denotes the structural mask of image x_B. In the source domain, only images $x_A \sim \mathcal{X}_A$ are required while

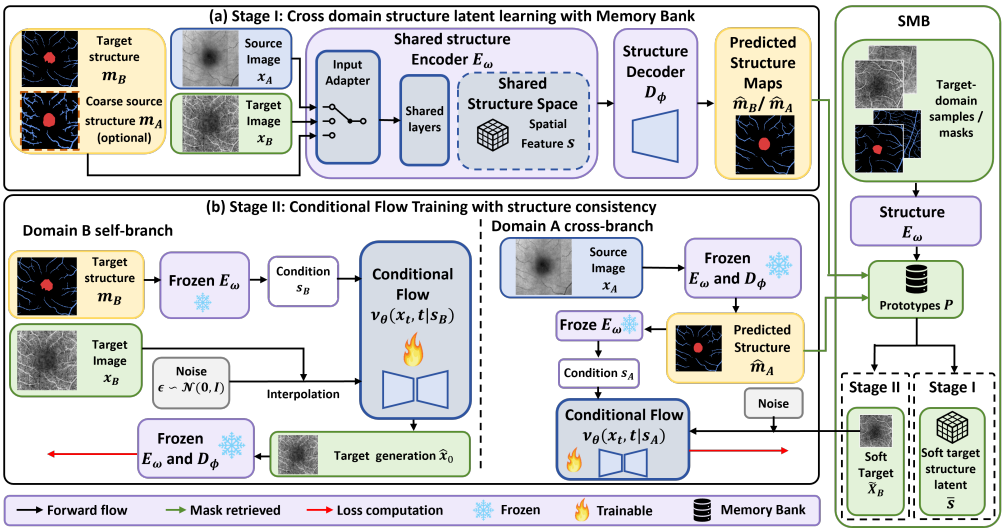


Figure 2: Overview of the training procedure of our proposed two-stage framework. Stage I learns a shared structure encoder-decoder that maps images and structural masks into a common structure space, with memory-bank-based soft supervision for source images. After freezing the encoder-decoder, Stage II trains a conditional flow model for target-domain generation. The self-branch learns target-domain appearance rendering, while the cross-branch uses source-derived structures and memory-bank prototypes for cross-domain synthesis.

coarse source masks m_A (obtained using threshold-based methods) may be used when available. Given a source image x_A, the goal is to generate the corresponding target-domain image $\hat{x}_B$ that preserves the anatomical structure of x_A while matching the appearance distribution of $\mathcal{X}_B$.

We formulate unpaired translation as latent structure-conditioned target-domain generation, and we tackle each of them in separate stages. We first briefly review the flow matching [8] formulation used for target-domain generation. In Stage I, described in Sec. 3.1, we learn a shared structure encoder-decoder (E_ω, D_ϕ). The encoder maps source images, target images, target-domain masks, and optional source-domain masks into a common structure space $\mathcal{Z}_{\text{str}}$, while the decoder reconstructs structural maps, providing an explicit structure prior shared across domains. In Stage II, described in Sec. 3.2, we freeze E_ω and D_ϕ and train a conditional flow model to generate a target-domain image x_B from the learned structural conditions $s_u \in \mathcal{Z}_{\text{str}}$. Sec. 3.3 summarizes the inference procedure. We provide an overview in Fig. 2.

Background: Flow Matching. Flow matching [8] learns a velocity field that maps Gaussian noise to the data distribution. For $x \sim \mathcal{X}$, $\epsilon \sim \mathcal{N}(0, I)$, and $t \in [0, 1]$, we define $x_t = (1-t)\epsilon + tx$, giving target velocity $v_t = x - \epsilon$. The unconditional and conditional objectives are

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{\epsilon,t}\left[\|v_\theta(x_t, t) - v_t\|_2^2\right], \quad \mathcal{L}_{\text{cond}} = \mathbb{E}_{\epsilon,t}\left[\|v_\theta(x_t, t \mid s) - v_t\|_2^2\right]. \tag{1}$$

At inference, samples are generated by solving $\frac{dx_t}{dt} = v_\theta(x_t, t)$ from $t = 0$ to $t = 1$.

3.1 Stage I: Shared Structure Space Learning

Stage I learns an encoder-decoder to construct a modality-agnostic latent structure space. Given an input u, the structure encoder E_ω produces a latent spatial feature $s_u = E_\omega(u)$ where $s_u \in \mathcal{Z}_{\text{str}}$, and u can be either source image x_A, target image x_B, or target mask m_B. The structure decoder D_ϕ reconstructs a multi-class structural map, $\hat{m}_u = D_\phi(s_u)$. Since images and masks differ in their low-level form, we use input-specific adapters before the shared encoder layers. These adapters absorb low-level input differences, while the shared layers enforce a common latent structure representation. Stage I uses different supervision depending on the input type. Target mask inputs m_B anchor the latent space through reconstruction, target images x_B are supervised by their structural masks m_B, and source images x_A are guided by memory-bank-retrieved target-domain prototypes when reliable source-domain masks are unavailable. When a coarse source mask m_A is available, it provides additional weak decoding supervision. These complementary objectives encourage all inputs to be encoded into a shared structure space $\mathcal{Z}_{\text{str}}$, as detailed below.

Mask decoding. To anchor $\mathcal{Z}_{\text{str}}$ to anatomical structure, we supervise the decoder output $\hat{m}$ with an available structural mask m, using a mask decoding loss:

$$\mathcal{L}_{\text{dec}}(\hat{m}, m) = \text{CE}(\hat{m}, m) + \frac{1}{C-1}\sum_{c=1}^{C-1}\left(1 - \text{Dice}(\hat{m}^c, m^c)\right), \tag{2}$$

where $\text{CE}(\cdot,\cdot)$ denotes the cross-entropy loss, $\text{Dice}(\cdot,\cdot)$ denotes the class-wise Dice coefficient, C is the number of structure classes, and the background class is excluded from the Dice term. This loss is applied to different inputs depending on the available supervision: m_B defines the mask-based structure space by self-reconstruction, x_B is supervised by its corresponding annotation m_B, and x_A is supervised by the coarse mask m_A only when it is available.

Image-mask latent alignment. To encourage the target-domain image and corresponding mask to share a canonical structure space, we align their latent features via a latent consistency loss:

$$\mathcal{L}_{\text{align}} = \|s_B - s_m\|_1, \quad s_B = E_\omega(x_B), \quad s_m = E_\omega(m_B), \tag{3}$$

where s_B and s_m denote the spatial features of the image-induced and mask-induced latent representations, respectively.

Soft prototype structure latent alignment. It is essential for (E_ω, D_ϕ) to encode structural information from source images x_A, as this is the only input available during inference time. However, source-domain images often lack reliable mask annotations. To provide structural supervision in this setting, we propose a Structural Memory Bank (SMB), which stores target masks, target images, and their corresponding image latent representations as a prototype set $\mathcal{P}$. For each source image, SMB retrieves structurally similar target prototypes and aggregates them into a soft structural latent $\bar{s}$, providing pseudo structural supervision without requiring source-domain masks.

$$\mathcal{P} = \{(x_B^j, m_B^j, s_B^j)\}_{j=1}^N, \quad s_B^j = E_\omega(x_B^j), \tag{4}$$

where N is the number of stored target domain samples. Given a source image x_A, we predict its structure map $\hat{m}_A = D_\phi(E_\omega(x_A))$, then compare $\hat{m}_A$ with target masks $\{m_B^j\}_{j=1}^N$ in the bank and retrieve the top-K most similar prototypes (similarity measured via Dice). For each retrieved prototype $\{s_B^j\}_{j=1}^K$, we compute a similarity-based weight α_j, and the retrieved prototypes are then aggregated into the soft prototype structural latent $\bar{s}$:

$$\bar{s} = \sum_{j=1}^K \alpha_j s_B^j, \quad \alpha_j = \frac{\exp\left(\text{Dice}(\hat{m}_A, m_B^j)^\gamma/\tau\right)}{\sum_{\ell\in K}\exp\left(\text{Dice}(\hat{m}_A, m_B^\ell)^\gamma/\tau\right)}, \tag{5}$$

where γ controls the sharpness of similarity scores, and τ is a temperature parameter. The source-derived latent is then aligned with this soft latent using an L_1 loss:

$$\mathcal{L}_{\text{bank}} = \|E_\omega(x_A) - \bar{s}\|_1. \tag{6}$$

This bank-based supervision encourages source structures to lie close to the target structural distribution without requiring paired image supervision.

Cycle consistency. To further suppress modality-specific appearance in the structure latent space $\mathcal{Z}_{\text{str}}$, we introduce an asymmetric latent-cycle consistency constraint. Specifically, we decode the input-derived latent into a structural mask and then re-encode this mask back into the latent space. The original input latent is encouraged to match the re-encoded structural latent.

$$\mathcal{L}_{\text{cycle}} = \|s_u - \text{sg}(\tilde{s}_u)\|_1, \quad s_u = E_\omega(u), \quad \tilde{s}_u = E_\omega(D_\phi(s_u)). \tag{7}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation, and $u \in \{x_A, x_B\}$. Therefore, the cycle loss only updates the first encoded latent s_u, while the re-encoded latent $\tilde{s}_u$ is treated as fixed. This asymmetric image-to-mask-to-latent cycle encourages s_u to retain information that can be consistently expressed through a structural map, while suppressing modality-specific appearance cues.

Overall Objective in Stage I. To summarize, the Stage I objective depends on the type of input. For a target structural mask m_B, we use only the reconstruction loss:

$$\mathcal{L}_{m_B} = \lambda_{\text{dec}}\mathcal{L}_{\text{dec}}\big(D_\phi(E_\omega(m_B)), m_B\big). \tag{8}$$

For a target image x_B with its structural annotation m_B, we use reconstruction, image-mask latent alignment, and latent cycle consistency:

$$\mathcal{L}_{x_B} = \lambda_{\text{dec}}\mathcal{L}_{\text{dec}}\big(D_\phi(E_\omega(x_B)), m_B\big) + \lambda_{\text{align}}\mathcal{L}_{\text{align}}(x_B, m_B) + \lambda_{\text{cycle}}\mathcal{L}_{\text{cycle}}(x_B). \tag{9}$$

For a source image x_A, accurate annotations are unavailable. We therefore use the memory-bank loss to pull its structure latent toward the target structural distribution, together with latent cycle consistency. If an optional coarse source mask m_A is available, we also include a decode loss:

$$\mathcal{L}_{x_A} = \lambda^A_{\text{dec}}\mathcal{L}_{\text{dec}}\big(D_\phi(E_\omega(x_A)), m_A\big) + \lambda_{\text{bank}}\mathcal{L}_{\text{bank}}(x_A) + \lambda_{\text{cycle}}\mathcal{L}_{\text{cycle}}(x_A), \tag{10}$$

where $\lambda^A_{\text{dec}} = \lambda_{\text{dec}}$ when the optional coarse source mask m_A is available, and $\lambda^A_{\text{dec}} = 0$ otherwise. The overall Stage-I objective is optimized by applying the corresponding loss to each type of input:

$$\mathcal{L}_{\text{stageI}} = \mathcal{L}_{m_B} + \mathcal{L}_{x_B} + \mathcal{L}_{x_A}. \tag{11}$$

The reconstruction losses anchor the structure space using available masks, the alignment loss pulls annotated target image representations into this space, the bank loss aligns source-derived structures with the target structural distribution, and the cycle loss suppresses modality-specific appearance information through image-to-mask-to-latent consistency.

3.2 Stage II: Structure-Conditioned Target Generation

In Stage II, we freeze the encoder-decoder (E_ω, D_ϕ) learned in Stage I, and train a target-domain flow generator conditioned on structural representations from $\mathcal{Z}_{\text{str}}$. This decouples the two objectives: Stage I extracts structure, while Stage II learns to generate target-domain appearance. Given a latent structural condition $s \in \mathcal{Z}_{\text{str}}$, the conditional velocity field is defined as $v_\theta(x_t, t \mid s)$, where s provides structural guidance. The generator is optimized to render target-domain appearance while respecting the structural condition from the frozen encoder. As in Stage I, Stage II's training objective varies with the input type; we describe these as self-branch and cross-branch below.

Target-domain self-branch. This branch learns to render target appearance under reliable structural supervision. Given a target image x_B and its mask m_B, we obtain the flow model condition as $s_B = E_\omega(m_B)$. The conditional flow model is then trained to reconstruct x_B under this condition using the conditional flow matching objective (from Eq. 1):

$$\mathcal{L}_{\text{self}} = \mathcal{L}_{\text{cond}}(x_B, s_B). \tag{12}$$

This branch provides in-domain supervision for the generator, teaching it to render target-domain appearance when the corresponding structure is provided as condition.

Source-to-target cross-branch. This branch teaches the flow model to render target appearance when the condition is derived from the source image. Given a source image x_A, we first predict its structural map using the frozen encoder-decoder as $\hat{m}_A = D_\phi(E_\omega(x_A))$. To make the condition closer to the mask-defined structure space, we do not directly use the image-derived latent representation $E_\omega(x_A)$ to condition the flow model. Instead, the predicted structural map is *re-encoded* into the shared structure space $s_A = E_\omega(\hat{m}_A)$. This image-to-mask-to-condition pathway ensures that the generator is conditioned mainly on source anatomy rather than source-domain appearance. Since no corresponding target image is available for x_A, we reuse the target-domain structure bank introduced in Stage I to construct a structure-matched soft target image. Based on the mask similarity between $\hat{m}_A$ and target masks in the bank, we retrieve the top-K target images and use the similarity weights α_j to construct a structure-matched soft target image. The cross-domain training objective is then defined as

$$\tilde{x}_B = \sum_{j \in K} \alpha_j x^j_B, \quad \mathcal{L}_{\text{cross}} = \mathcal{L}_{\text{cond}}(\tilde{x}_B, s_A). \tag{13}$$

The soft target image $\tilde{x}_B$ is not treated as ground-truth supervision for x_A. Instead, it provides a target-domain appearance reference with a similar structure, enabling the generator to learn target-style rendering for source-derived structure conditions. This facilitates unpaired supervision. Given a flow prediction at time t, we estimate the clean target image as $\hat{x}_0 = x_t + (1-t)v_\theta(x_t, t \mid s)$. The estimated image is passed through the frozen structure encoder-decoder, and the recovered structural map is compared with the conditioning structure:

$$\mathcal{L}_{\text{struct}} = 1 - \text{Dice}\left(D_\phi(E_\omega(\hat{x}_0)), m_s\right), \tag{14}$$

where m_s denotes the structure map used to define the condition, such as m_B for the target self-domain branch and $\hat{m}_A$ for the source cross-domain branch. This loss encourages the generator to produce target-domain images whose recovered structures remain consistent with the conditioning structure.

The overall Stage II objective is:

$$\mathcal{L}_{\text{stageII}} = \mathcal{L}_{\text{self}} + \lambda_{\text{cross}}\mathcal{L}_{\text{cross}} + \lambda_{\text{struct}}\mathcal{L}_{\text{struct}}. \tag{15}$$

The self-domain branch learns target-domain appearance generation from real target images, the cross-domain branch enables unpaired source-to-target translation through structure-matched soft targets, and the structure consistency loss encourages the generated image to preserve the specified anatomical structure.

3.3 Inference

We illustrate the inference procedure in Fig. 3. During inference, only the source image x_A is available, and our aim is to generate the corresponding target-domain image $\hat{x}_B$. From x_A, we first predict its structural map $\hat{m}_A = D_\phi(E_\omega(x_A))$. This map is then *re-encoded* into the shared structure space to obtain the flow model condition, $s_A = E_\omega(\hat{m}_A)$. Starting from Gaussian noise, the translated target-domain image is generated by solving the conditional flow ODE: $\frac{dx_t}{dt} = v_\theta(x_t, t \mid s_A)$ from $t = 0$ to $t = 1$. The final output is $\hat{x}_B = x_{t=1}$. No target image/mask or prototype retrieval is required during inference.

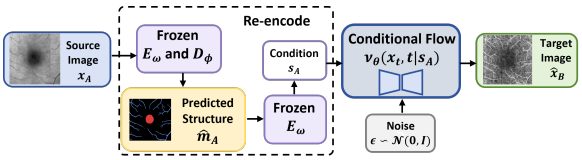


Figure 3: Inference pipeline of our structure-conditioned generation framework.

4 Experiments

Datasets. We used the **OCTA-500** [26] dataset for 2D retinal OCT-to-OCTA translation. The original projections have resolutions of 304×304 or 400×400 pixels. We split the dataset at the case level into 450 training and 50 test cases, keeping all projections from each case in the same partition. In Stage I, the official full, ILM-OPL, and OPL-BM projections were used as natural augmentation on the same split, yielding 1,350 training and 150 test projection images; in Stage II, only the full projection was used, yielding 450 training and 50 test images. All images were resized to 256×256. The target-domain mask m_B corresponds to the official OCTA large-vessel and FAZ annotations. We also used the Task 1 brain subset of **SynthRAD2023** [27] for MRI-to-CT synthesis. The released data includes 180 subjects with registered MR-CT volumes and 30 MR-only validation subjects; since validation CT is unavailable, we randomly selected 20 registered subjects as an internal test set. We extracted 9,646 2D axial slices from the supratentorial brain and skull regions, yielding 8,685 training and 961 test slices. All slices were resized with aspect-ratio preservation and padded to 256×256. Anatomical labels were generated using TotalSegmentator [28] and merged into six coarse classes. Here, the target-domain mask m_B corresponds to CT anatomical labels, while the source-domain coarse mask m_A corresponds to MR coarse anatomical labels. Within each training stage, source and target images for our method are drawn from non-overlapping cases. Further details are provided in the Appendix C.

Baselines. We compare our method with representative image-to-image translation baselines under both paired and unpaired settings, covering GAN-based, contrastive, disentanglement-based, diffusion-based, and bridge-based methods. For unpaired translation, we include CycleGAN [6], CUT [7], MUNIT [9], SynDiff [1], and UNSB [15]. CycleGAN is a cycle-consistent GAN, CUT uses patch-wise contrastive learning to preserve content correspondence, and MUNIT disentangles content and style representations for multimodal translation. In addition, SynDiff introduces an adversarial diffusion framework for unpaired medical image translation, while UNSB models unpaired translation through a Neural Schrödinger Bridge formulation. For paired translation, we compare against Pix2pix [3], DDBM [4], and MuTri [2]. Pix2pix serves as a classical conditional GAN baseline, DDBM learns a diffusion bridge between paired endpoint distributions, and MuTri [2] is a recent 3D OCT-to-OCTA translation framework based on VQ-VAE [29] and multi-view tri-alignment. We use official implementations and recommended configurations, with common train/test partitions, preprocessing, resolution, and evaluation protocols.

Table 1: Comparison on the OCT to OCTA translation task. We report mean ± 95% confidence interval for whole-image, vessel-region, and FAZ-region metrics. **Bold** denotes the best result, significantly outperforming the best unpaired baseline.

Method	Paired	Whole Image		Vessel		FAZ	
		PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
CycleGAN [6]	×	14.21±0.18	5.71±1.11	9.74±0.27	6.34±1.81	16.79±0.77	23.15±1.87
CUT [7]	×	13.62±0.22	7.91±2.01	14.17±0.37	8.24±1.47	15.79±0.54	19.87±2.23
MUNIT [9]	×	16.12±0.16	15.44±1.02	13.81±0.35	32.82±2.41	16.51±0.75	10.57±3.08
SynDiff [1]	×	14.41±0.20	6.48±1.46	12.06±0.30	5.70±1.52	18.09±0.81	27.99±2.02
UNSB [15]	×	14.31±0.23	7.16±1.31	10.19±0.36	5.32±1.62	18.27±0.88	32.46±2.42
DDBM [4]	✓	15.84±0.21	27.15±0.73	14.84±0.42	52.46±3.27	20.18±1.15	33.01±2.55
Pix2pix [3]	✓	15.61±0.20	23.83±0.91	14.82±0.31	50.66±3.09	20.15±0.74	29.79±1.86
MuTri [2]	✓	18.59±0.46	51.16±1.76	17.02±0.49	70.09±2.73	20.21±1.02	34.16±2.88
Ours	×	**19.05±0.72**	**64.65±5.40**	**18.92±0.88**	**75.72±5.81**	**20.48±1.00**	**35.05±4.03**

Table 2: Comparison on the brain MR-to-CT translation task. We report mean ± 95% confidence interval for whole-image, soft-tissue-region, and bone-region metrics. **Bold** denotes the best result, significantly outperforming the best unpaired baseline.

Method	Paired	Whole Image		Soft Tissue		Bone	
		PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
CycleGAN [6]	×	19.48±0.10	88.52±0.16	23.76±0.18	60.56±0.62	9.86±0.09	22.60±1.05
CUT [7]	×	20.52±0.11	90.13±0.13	20.95±0.18	59.25±0.57	9.81±0.09	22.04±0.90
MUNIT [9]	×	21.79±0.13	90.80±0.16	21.78±0.18	51.02±0.61	11.75±0.09	44.14±1.04
SynDiff [1]	×	24.58±0.14	94.50±0.12	23.33±0.19	54.11±0.86	14.31±0.12	66.72±0.83
UNSB [15]	×	19.51±0.10	88.34±0.17	20.90±0.21	61.02±0.51	8.89±0.09	11.70±1.15
DDBM [4]	✓	27.67±0.14	97.01±0.09	27.20±0.12	68.94±0.52	16.92±0.13	79.05±0.60
Pix2pix [3]	✓	27.75±0.14	96.66±0.15	26.96±0.13	66.80±0.49	16.91±0.13	78.42±0.65
Ours	×	**28.64±0.20**	**97.15±0.12**	**28.45±0.15**	**69.20±0.55**	**17.62±0.17**	**80.30±0.89**

Evaluation Metrics. We evaluate translation quality using PSNR and SSIM [30, 31] against ground-truth target images, reported over the whole image and task-specific regions: vessel/FAZ for OCT-to-OCTA and soft-tissue/bone for MR-to-CT. Soft tissue includes the cerebrum, ventricle, thalamus, caudate, and lentiform. Results are reported as mean ± 95% confidence interval, with SSIM shown as a percentage. We also use Dice [32] to assess Stage-I consistency between decoded/re-encoded structure masks and the target-domain mask m_B.

4.1 Results

Quantitative results. Although our method is trained in an unpaired setting, evaluation is conducted on test pairs by comparing each translated image with its ground-truth target image using PSNR and SSIM over the whole image and task-specific regions. Tab. 1 shows the results on OCT-to-OCTA translation. Our method consistently outperforms existing unpaired baselines on the whole image, vessel region, and FAZ region, demonstrating improved vascular recovery and structural correspondence. Notably, despite using no paired source-target supervision during training, it also surpasses paired baselines, improving over the best paired result by up to 11.2% in PSNR and 26.4% in SSIM. Tab. 2 further shows consistent gains on brain MR-to-CT translation. Compared with the best unpaired baseline, our method improves PSNR by 16.5%, 19.7%, and 23.1% on the whole image, soft-tissue region, and bone region, respectively, with SSIM gains across all regions. It also outperforms paired baselines on all metrics, achieving PSNR gains of 3.2%–4.6% with consistent SSIM improvements. Overall, these results show that our structure-conditioned formulation improves both target-domain fidelity and structure preservation, especially in anatomy-critical regions.

Table 3: Stage I effectiveness of the shared structure encoder.

Setting	$\mathrm{Dice}_{\mathrm{OCT}}$↑	$\mathrm{Dice}_{\mathrm{OCTA}}$↑	$\mathrm{Dice}_{\mathrm{re}}$↑	$\mathrm{Dice}_{\mathrm{pair}}$↑
w/o m_A	73.33±0.02	75.78±0.03	75.10±0.04	71.23±0.03
w/o $\mathcal{L}_{\mathrm{cycle}}$	72.96±0.04	74.66±0.03	75.06±0.04	69.56±0.03
Ours	**75.89±0.03**	**74.75±0.02**	**76.92±0.04**	**72.62±0.05**

Qualitative Results. Fig. 4 presents qualitative comparisons under both unpaired and paired settings on OCT-to-OCTA and brain MR-to-CT translation. For each method, the first row shows the translated target-domain image, and the second row shows the MAE map between the prediction and ground-truth target image. Overall, our method produces more realistic target-domain images with better anatomical preservation and lower reconstruction errors. As highlighted by the red boxes, existing baselines often lose fine structures, such as missing vessel branches in OCTA images and inconsistent ventricle shapes in brain CT images; these errors are also reflected by stronger MAE responses around vessel regions and anatomical boundaries. In contrast, our method better preserves these

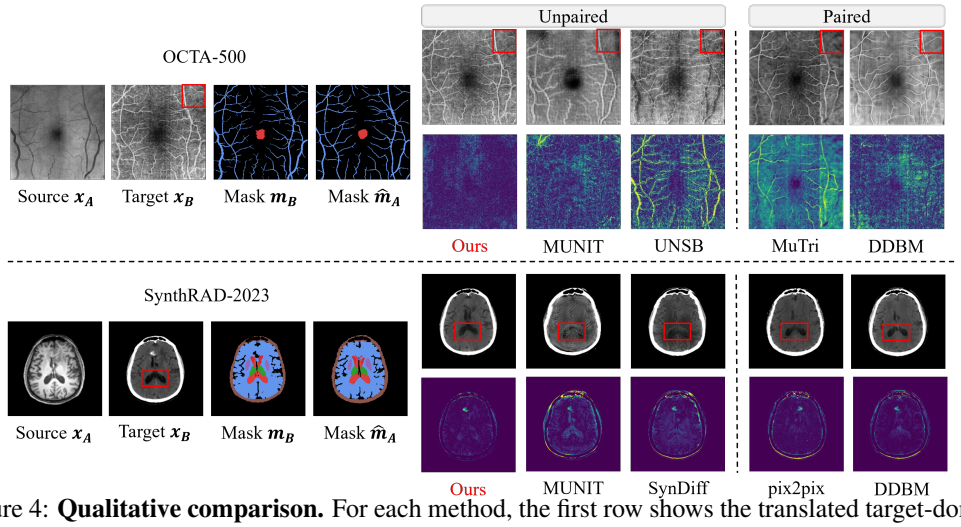
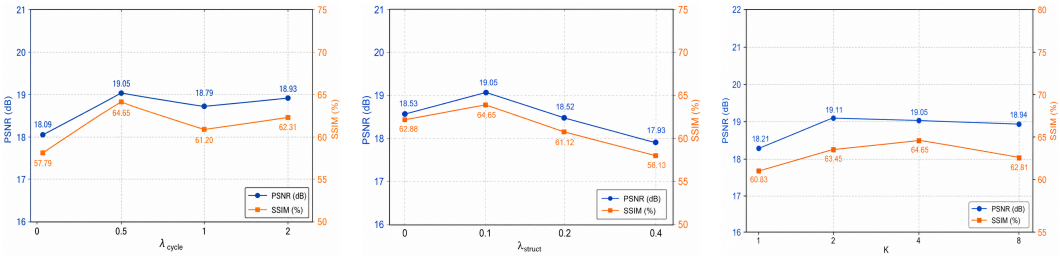


Figure 4: **Qualitative comparison.** For each method, the first row shows the translated target-domain image and the second row shows the mean absolute error (MAE) map against the ground truth. For our method, the source structural condition is obtained as $\hat{m}_A = D_\phi(E_\omega(x_A))$. For visualization, CT images are re-windowed to a brain window of $[0, 100]$ HU to enhance soft-tissue contrast, while quantitative metrics are computed in the original normalized CT space. Our method better preserves anatomy and reduces errors in structure-critical regions (red boxes).



(a) λ_{cycle} (b) λ_{struct} (c) Top-K prototypes

Figure 5: **Ablation studies on the OCT-to-OCTA translation task.** We evaluate the effects of key hyperparameters on translation performance using PSNR and SSIM. **(a)** Varying the weight λ_{cycle} of the cycle loss $\mathcal{L}_{\text{cycle}}$. **(b)** Varying the weight λ_{struct} of the structure consistency loss $\mathcal{L}_{\text{struct}}$ in Stage II. **(c)** Varying the number of retrieved prototypes K in the target-domain structure bank SMB.

structures, generating more continuous vascular patterns for OCT-to-OCTA translation and clearer CT-like soft-tissue and bone structures for brain MR-to-CT translation. These improvements come from explicitly separating structure preservation from appearance generation: the shared structure encoder provides domain-invariant anatomical guidance, while the target-domain flow model renders realistic appearance under this guidance. More qualitative results are provided in the Appendix G.

Ablation study. We analyze the effects of λ_{cycle}, λ_{struct}, and the number of retrieved prototypes K on OCT-to-OCTA translation, as shown in Fig. 5. Removing $\mathcal{L}_{\text{cycle}}$ clearly degrades performance, while non-zero values yield stable results, indicating that latent cycle consistency helps suppress appearance cues and align OCT- and OCTA-derived structure conditions. For λ_{struct}, a moderate weight improves structural guidance, but an overly large value reduces performance, suggesting that excessive regularization can bias the conditional flow away from the target OCTA distribution. For

Table 4: Center-held-out MR-to-CT evaluation. Results are patient-level means; PSNR is in dB and SSIM in %. "Paired" denotes image-pair supervision. Our method uses target-domain structural masks.

Method	**Paired**	**Whole Image**		**Soft Tissue**		**Bone**	
		PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
SynDiff	No	24.10	94.68	23.18	50.71	13.71	62.31
DDBM	Yes	25.13	95.79	23.27	58.71	14.10	67.49
Ours	No	**27.98**	**97.05**	**27.15**	**59.17**	**17.14**	**80.45**

Table 5: Downstream classification on OCTA-500. Brackets denote 95% case-level bootstrap confidence intervals. $|\Delta p|$ and flip rate measure probability and decision changes relative to acquired OCTA. Bold indicates the best mean among translated inputs.

Input	AUROC ↑	Bal. Acc. ↑	$\lvert\Delta p\rvert$ ↓	Flip rate ↓
Acquired OCTA	0.928 [0.848, 0.984]	0.850 [0.750, 0.944]	–	–
MuTri	0.702 [0.555, 0.840]	0.533 [0.469, 0.612]	0.288	0.30
SynDiff	**0.888** [0.773, 0.976]	0.533 [0.468, 0.612]	0.315	0.30
UNSB	0.525 [0.353, 0.700]	0.550 [0.500, 0.625]	0.355	0.28
MUNIT	0.455 [0.297, 0.619]	0.500 [0.500, 0.500]	0.416	0.28
Ours	0.887 [0.740, 0.995]	**0.867** [0.762, 0.958]	**0.132**	**0.08**

Table 6: Independent vessel evaluation on 50 held-out OCT-to-OCTA cases. Values are mean ± 95% case-level bootstrap CI half-width. CC denotes connected components. Bold indicates the best mean.

Method	Vessel Dice ↑	clDice ↑	Density Err. ↓	Halluc. ↓	Omiss. ↓	#CC Err. ↓
MUNIT	0.501 ± 0.029	0.493 ± 0.032	0.0304 ± 0.0037	0.362 ± 0.030	0.584 ± 0.029	28.9 ± 5.4
MuTri	0.787 ± 0.017	$\mathbf{0.818} \pm 0.019$	0.0118 ± 0.0026	0.162 ± 0.010	0.253 ± 0.028	25.3 ± 7.3
Ours	$\mathbf{0.797} \pm 0.026$	$\mathbf{0.818} \pm 0.027$	$\mathbf{0.0113} \pm 0.0031$	$\mathbf{0.152} \pm 0.014$	$\mathbf{0.243} \pm 0.036$	$\mathbf{17.8} \pm 6.0$

the structure bank, a single prototype provides an unreliable hard target, whereas multiple prototypes offer more robust soft structural guidance and stable performance across K. Overall, the results show that our framework benefits from cycle consistency, a non-excessive structure constraint, and multi-prototype aggregation, while remaining relatively insensitive to exact hyperparameter settings.

Tab. 3 evaluates the Stage I shared structure encoder. Let $\hat{m}_A = D_\phi(E_\omega(x_A))$, $\hat{m}_B = D_\phi(E_\omega(x_B))$, and $\widetilde{m}_A = D_\phi(E_\omega(\hat{m}_A))$. We define $\mathrm{Dice}_{\mathrm{OCT}} = \mathrm{Dice}(\hat{m}_A, m_B)$, $\mathrm{Dice}_{\mathrm{OCTA}} = \mathrm{Dice}(\hat{m}_B, m_B)$, $\mathrm{Dice}_{\mathrm{re}} = \mathrm{Dice}(\widetilde{m}_A, m_B)$, and $\mathrm{Dice}_{\mathrm{pair}} = \mathrm{Dice}(\hat{m}_A, \hat{m}_B)$. Removing the source coarse mask input m_A causes only limited changes: although the full model performs better on $\mathrm{Dice}_{\mathrm{OCT}}$, $\mathrm{Dice}_{\mathrm{re}}$, and $\mathrm{Dice}_{\mathrm{pair}}$, the variant without m_A remains comparable and even improves $\mathrm{Dice}_{\mathrm{OCTA}}$. This indicates that m_A provides auxiliary source-side guidance but is not the dominant factor for structure extraction. In contrast, removing $\mathcal{L}_{\mathrm{cycle}}$ more clearly reduces cross-domain structural consistency, with $\mathrm{Dice}_{\mathrm{pair}}$ dropping from 72.62 to 69.56. Overall, the encoder can learn reasonable structural representations even without m_A, while cycle consistency is more important for aligning OCT and OCTA structures in the shared space. Qualitative mask visualizations are provided in the Appendix Fig. 7.

Cross-center generalization. We retrain the evaluated methods on SynthRAD2023 centers A and B and test on unseen center C without adaptation. No data or prototypes from center C are used during training or model selection. As shown in Tab. 4, our method achieves the best mean performance across all three anatomical regions.

Downstream and structural evaluation. We evaluate generated OCTA using a normal-versus-abnormal classifier trained exclusively on acquired OCTA from 450 training cases and frozen for testing on 50 held-out cases. As shown in Tab. 5, our method achieves the highest balanced accuracy among translated inputs and the lowest probability deviation and decision-flip rate relative to acquired OCTA. On the same cases, we assess vessel fidelity using a frozen evaluator independent of translation training and model selection. Tab. 6 shows that our method achieves the highest mean vessel Dice and lowest mean structural errors, with comparable clDice [33] to MuTri.

5 Conclusion

We proposed a two-stage unpaired medical image translation framework that separates structure representation learning from target-domain generation. Stage I learns a shared structure space through mask reconstruction and prototype-based soft supervision to capture modality-agnostic anatomy. Stage II freezes the structure encoder-decoder and uses its structural representations to condition a target-domain flow model, enabling appearance synthesis under fixed anatomical guidance. Experiments on OCT-to-OCTA and brain MR-to-CT translation show that our method outperforms both unpaired and paired baselines, preserving clinically meaningful structures while generating realistic target-domain images. These results highlight the benefit of decoupling structural guidance from appearance rendering for robust unpaired medical image translation.

Acknowledgments

This research was partially supported by grants NSF CCF-2144901, NIH R01NS143143 and R01CA297843.

References

[1] M. Özbey, O. Dalmaz, S. U. Dar, H. A. Bedel, Ş. Özturk, A. Güngör, and T. Cukur, "Unsupervised medical image translation with adversarial diffusion models," *IEEE Transactions on Medical Imaging*, vol. 42, no. 12, pp. 3524–3539, 2023.

[2] Z. Chen, H. Wang, C. Ou, and X. Li, "Mutri: Multi-view tri-alignment for oct to octa 3d image translation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 20 885–20 894.

[3] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[4] L. Zhou, A. Lou, S. Khanna, and S. Ermon, "Denoising diffusion bridge models," *arXiv preprint arXiv:2309.16948*, 2023.

[5] X. Zhang, J. Liang, P. Cao, J. Yang, and O. R. Zaiane, "Structure-aware mri translation: Multi-modal latent diffusion model with arbitrary missing modalities," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2025, pp. 508–518.

[6] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[7] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *European conference on computer vision*. Springer, 2020, pp. 319–345.

[8] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," *arXiv preprint arXiv:2210.02747*, 2022.

[9] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 172–189.

[10] H. Fu, M. Gong, C. Wang, K. Batmanghelich, K. Zhang, and D. Tao, "Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2427–2436.

[11] S. Benaim and L. Wolf, "One-sided unsupervised domain mapping," *Advances in neural information processing systems*, vol. 30, 2017.

[12] C. Jung, G. Kwon, and J. C. Ye, "Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 260–18 269.

[13] W. Wang, W. Zhou, J. Bao, D. Chen, and H. Li, "Instance-wise hard negative example generation for contrastive learning in unpaired image-to-image translation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 14 020–14 029.

[14] C. Zheng, T.-J. Cham, and J. Cai, "The spatially-correlative loss for various image translation tasks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 407–16 417.

[15] B. Kim, G. Kwon, K. Kim, and J. C. Ye, "Unpaired image-to-image translation via neural schrödinger bridge," in *International Conference on Learning Representations*, vol. 2024, 2024, pp. 19 312–19 331.

[16] J. Yi, Q. Bi, H. Zheng, H. Huang, Y. Shen, H. Zhan, W. Ji, Y. Huang, Y. Li, X. Wu *et al.*, "Degradation-aware dynamic schrödinger bridge for unpaired image restoration," *Advances in Neural Information Processing Systems*, vol. 38, pp. 2415–2439, 2026.

[17] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[18] H. Zhang, Y. Liu, J. Yang, S. Wan, X. Wang, W. Peng, and P. Fua, "Lefusion: Controllable pathology synthesis via lesion-focused diffusion models," *arXiv preprint arXiv:2403.14066*, 2024.

[19] C. Hansen, S. Glinskis, A. Raju, M. Kornreich, J. Park, J. Pawar, R. Herzog, L. Zhang, and B. Odry, "Inpainting pathology in lumbar spine mri with latent diffusion," *arXiv preprint arXiv:2406.02477*, 2024.

[20] S. Ahmed, J. Feng, A. Mehmood, M. U. Ali, M. Yaqub, M. A. Manan, and A. Raheem, "Pinn-dadif: Physics-informed deep adaptive diffusion network for robust and efficient mri reconstruction," *Digital Signal Processing*, vol. 160, p. 105085, 2025.

[21] H. Xie, W. Gan, B. Zhou, M.-K. Chen, M. Kulon, A. Boustani, B. A. Spencer, R. Bayerlein, X. Chen, Q. Liu *et al.*, "Dose-aware diffusion model for 3d low-dose pet: multi-institutional validation with reader study and real low-dose data," *arXiv preprint arXiv:2405.12996*, 2024.

[22] D. Hein, S. Holmin, T. Szczykutowicz, J. S. Maltz, M. Danielsson, G. Wang, and M. Persson, "Ppfm: Image denoising in photon-counting ct using single-step posterior sampling poisson flow generative models," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 8, no. 7, pp. 788–799, 2024.

[23] S. Gupta, D. Samaras, and C. Chen, "Topodiffusionnet: A topology-aware diffusion model," *arXiv preprint arXiv:2410.16646*, 2024.

[24] M. Xu, S. Gupta, X. Hu, C. Li, S. Abousamra, D. Samaras, P. Prasanna, and C. Chen, "Topocellgen: Generating histopathology cell topology with a diffusion model," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 20 979–20 989.

[25] B. Asadi, P. Wu, M. Golparvar-Fard, V. Shah, and R. Hajj, "FMS^2: Unified flow matching for segmentation and synthesis of thin structures," *arXiv preprint arXiv:2603.13659*, 2026.

[26] M. Li, K. Huang, Q. Xu, J. Yang, Y. Zhang, Z. Ji, K. Xie, S. Yuan, Q. Liu, and Q. Chen, "Octa-500: a retinal dataset for optical coherence tomography angiography study," *Medical image analysis*, vol. 93, p. 103092, 2024.

[27] A. Thummerer, E. Van der Bijl, A. Galapon Jr, J. J. Verhoeff, J. A. Langendijk, S. Both, C. N. A. van den Berg, and M. Maspero, "Synthrad2023 grand challenge dataset: Generating synthetic ct for radiotherapy," *Medical physics*, vol. 50, no. 7, pp. 4664–4674, 2023.

[28] J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. W. Sauter, T. Heye, D. T. Boll, J. Cyriac, S. Yang *et al.*, "Totalsegmentator: robust segmentation of 104 anatomic structures in ct images," *Radiology: Artificial Intelligence*, vol. 5, no. 5, p. e230024, 2023.

[29] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[30] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th international conference on pattern recognition*. IEEE, 2010, pp. 2366–2369.

[31] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[32] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[33] S. Shit, J. C. Paetzold, A. Sekuboyina, I. Ezhov, A. Unger, A. Zhylka, J. P. Pluim, U. Bauer, and B. H. Menze, "cldice-a novel topology-preserving loss function for tubular structure segmentation," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021, pp. 16 555–16 564.

Appendix

A Additional Ablation Studies

We further ablate the weights of $\mathcal{L}_{\text{align}}$, $\mathcal{L}_{\text{dec}}$, and $\mathcal{L}_{\text{bank}}$ to study their effects on the learned shared structure space, as shown in Tabs. 7–9. For the latent structure alignment loss, Tab. 7 shows that $\lambda_{\text{align}} = 0.5$ gives the best result, achieving 19.05 dB PSNR and 64.65% SSIM, while both smaller and larger weights degrade performance. This suggests that proper latent alignment is important for learning domain-shared structural representations, but excessive alignment may over-constrain the latent space. For the structure reconstruction loss, Tab. 8 shows that $\lambda_{\text{dec}} = 1$ performs best, indicating that decoding the shared latent representation into structure masks provides useful anatomical supervision. However, increasing the weight to 2 reduces performance, suggesting that overly strong reconstruction supervision may limit the flexibility of the structure latent. For the structure-bank consistency loss, Tab. 9 shows that $\lambda_{\text{bank}} = 0.2$ achieves the best performance. A smaller weight provides insufficient target-domain structural guidance, whereas a larger weight may introduce less relevant prototype constraints. Overall, these results demonstrate that the three Stage-I losses are complementary and require balanced weighting to learn an effective shared structure representation for downstream OCT-to-OCTA translation.

Table 10 studies the effects of the Stage I coarse source mask m_A and the Stage II cross-domain branch. Removing both components leads to the lowest performance, with 16.70 dB PSNR and 61.29% SSIM on the whole image. Adding only the cross-domain branch substantially improves the whole-image performance to 18.23 dB PSNR and 63.24% SSIM, suggesting that Stage II cross-domain guidance helps the model better match target-domain structures. Using only the coarse mask m_A also improves the result to 17.82 dB PSNR and 62.96% SSIM, showing that the Stage I source structural prior contributes to anatomical preservation. Combining both components achieves the best performance across all evaluated regions, including the whole image, vessel region, and FAZ region. In particular, the full model reaches 19.05 dB PSNR and 64.65% SSIM on the whole image, 18.92 dB PSNR and 75.72% SSIM on vessels, and 20.48 dB PSNR and 35.05% SSIM on FAZ. These results indicate that the coarse source mask and cross-domain branch provide complementary structural guidance for OCT-to-OCTA translation.

We further analyze the design choice of performing structure conditioning with the memory bank in the shared latent space. As shown in Fig. 6, the decoded masks of the top-K retrieved prototypes provide relevant structural candidates, but they are not perfectly aligned in the image domain. Therefore, directly aggregating these decoded masks with similarity weights leads to an unreasonable structure, as shown in the last column: the direct combination produces noisy and overly dense vessel responses that do not preserve a plausible anatomical layout. In contrast, our method first performs similarity-weighted aggregation over the retrieved prototypes in the shared structure latent space and then decodes the aggregated representation into the estimated structure mask. This latent-space aggregation yields a more reasonable and coherent structure that is closer to the target mask. The comparison indicates that the shared structure latent space provides a better manifold for combining anatomical priors, while direct image-domain aggregation fails to respect spatial and topological consistency. This supports our design of using a memory bank together with latent-space structure conditioning for faithful cross-domain translation.

B Implementation details

OCT-to-OCTA experiments: In Stage I, the shared structure encoder-decoder is trained progressively: we first warm up the model with target masks m_B for 10 epochs, then introduce annotated target images x_B for another 10 epochs, and finally add source images x_A together with the SMB. The Stage I loss weights are set to $\lambda_{\text{dec}} = 1.0$, $\lambda_{\text{align}} = 0.5$, $\lambda_{\text{bank}} = 0.2$, and $\lambda_{\text{cycle}} = 0.5$. In Stage II, after freezing the structure encoder-decoder, the conditional flow generator is trained with $\lambda_{\text{self}} = 1.0$, $\lambda_{\text{cross}} = 0.3$, and $\lambda_{\text{struct}} = 0.1$. All weights are kept fixed throughout training. For consistency across the two stages, we retrieve the top-K prototypes from the memory bank with $K = 4$.

Brain MR-to-CT experiments: We follow the same two-stage training strategy as described in the main paper, while using task-specific loss weights. In Stage I, the shared structure encoder-decoder is trained with $\lambda_{\text{dec}} = 1.0$, $\lambda_{\text{align}} = 0.5$, $\lambda_{\text{bank}} = 0.2$, and $\lambda_{\text{cycle}} = 0.5$. In Stage II, the

Table 7: Ablation on the weight λ_{align} of the latent structure alignment loss $\mathcal{L}_{\text{align}}$ on the OCT-to-OCTA translation task. We report PSNR↑ and SSIM↑ (mean ± 95% CI). Best results are in **bold**.

λ_{align}	PSNR↑	SSIM↑
0.25	18.35±0.64	59.39±5.33
0.5	**19.05±0.72**	**64.65±5.40**
1	18.27±0.56	58.43±4.97

Table 8: Ablation on the weight λ_{dec} of the structure reconstruction loss $\mathcal{L}_{\text{dec}}$ on the OCT-to-OCTA translation task. We report PSNR↑ and SSIM↑ (mean ± 95% CI). Best results are in **bold**.

λ_{dec}	PSNR↑	SSIM↑
0.5	18.40±0.57	61.31±4.87
1	**19.05±0.72**	**64.65±5.40**
2	18.64±0.65	59.42±5.36

structure encoder-decoder is frozen, and the conditional flow generator is optimized with $\lambda_{\text{self}} = 1.0$, $\lambda_{\text{cross}} = 1.0$, and $\lambda_{\text{struct}} = 0.5$. For consistency across the two stages, we retrieve the top-K prototypes from the memory bank with $K = 4$. For the similarity weighting in the memory bank, we use the same parameters for both OCT-to-OCTA and brain MR-to-CT experiments, setting $\gamma = 2$ and $\tau = 0.05$.

Training procedure. During training, we first warm up the structure encoder-decoder using only target masks m_B to construct a mask-defined structural latent space. We then introduce target images x_B with annotations m_B to align image-derived latents with this space. Finally, we add source images x_A and construct the SMB from target images and masks to weakly align source structures with the target structural distribution. After Stage I, the structure encoder-decoder is frozen to provide structural conditions for the target-domain flow generator. Before training the structure-conditioned generator in Stage II, we pretrain an unconditional pixel-space flow model using only target-domain images. This model learns the marginal appearance distribution of the target modality and serves as the initialization for the conditional flow generator. With this target-domain appearance prior, the conditional generator does not need to learn target appearance from scratch, and can focus on rendering target-domain images under the structural conditions provided by the frozen structure encoder-decoder. During Stage II, the self-branch and cross-branch are trained jointly, where the self-branch learns target-domain reconstruction under target structural conditions, and the cross-branch encourages source-to-target generation using source-derived structural conditions within the same conditional flow model.

Construction of coarse source masks. In the source domain, the coarse mask m_A is not assumed to be a manual annotation, and its construction depends on the translation task. The purpose of m_A is to provide a weak structural cue for source images, rather than an accurate pixel-level segmentation target. For OCT-to-OCTA translation, we obtain coarse OCT structural masks by applying intensity-based thresholding to highlight major vessel-like responses and the FAZ region, followed by basic morphological operations to remove small noisy components and improve spatial continuity. These masks are therefore expected to be incomplete and noisy compared with the target-domain OCTA structural masks. For MR-to-CT translation, source-domain structural masks are constructed from anatomical segmentation labels, with the label sources and class definitions detailed in Appendix C. For MR subjects with paired CT, the bone label is derived from the corresponding CT skull segmentation. These coarse masks may contain missing regions or inaccurate boundaries.

Table 9: Ablation on the weight λ_{bank} of the structure-bank consistency loss $\mathcal{L}_{\text{bank}}$ on the OCT-to-OCTA translation task. We report PSNR↑ and SSIM↑ (mean ± 95% CI). Best results are in **bold**.

λ_{bank}	PSNR↑	SSIM↑
0.1	18.32±0.65	58.66±5.41
0.2	**19.05±0.72**	**64.65±5.40**
0.4	18.89±0.63	62.36±4.95

Table 10: Ablation of coarse source mask input m_A and cross-domain branch. We report PSNR↑ and SSIM↑ for the whole image, vessel region, and FAZ region. Best results are in **bold**.

m_A	Cross branch	Whole Image		Vessel		FAZ	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
×	×	16.70±0.74	61.29±5.65	16.44±0.95	70.57±6.08	18.19±0.99	31.46±3.89
×	✓	18.23±0.99	63.24±5.59	17.72±0.90	74.26±5.91	19.34±0.96	33.36±4.00
✓	×	17.82±0.73	62.96±5.26	17.18±0.90	72.69±5.63	19.08±1.00	32.10±3.88
✓	✓	**19.05±0.72**	**64.65±5.40**	**18.92±0.88**	**75.72±5.81**	**20.48±1.00**	**35.05±4.03**

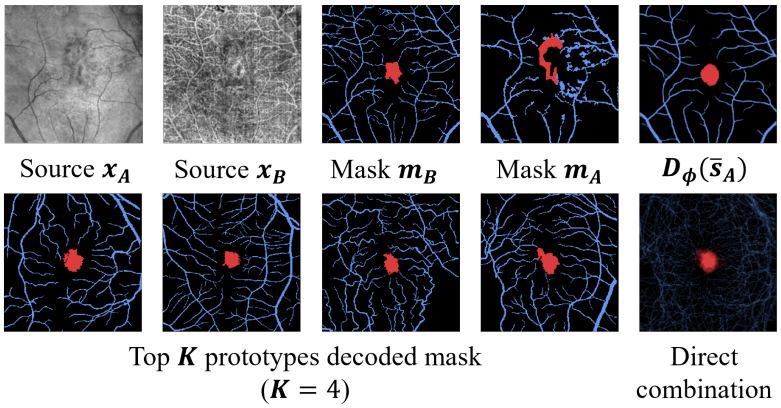


Figure 6: **Memory bank analysis.** Given a source image x_A, we retrieve the top-K target-domain prototypes from the structure memory bank based on mask similarity. The bottom row shows the decoded masks of the retrieved prototypes, while the rightmost result shows direct similarity-weighted combination in the mask space. In contrast, our method first aggregates the retrieved prototypes in the shared structure latent space and then decodes the aggregated latent as $D_\phi(\bar{s}_A)$. Compared with direct mask-space combination, latent-space aggregation produces a cleaner and more coherent structural map that better matches the source-derived structure, supporting our design of using the shared structure latent space to construct structural conditions.

Our method treats them only as optional weak supervision; the core training objective relies on the target-domain structural masks m_B to define the shared structure space.

C Dataset details

For OCTA-500 [26], the original 2D OCTA projection images contain two in-plane resolutions, 304×304 and 400×400 pixels. We used the official layer-wise projections defined by the internal limiting membrane (ILM), outer plexiform layer (OPL), and Bruch's membrane (BM). Specifically, the full, ILM-OPL, and OPL-BM projections were used in Stage 1 as natural augmentation, while only the full projection was used in Stage 2. We randomly selected 50 samples as the test set and used the remaining 450 samples for training. Therefore, Stage 1 contains 1,350 training images and 150 test images, and Stage 2 contains 450 training images and 50 test images. The official OCTA-500 segmentation annotations were used, and we focused on the large-vessel and foveal avascular zone (FAZ) regions. All projection images and their corresponding masks were resized with aspect-ratio preservation and padded to 256×256.

For SynthRAD2023 [27], we used the Task 1 brain subset for MRI-to-CT synthesis. The released data contain 180 paired MR/CT training subjects and 30 MR-only validation subjects. Since the CT images are unavailable for the official validation subjects, we randomly selected 20 paired training subjects as an internal test set and used the remaining paired subjects for training. The official preprocessing resampled the brain CT volumes to isotropic $1 \times 1 \times 1\ \mathrm{mm}^3$ spacing and rigidly registered the MR volumes to the resampled CT space. The released volumes have case-dependent matrix sizes due to subject-specific cropping, with an average training-set size of approximately $213 \times 243 \times 194$ voxels. From these 3D volumes, we extracted 2D axial slices from the supratentorial brain and skull regions,

resulting in 9,646 slices in total, including 8,685 training slices and 961 test slices. The original axial slices therefore have variable in-plane sizes, with an average size of approximately 213×243 pixels. All slices and their corresponding masks were resized with aspect-ratio preservation and padded to 256×256. CT-side anatomical labels were generated using TotalSegmentator [28] and merged into six foreground classes plus background. No additional registration was applied beyond the official SynthRAD2023 preprocessing.

For Stage I training on the OCTA-500 dataset under the unpaired setting, we split the data into 225 OCT cases and 225 non-overlapping OCTA cases, ensuring no sample overlap between the two domains. For Stage II, since the cross-domain branch serves as an auxiliary augmentation to the main target-domain training, we used 100 OCT cases and 350 non-overlapping OCTA cases. For the SynthRAD2023 dataset, we used 90 CT cases and 100 MR cases in Stage I, and 140 CT cases and 50 MR cases in Stage II. MR cases from the official SynthRAD2023 validation subjects were included in both Stage I and Stage II MR sets because they do not have paired CT scans.

During the conversion from 3D volumes to 2D axial slices, we restricted the field of view to the supratentorial brain and skull region. Slices containing inferior or non-target structures, including the cerebellum, brainstem, orbits, nasal cavity, and neck, were excluded. Specifically, the lower boundary was determined by the highest axial slice containing the cerebellum plus 21 slices, which removed the cerebellum and inferior anatomy with an additional 20-slice margin. The upper boundary was set to $z_{\mathrm{dim}} - 21$, removing the last 20 slices near the top of the head. As a result, the retained slices mainly covered the cerebrum, ventricles, selected deep gray matter structures, and skull.

For anatomical label generation, the original segmentation outputs were merged into six foreground classes plus background: cerebrum, ventricle, thalamus, caudate, lentiform nucleus, and bone, with background treated as a separate class. CT-side labels were obtained from TotalSegmentator, while MR-side brain labels were generated using SynthSeg-compatible anatomical labels. For MR subjects with paired CT, the bone label was derived from the corresponding CT skull segmentation and used for the MR image. This merging strategy was designed to provide coarse anatomical guidance rather than fine-grained neuroanatomical parcellation.

D Limitations

Although we evaluate generalization to one held-out SynthRAD2023 center, this setting does not cover the full range of institutional, scanner, protocol, and population differences. Our experiments use 2D en-face OCT/OCTA projections and independently translated axial MR/CT slices, without explicitly enforcing inter-slice consistency or 3D anatomical topology. The framework also relies on target-domain structural masks during training, and mask errors may affect the learned representation. Local anatomical hallucination and omission remain possible despite the observed improvements in structural fidelity. Broader external validation and expert assessment are needed to establish reliability for clinical applications.

E Broader impacts

The proposed framework may help improve the accessibility of multi-modal medical imaging by synthesizing target-domain images from more readily available source modalities, potentially reducing acquisition burden, cost, and patient discomfort in research and clinical-support settings. By explicitly modeling anatomical structure, the method aims to improve structural fidelity compared with direct translation approaches, which is important for downstream analysis and visualization. However, synthesized medical images should not be used as a substitute for acquired clinical scans without rigorous validation. Potential risks include hallucinated or missing structures, dataset bias, and reduced reliability under protocol or institution shifts. Therefore, any clinical deployment should require careful multi-center evaluation, uncertainty assessment, and expert oversight.

F Training details

Stage I trains the structure encoder–decoder with a batch size of 32 for 10K optimization steps, taking approximately 2 hours on 4 NVIDIA RTX A6000 GPUs under mixed precision. Stage II trains the structure-conditioned flow-matching generator for 80K steps, finishing in roughly 2 hours on the

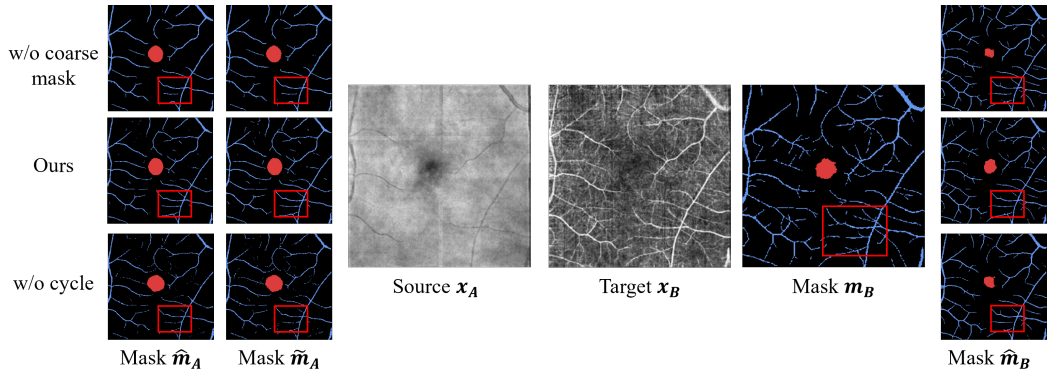


Figure 7: **Qualitative analysis of Stage I structure encoding.** We visualize the predicted structure masks under different Stage I settings, corresponding to Table 3. Red boxes highlight local vascular regions. Compared with the ablated variants, the full model preserves more continuous vessel structures with fewer broken or missing segments, demonstrating the benefit of using the source coarse mask m_A and the latent cycle consistency loss $\mathcal{L}_{\text{cycle}}$ for learning a shared structure space.

same hardware. All otherwise unspecified hyperparameters follow their default values, including the AdamW optimizer with learning rate 2×10^{-4} and weight decay 10^{-4}, gradient clipping at 1.0, and EMA for the flow model.

G Additional qualitative and robustness analyses

Fig. 7 provides qualitative analysis of the Stage-I structure encoder under different training settings. The full model produces more continuous and complete vessel structures than the ablated variants, especially in the local regions highlighted by red boxes. Compared with the setting without the coarse source mask m_A, our model better preserves source-side vascular details, indicating that m_A provides useful auxiliary structural guidance. Compared with the setting without the latent cycle consistency loss $\mathcal{L}_{\text{cycle}}$, our model also reduces broken or missing vessel segments, suggesting that the cycle constraint helps maintain a more consistent shared structure space.

Figs. 8 and 9 show additional qualitative results for OCT-to-OCTA and MR-to-CT translation, respectively. For OCT-to-OCTA translation, our method preserves vessel-like anatomical structures while generating target-domain OCTA appearance, whereas baseline methods often introduce noisy responses or lose fine vascular details. For MR-to-CT translation, our method produces more anatomically consistent CT-like images across different brain slices, with better preservation of tissue layouts and structure boundaries. These results further demonstrate that the proposed structure-conditioned generation framework improves target-domain appearance synthesis while maintaining clinically meaningful anatomical structures.

G.1 Robustness to automatically generated target masks

We assess sensitivity to target-domain mask quality on OCT-to-OCTA translation. Automatic vessel and FAZ masks are extracted from target-domain OCTA training images using fixed thresholding and morphological post-processing, without using the official annotations during mask generation. We retrain the full framework from scratch using either the official annotations or the automatic masks, keeping all other settings fixed. Target-domain masks are used only during training.

Tab. 11 reports the quality of the automatic masks and the resulting translation performance. Despite imperfect agreement with the official annotations, automatic masks yield similar mean PSNR and SSIM across the whole image, vessel, and FAZ regions. These results support robustness to the mask errors introduced by this automatic extraction procedure on OCT-to-OCTA.

G.2 Memory-bank bias and subject-specific fidelity

We examine whether outputs resemble retrieved training prototypes at the expense of subject-specific anatomy. The memory bank is used only during training; retrieval for this analysis is performed post

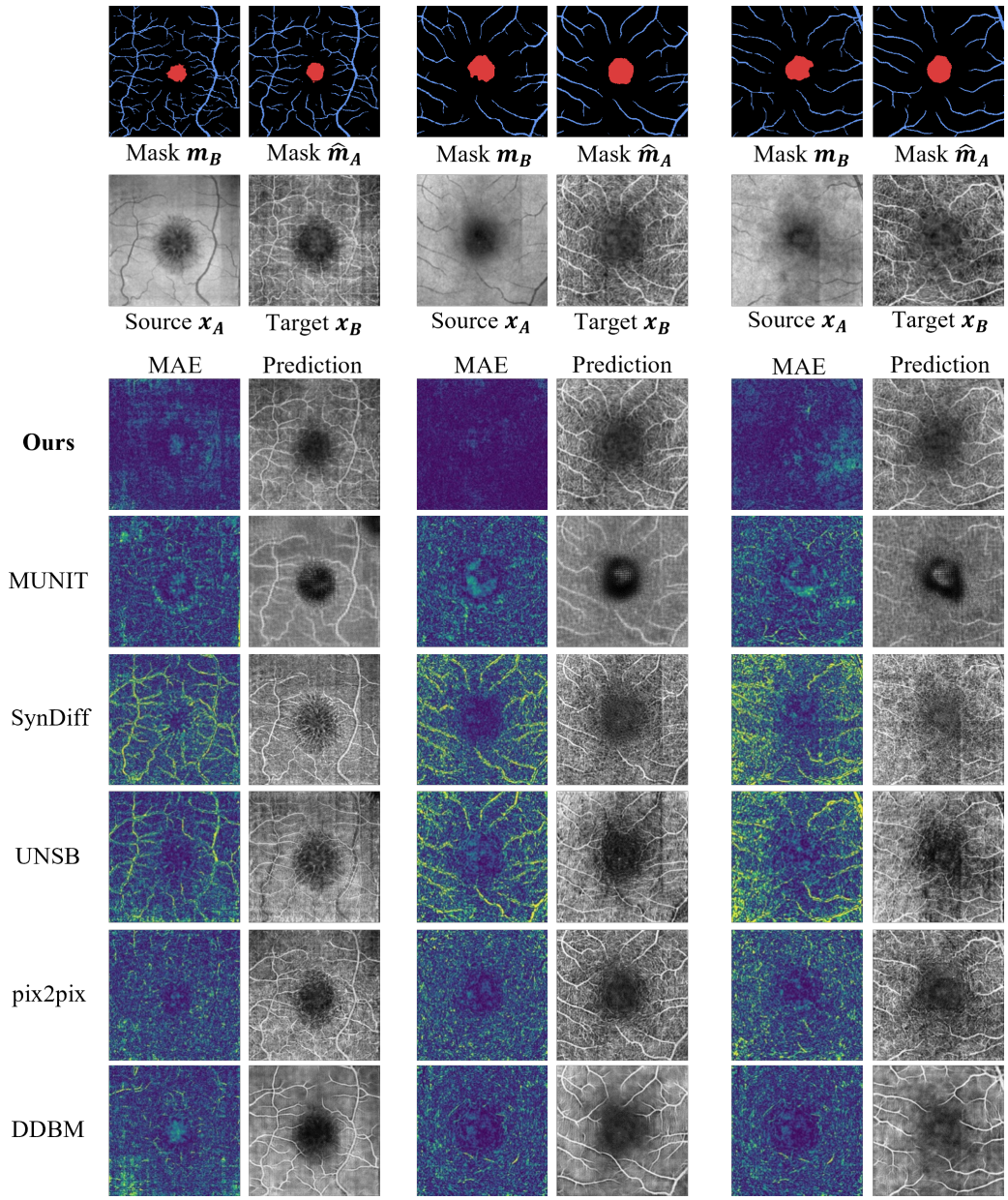


Figure 8: Additional qualitative results for OCT-to-OCTA translation. The figure shows multiple representative examples with the corresponding structural annotations, source/target images, intermediate response maps, and generated OCTA results. These examples further demonstrate that the proposed structure-conditioned generation preserves vessel-like anatomical structures while rendering target-domain OCTA appearance.

Table 11: Robustness to automatically generated target masks on OCT-to-OCTA. Mask Dice is measured against official annotations and reported as mean ± standard deviation. Translation metrics are reported as means. PSNR is in dB; Dice and SSIM are in %.

Target masks	**Mask Dice**		**Whole Image**		**Vessel**		**FAZ**	
	Vessel	**FAZ**	**PSNR ↑**	**SSIM ↑**	**PSNR ↑**	**SSIM ↑**	**PSNR ↑**	**SSIM ↑**
Official	100.00	100.00	19.05	64.65	18.92	75.72	20.48	35.05
Automatic	81.65 ± 3.10	80.34 ± 13.30	19.03	64.39	18.95	75.34	20.42	35.19

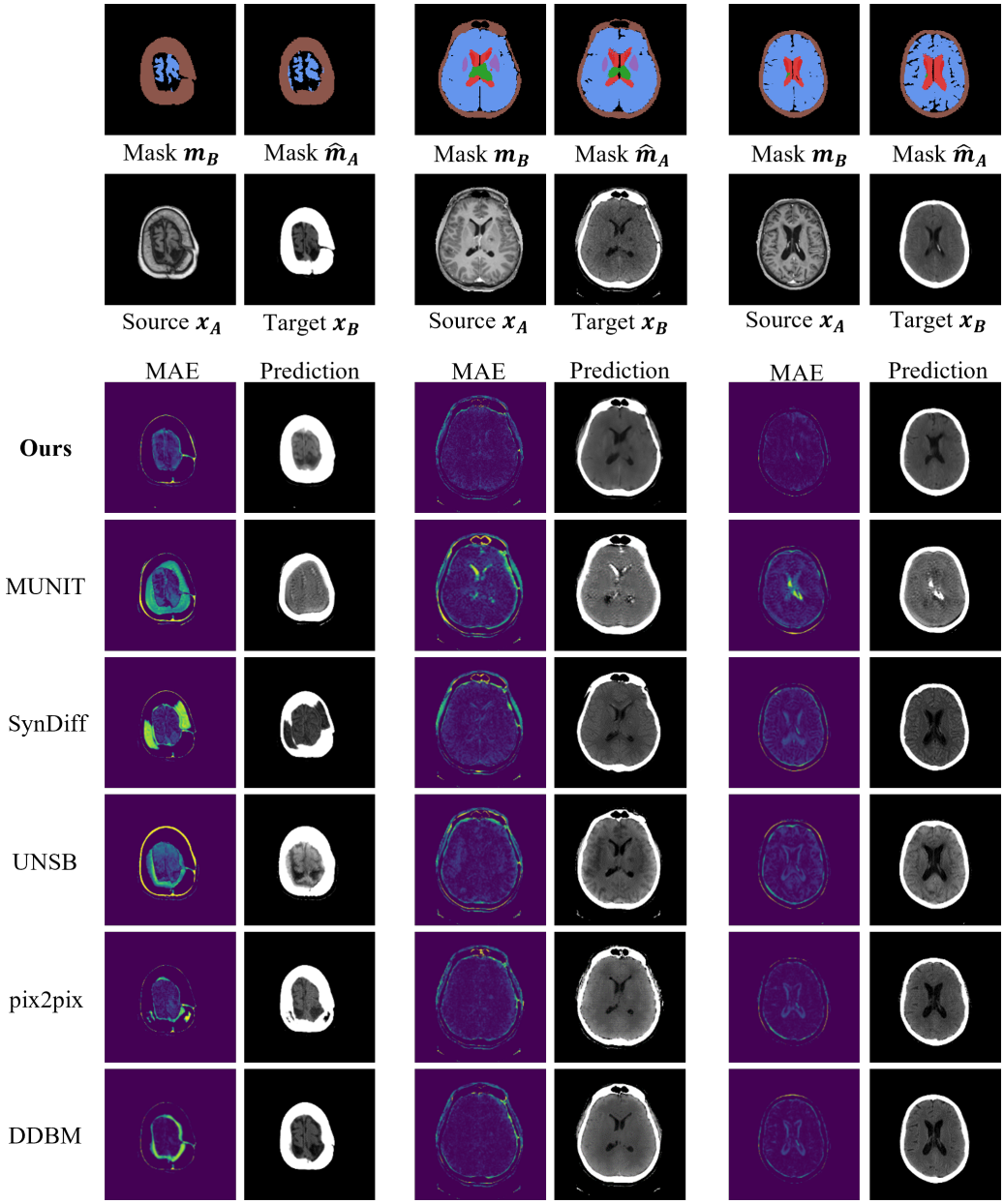


Figure 9: Additional qualitative results for MR-to-CT translation. The figure presents representative examples with structural annotations, source MR images, target-domain CT images, intermediate response maps, and generated CT results. These examples further show that the proposed structure-conditioned generation preserves anatomical structure while producing CT-like appearance across different brain slices.

Table 12: Post-hoc analysis of subject-specific fidelity and prototype affinity on OCT-to-OCTA. All methods are evaluated against the same retrieved training prototypes for each case. Subject Margin is Own-subject Dice minus Prototype Affinity. Values are mean ± 95% case-level bootstrap CI half-width. Bold indicates the best mean among translation methods for Own-subject Dice and Subject Margin.

Subset	Image / Method	Own-subject Dice ↑	Prototype Affinity	Subject Margin ↑
All 50	Acquired OCTA	0.859 ± 0.008	0.124 ± 0.004	0.735 ± 0.009
	MuTri	0.787 ± 0.017	0.112 ± 0.003	0.675 ± 0.017
	Ours	$\mathbf{0.797} \pm 0.026$	0.118 ± 0.004	$\mathbf{0.679} \pm 0.025$
Lowest retrieval-confidence 10	Acquired OCTA	0.867 ± 0.016	0.111 ± 0.006	0.755 ± 0.021
	MuTri	0.810 ± 0.039	0.104 ± 0.005	0.706 ± 0.040
	Ours	$\mathbf{0.827} \pm 0.038$	0.107 ± 0.004	$\mathbf{0.720} \pm 0.041$

Table 13: Blinded reader study on 10 held-out OCT-to-OCTA cases with five readers. Fidelity and Acceptability denote clinical vascular fidelity and diagnostic acceptability; Missing and Unsupported denote missing-vessel and unsupported-vessel severity. Values are mean ± 95% case-level bootstrap CI half-width. Bold indicates the best mean.

Method	Fidelity ↑	Missing ↓	Unsupported ↓	Acceptability ↑
SynDiff	2.74 ± 0.34	1.74 ± 0.28	1.42 ± 0.24	2.64 ± 0.39
MuTri	3.00 ± 0.17	1.64 ± 0.15	1.18 ± 0.16	2.94 ± 0.21
Ours	$\mathbf{3.38} \pm 0.60$	$\mathbf{1.42} \pm 0.49$	$\mathbf{0.88} \pm 0.29$	$\mathbf{3.38} \pm 0.75$

hoc. For each held-out source image, we fix the retrieved training prototypes and their weights across all compared methods. Using the frozen vessel evaluator, we compute Own-subject Dice against the subject-specific annotation. Prototype Affinity is the retrieval-weighted mean Dice between the predicted vessel mask and the individual retrieved masks. Subject Margin is Own-subject Dice minus Prototype Affinity. Acquired OCTA provides a reference, while MuTri provides a comparison method that does not use the memory bank.

Tab. 12 reports results for all 50 held-out cases and the ten cases with the lowest retrieval confidence. Here, confidence measures source-to-bank similarity. Our method achieves higher mean Own-subject Dice than MuTri, with similar Prototype Affinity and a slightly larger mean Subject Margin. The same pattern holds in the lowest-confidence subset. These results are consistent with preservation of subject-specific structure at the global level, although global overlap measures cannot exclude local retrieval-induced hallucination or omission.

H Blinded reader study

We conduct a blinded reader study on 10 held-out OCT-to-OCTA cases with five readers, including three retina specialists with 2–10 years of experience interpreting OCT/OCTA. Each reader independently assesses anonymized outputs from SynDiff, MuTri, and our method relative to the same source OCT and acquired OCTA. Clinical vascular fidelity and diagnostic acceptability are scored from 1 to 5, with higher scores indicating better quality. Missing-vessel and unsupported-vessel severity are scored from 0 to 3, with lower scores indicating fewer perceived structural errors. We report mean scores with 95% case-level bootstrap confidence intervals, retaining all five reader ratings within each resampled case.

As shown in Tab. 13, our method achieves the highest mean vascular fidelity and diagnostic acceptability, together with the lowest mean missing-vessel and unsupported-vessel severity among the evaluated methods. These results complement the independent vessel evaluation in Tab. 6 by providing reader-based evidence of improved structural fidelity. The small cohort limits the generalizability of these findings.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See Sec.3 and Sec.4.1

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Appendix.D

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not present theoretical results or formal proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Appendix B and C

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code, pretrained weights, and instructions for reproducing the main experiments are publicly available in our GitHub repository.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See Sec. 4 and Appendix. C

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All quantitative results are reported as mean $\pm$ 95% confidence intervals.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Appendix. F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics. We use de-identified medical imaging data and do not release identifiable patient information or make clinical deployment claims.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Appendix. E.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We plan to release the code and trained weights for research use only. The release will contain no private or identifiable patient data and will not be intended for clinical diagnosis or decision-making. Users will be required to follow the licenses of the original public datasets, and generated images should be treated as synthetic research outputs.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use the code and models from Syndiff and UNSB and properly cite them. Datasets are also cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Documentation for the released code and pretrained weights is provided in the public repository linked above.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: The reader-study protocol and scoring scales are described in Appendix H, and readers received no compensation. The full text of the participant instructions is not included.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper includes a blinded reader study, but does not report an IRB approval or an institutional exemption determination for this study.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We only use LLMs for writing, editing, or formatting purposes.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.