Enhancing Retinal Vessel Segmentation Generalization via Layout-Aware Generative Modelling

Anonymous Author(s)
Affiliation
Address
email

Abstract

Generalization in medical segmentation models is challenging due to limited annotated datasets and imaging variability. To address this, we propose **R**etinal **L**ayout-**A**ware **D**iffusion (RLAD), a novel diffusion-based framework for generating controllable layout-aware images. RLAD conditions image generation on multiple key layout components extracted from real images, ensuring high structural fidelity while enabling diversity in other components. Applied to retinal fundus imaging, we augmented the training datasets by synthesizing paired retinal images and vessel segmentations conditioned on extracted blood vessels from real images, while varying other layout components such as lesions and the optic disc. Experiments demonstrated that RLAD-generated data improved generalization in retinal vessel segmentation by up to 8.1%. Furthermore, we present **REYIA**, a comprehensive dataset comprising 585 manually segmented retinal images. We make the REYIA dataset and our source code open (upon publication) .

1 Introduction

Deep learning has achieved remarkable success across various domains, but its progress often depends on access to large annotated datasets. In fields such as natural language processing, vision-language modeling, and image generation, synthetic data from large models has driven significant advancements [1–6]. However, in medical imaging, particularly retinal vessel segmentation, data scarcity and variability in imaging conditions remain persistent limitations [7–10]. Retinal vessel segmentation is critical for the diagnosis of ocular and systemic diseases [11–14], yet the creation of annotated datasets demands a considerable amount of time, specialized expertise, and consistency across imaging devices [15].

Retinal vessel segmentation involves two tasks: general vessel segmentation, which identifies the vasculature, and artery/vein (AV) segmentation, which also differentiates arteries from veins. This distinction provides insights into vessel-specific pathologies[16, 17]. However, AV segmentation requires complex annotations, making it challenging to obtain sufficient labeled data for robust training.

Generative models like GANs and VAEs have been explored to address data scarcity in medical imaging [18, 19]. When applied to retinal images, these models often encounter challenges, including difficulties in preserving anatomical fidelity and issues with training stability [20]. Diffusion models have recently emerged as powerful tools for generating diverse high-fidelity images, with superior stability and detail preservation, compared to GANs and VAEs [21, 22]. Despite their success in image synthesis tasks across domains, e.g., natural image generation and text-to-image modeling, their application in medical imaging has largely focused on generating synthetic images rather than directly enhancing segmentation performance through data augmentation.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

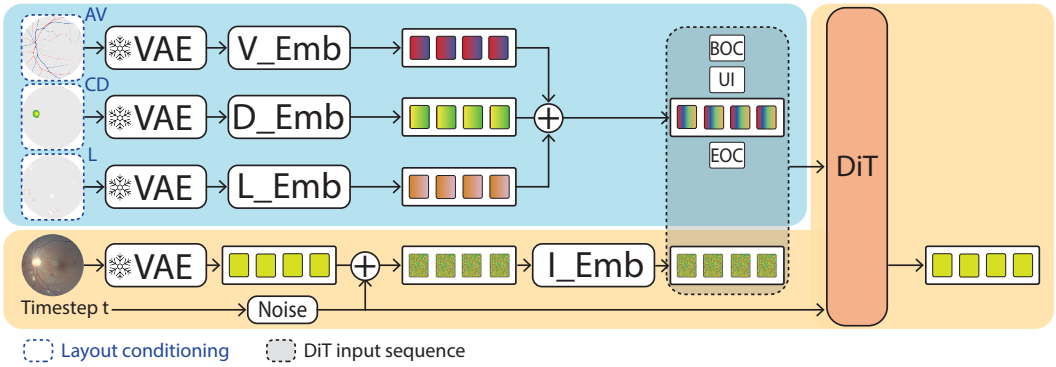

Figure 1: **RLAD Architecture.** The original fundus image and segmentation maps for artery/vein (AV), the optic cup/disc (CD), and lesions (L) are encoded into latent representations using a frozen VAE. Gaussian noise is added to the image latent, and each latent (image, CD, AV, and L) is projected into the DiT [25] input space via distinct projections. Condition embeddings for AV, CD, and L are summed into a single embedding, c. The DiT input consists of a beginning-of-conditioning (BOC) token, user input (UI), c, an end-of-conditioning (EOC) token, and the noised image latent. The DiT outputs the corresponding denoised image latent. The UI token specifies whether a layout component is guided by user input or defaults to a neutral embedding when absent.

To address these limitations, we propose Retinal Layout-Aware Diffusion **(RLAD)**, a diffusion-based framework for the controllable generation of synthetic retinal images. By conditioning on multiple key retinal structures—such as artery/vein (AV), the optic cup/disc (CD), and lesions (L)—RLAD preserves essential vascular layouts while introducing variability in other regions. This enables the creation of paired image-segmentation maps that expand training datasets without compromising structural integrity. Synthetic data generated by RLAD improve segmentation model robustness across diverse imaging conditions and acquisition settings.

We evaluated RLAD-generated data using state-of-the-art visual encoders such as Vision Transformers [23] and Swin Transformers [24], and demonstrate consistent improvements in generalization performance under distribution shifts (up to 8.1%). Additionally, we introduce **REYIA**, the largest multi-source collection of 585 retinal images with human reference AV segmentation, which not only complements our synthetic data but also demonstrates strong baseline performance, further validating the effectiveness of our synthetic data. In summary, the main contributions of this work are:

- A novel multi-layout-aware generative model (**RLAD**) that synthesizes diverse yet anatomically accurate retinal images while preserving semantic structures.
- Demonstrating consistent segmentation performance improvements across state-of-the-art architectures using RLAD-generated data.
- Introducing **REYIA**, the largest multi-source collection of datasets for AV-segmented retinal fundus images.

2 Related Work

Retinal AV segmentation plays a critical role in diagnosing microvascular pathologies [26–30]. Early methods [8, 31–34], such as Little W-Net [7], focused on compact convolutional neural networks to reduce computational complexity. More recently, LUNet achieved state-of-the-art performance on optic disc-centered images but struggled to generalize to macula-centered images [9]. This underscores the primary challenge of achieving robust generalization across diverse retinal imaging conditions.

Generative adversarial networks have been extensively used for retinal image synthesis, often conditioning the generation process on features such as vessel or lesion masks [35, 36]. While these methods produced visually realistic images, they frequently lacked anatomical accuracy and robustness [20], limiting their effectiveness for downstream tasks like AV segmentation. To address these issues, Go et al. [20] proposed a hybrid approach that combined a diffusion model for generating AV masks with a conditional GAN for synthesizing retinal images. Their method preserved patient privacy and demonstrated that synthetic images could lead to AV segmentation performance comparable to models trained on real data. However, it failed to further enhance AV segmentation performance

further, possibly due to limited variability in the generated AV masks, which may have propagated to the synthesized images.

Diffusion models have demonstrated remarkable generative capabilities across various domains, including image synthesis, video generation, layout and 3D modeling [1, 21, 37–43]. Recent advancements, such as classifier-free guidance [44] enable precise control over conditioning signals during generation, making these models well-suited for structured image synthesis tasks. Transformer-based architectures such as DiT [25] further enhance performance by capturing long-range dependencies.

Building on these developments, we propose a multi-layout-aware diffusion framework specifically designed for retinal fundus image synthesis. Unlike prior approaches, our method conditions generation on multiple retinal layout components —AV, CD, and L—extracted from real, non-annotated images using pretrained segmentation models. This minimizes error propagation and enhances realism while addressing domain generalization challenges in AV segmentation tasks through synthetic data augmentation.

3 Datasets

This section introduces the new datasets created for this study and provides an overview of the datasets used for diffusion model training and downstream segmentation tasks. For additional details, please refer to the appendix.

3.1 New Datasets

We introduce REYIA, a curated set of 585 retinal fundus images annotated with AV blood vessel segmentations using the open-access Lirot.ai software [15] and summarized in Table 1. To enhance diversity, REYIA includes manually segmented images as part of this research from nine datasets: FIVES [45], TREND [46], GRAPE [47], MESSIDOR [48], MAGRABIA [49], PAPILA [50], MBRSET [51] AV-WIDE [52] and ENRICH. ENRICH is a new dataset collected for this study, consisting of 111 retinal fundus images (IRB S60649). AV-WIDE, which initially contained only skeletonized vessels, was reannotated to include complete vessel segmentations.

Dataset	**# Samples**	**Image Center**	**FOV** (°)	Region	Resolution (px)
GRAPE† [47]	81	M	50	China	1444x1444
MESSIDOR† [48]	67	M	45	France	1444x1444
PAPILA† [50]	78	D	30	Spain	1444x1444
MAGHREBIA† [49]	69	M, D	30	Maghreb	1444x1444
ENRICH*	111	D	45	Belgium	1958x2196
FIVES† [45]	75	M	45	China	1444x1444
AV-WIDE† [52]	26	D	Ultra wide	USA	829x1531
TREND † [46]	48	M	30	Montenegro	2560x2560
MBRSET† [51]	30	M	30	Brazil	1444x1444

Table 1: **REYIA datasets collection** released with this work. Datasets marked with † were annotated specifically for this work, and those marked with * were both introduced and annotated here.

3.2 Diffusion Model Datasets

To train RLAD, we curated 112,320 retinal fundus images from publicly available datasets spanning diverse imaging conditions, fields of view (FOV), and pathologies. The sources include widely used datasets: UZLF [53], GRAPE [47], MESSIDOR [48], PAPILA [50], MAGRABIA [49], ENRICH, 1000images [54], DDR [55], EYEPACS [56], G1020 [57], IDRID [58] and ODIR [59]. Evaluation of the realism of the generated images, in comparison to real images, was performed on the DRTiD dataset [60].

3.3 AV Segmentation Datasets

3.3.1 Datasets for Segmentation Model Training

To train our segmentation models, we constructed a composite dataset combining the UZLF dataset with newly annotated versions of GRAPE, MESSIDOR, ENRICH, MAGRABIA, and PAPILA. These datasets feature high-resolution retinal fundus images with FOVs ranging from 30° to 45° and encompass a variety of ophthalmic conditions and patient populations.

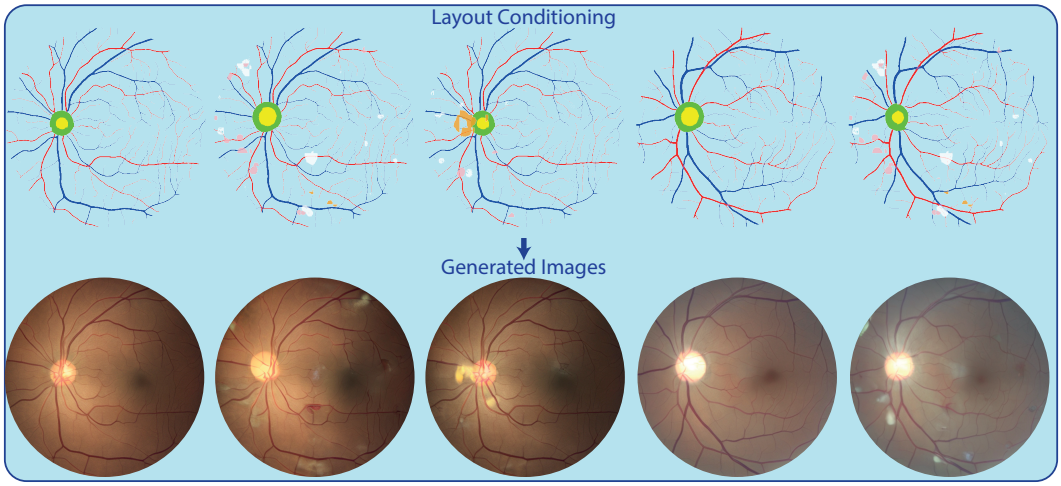

Figure 2: **Retinal Layout-Aware Diffusion Qualitative Examples.** Top: user-defined layout components inputs (artery/vein in red/blue, optic disc/cup in green/yellow, and lesions in white/pink/orange). Bottom: corresponding generated fundus images.

3.3.2 Datasets for Segmentation Model Evaluation

To assess generalization performance under varying levels of distribution shift, we evaluated our segmentation models across three categories of datasets:

In-Domain (Local): Data collected from the same hospital under similar acquisition conditions to those as one of the training datasets, ensuring minimal distribution shifts.

Near-Domain (External): Data from different hospitals and environment, introducing moderate distribution shifts. This category includes HRF [61], INSPIRE [9, 62], UNAF [9, 63] and the reannotated FIVES dataset.

Out-of-Domain (OOD): Data that significantly differ from the training distribution, used to evaluate the model robustness across diverse imaging conditions. It includes AV-WIDE for ultra-wide-angle images, IOSTAR [64] for laser-based images, DRIVE [65, 66] for low-resolution images, RVD [10] for video frames from handheld devices, TREND and MBRSET for handheld device images.

4 Method

Our objective is to generate realistic retinal images based on key retinal layout components, specifically AV, CD, and L, extracted from real retinal fundus images.

4.1 Layout Extraction

We extract retinal layouts using open-source models for L segmentation [67] and CD segmentation [17, 68]. For AV segmentation, we retrained a SwinV2$_{\text{tiny}}$-based model on our annotated datasets with data augmentation techniques such as random color jitter, flips, and rotations. These extracted retinal layout components serve as input to the diffusion process. The impact of the layout extractor used is further discussed in the appendix.

4.2 Retinal Layout-Aware Diffusion

Our approach builds upon latent diffusion [69] and DiT [25]. The forward diffusion process [21, 37] gradually adds Gaussian noise to an image x_0, producing x_t. This process is defined as:

$$q(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\overline{\alpha}_t}x_0, (1 - \overline{\alpha}_t)I), \tag{1}$$

where the noise schedule $\{\overline{\alpha}_t\}$ follows a linear strategy as explored in [21]. The reverse process approximates the denoising steps to reconstruct x_0:

$$p_\theta(x_{t-1} \mid x_t, c) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, c), \Sigma_\theta(x_t, c)), \tag{2}$$

where c denotes conditioning information. Instead of operating directly in pixel space, we adopt latent diffusion and perform these operations in a compressed latent space of a frozen VAE. This

allows us to refine latent representations z_t iteratively towards z_0, improving computational efficiency and scalability.

To incorporate conditional information into the diffusion process, we extract the layout components (AV, CD and L) from the input data. These components are embedded into the transformer's latent space using dedicated projection heads: V_{emb}, D_{emb} and L_{emb}.

$$c_{\text{AV}} = V_{\text{emb}}\big(\text{AV}\big), \quad c_{\text{CD}} = D_{\text{emb}}\big(\text{CD}\big), \quad c_{\text{L}} = L_{\text{emb}}\big(\text{L}\big).$$

To handle both fully and partially conditional cases, we used user input (UI) tokens. Each token indicates whether a component is user-defined (guided) or neutral (unconditional). During training, each layout component is either provided or masked with a certain probability, allowing the model to learn both conditional and unconditional scenarios. This probabilistic masking is applied independently to each component. When a component is masked, it is replaced with a "black" image embedding, and its corresponding UI token is updated to signal the absence of guidance:

$$\text{UI} = [\text{UI}_{\text{AV}}, \text{UI}_{\text{CD}}, \text{UI}_{\text{L}}],$$

allowing flexible control over the conditioning process. The final conditioning vector is computed as:

$$c = c_{\text{AV}} + c_{\text{CD}} + c_{\text{L}}.$$

which is fed into the transformer as part of a sequence:

$$[\text{BOC}, \text{UI}, c, \text{EOC}, z_t],$$

where BOC and EOC mark the beginning and end of the conditioning tokens, respectively. After the transformer processes this sequence, only the image tokens are retained to produce z_{t-1}. This design ensures that conditioning signals guide the denoising process without remaining entangled in the final latent representation. A schematic overview of our architecture is provided in Figure 1.

Training Objective. Following DDPM [21], we adopt a noise prediction loss. Instead of directly modeling μ_θ and Σ_θ, our model predicts the noise ϵ added at a randomly chosen timestep t:

$$L_{\text{simple}} = \mathbb{E}_{z_0,t,\epsilon}\left[\|\epsilon - \hat{\epsilon}_\theta(z_t, t, c)\|^2\right]. \tag{3}$$

Minimizing this MSE loss enables the model to accurately denoise latent representations, effectively learning to reverse the diffusion process. By incorporating tokens that differentiate between user-defined and neutral embeddings for each layout component, the model can both generate anatomically guided images when specific conditions are provided, and produce diverse, unconstrained samples in the absence of such guidance. This flexibility ensures that the model adapts seamlessly to varying levels of conditional input, balancing anatomical fidelity with generative diversity.

Sampling. To generate new images, we start from a random Gaussian latent $z_T \sim \mathcal{N}(0, I)$ and iteratively remove noise at each diffusion step t. Our model predicts the added noise $\hat{\epsilon}_\theta(z_t, t, c)$, where c includes tokens for AV, CD, and L layouts.

We employ classifier-free guidance [44] to control how closely the model adheres to provided conditions. At each step, two predictions are made: one conditional (c) and one unconditional ($c = \emptyset$). These are combined as:

$$\hat{\epsilon}_\theta^{\text{guided}}(z_t, t, c) = \hat{\epsilon}_\theta(z_t, t, \emptyset) + w\big(\hat{\epsilon}_\theta(z_t, t, c) - \hat{\epsilon}_\theta(z_t, t, \emptyset)\big), \tag{4}$$

where w is a guidance scale. Higher w yields more faithful adherence to the conditions, lower w allows more diversity.

By iteratively applying guided noise predictions until reaching z_0, we decode z_0 using the VAE to produces a synthetic retinal fundus image. This approach balances anatomical fidelity when conditions are provided with greater diversity when they are neutral or absent. Examples of generated images are shown in Figure 2.

4.3 Backbone Pretraining

We investigate pretraining strategies to enhance segmentation performance, focusing on two key approaches: Masked Autoencoders (MAE) [74] and Windowed Contrastive Learning (WCL) [75].

Backbone	Local		External				OOD						Average	
	UZLF	LES-AV	HRF	INSPIRE	FIVES	UNAF	AV-WIDE	IOSTAR	DRIVE	RVD	TREND	MBRSET	External	OOD
RMHAS [8]	-	60.0	48.0	-	-	-	-	55.0	60.0	-	-	-	-	-
$RVD_{Swin\text{-}L}$ [10]	-	-	-	-	-	-	-	-	57.3	53.0	-	-	-	-
Little W-Net [7]	80.7	82.0	58.1	71.3	73.5	68.6	43.1	29.9	61.3	34.7	53.4	50.4	67.9	45.5
Automorph [34]	76.3	84.0†	77.4†	71.1	72.5	65.9	50.1	54.9	78.1†	34.1	66.6	63.7	71.7†	57.9†
VascX [70]	80.6	81.8	75.6	74.9	80.4	73.1	49.8	52.1	73.6	42.6	71.9	73.2	76.0	60.5
LUNet [9]	83.2	83.5	73.1	75.5	86.0	74.4	69.3	56.7	71.1	35.2	71.1	63.2	77.3	61.1
$DinoV2_{small}$ [71]	$81.6_{\pm 0.9}$	$82.4_{\pm 1.4}$	$74.2_{\pm 0.8}$	$76.6_{\pm 0.9}$	$82.7_{\pm 1.0}$	$72.9_{\pm 1.9}$	$59.4_{\pm 2.4}$	$57.2_{\pm 2.7}$	$75.0_{\pm 1.2}$	$45.4_{\pm 0.6}$	$67.1_{\pm 1.5}$	$79.6_{\pm 1.1}$	76.6	64.0
+ RLAD (Our)	$81.8_{\pm 0.9}$	$82.8_{\pm 1.3}$	$75.1_{\pm 0.8}$	$77.5_{\pm 0.7}$	$83.6_{\pm 1.0}$	$73.7_{\pm 1.8}$	$58.3_{\pm 2.1}$	$65.3_{\pm 3.2}$	$76.8_{\pm 1.1}$	$46.7_{\pm 0.6}$	$70.8_{\pm 1.5}$	$81.9_{\pm 1.5}$	77.5	66.6
Δ	+0.2	+0.4	+0.9	+0.9	+1.1	+0.8	-1.1	+8.1	+1.8	+1.3	+3.7	+2.3	+0.9	+2.6
RETFound [72]	$81.2_{\pm 1.0}$	$82.3_{\pm 1.5}$	$77.7_{\pm 1.1}$	$75.8_{\pm 0.9}$	$82.1_{\pm 1.0}$	$71.8_{\pm 1.9}$	$63.2_{\pm 1.9}$	$63.0_{\pm 3.3}$	$75.1_{\pm 1.2}$	$42.5_{\pm 0.8}$	$70.1_{\pm 1.4}$	$78.4_{\pm 1.7}$	76.9	65.2
+ RLAD (Our)	$83.1_{\pm 1.0}$	$83.6_{\pm 1.5}$	$80.2_{\pm 1.6}$	$78.4_{\pm 1.0}$	$86.3_{\pm 0.9}$	$74.6_{\pm 1.9}$	$69.5_{\pm 1.8}$	$70.5_{\pm 3.0}$	$77.1_{\pm 1.2}$	$46.4_{\pm 0.8}$	$76.9_{\pm 1.4}$	$79.1_{\pm 1.7}$	79.9	69.9
Δ	+0.9	+1.3	+2.5	+2.6	+4.2	+2.8	+6.3	+7.5	+2.0	+3.9	+6.8	+0.7	+3.0	+4.7
$SwinV2_{tiny}$ [73]	$82.8_{\pm 0.8}$	$83.4_{\pm 1.4}$	$79.9_{\pm 1.4}$	$78.1_{\pm 0.9}$	$85.9_{\pm 0.8}$	$74.3_{\pm 1.9}$	$68.1_{\pm 2.0}$	$67.6_{\pm 3.1}$	$76.0_{\pm 1.1}$	$44.1_{\pm 0.8}$	$76.2_{\pm 1.4}$	$81.5_{\pm 2.7}$	79.6	68.9
+ RLAD (Our)	$83.0_{\pm 0.8}$	$83.6_{\pm 1.4}$	$80.2_{\pm 1.3}$	$78.3_{\pm 0.9}$	$86.3_{\pm 0.8}$	$74.6_{\pm 1.9}$	$69.5_{\pm 2.0}$	$71.3_{\pm 2.7}$	$77.1_{\pm 1.4}$	$46.3_{\pm 0.7}$	$77.1_{\pm 1.1}$	$83.7_{\pm 2.0}$	79.9	70.8
Δ	+0.2	+0.2	+0.3	+0.2	+0.4	+0.3	+1.4	+3.7	+1.1	+2.2	+1.1	+2.0	+0.3	+1.9
$SwinV2_{large}$ [73]	$83.2_{\pm 0.9}$	$83.6_{\pm 1.4}$	$80.4_{\pm 1.3}$	$79.0_{\pm 0.9}$	$87.2_{\pm 0.8}$	$75.5_{\pm 1.7}$	$70.9_{\pm 2.1}$	$73.5_{\pm 3.1}$	$76.5_{\pm 1.1}$	$48.2_{\pm 0.7}$	$77.4_{\pm 1.4}$	$86.0_{\pm 1.6}$	80.5	72.1
+ RLAD (Our)	**83.2**$_{\pm 0.9}$	**83.6**$_{\pm 1.5}$	**80.4**$_{\pm 1.3}$	**79.1**$_{\pm 0.9}$	**87.3**$_{\pm 0.8}$	**75.8**$_{\pm 1.7}$	**71.2**$_{\pm 2.2}$	**74.5**$_{\pm 2.8}$	**77.1**$_{\pm 1.0}$	**48.2**$_{\pm 0.7}$	**77.6**$_{\pm 1.4}$	**86.2**$_{\pm 1.6}$	**80.7**	**72.5**
Δ	+0.0	+0.0	+0.0	+0.1	+0.1	+0.3	+0.3	+1.0	+0.6	+0.0	+0.2	+0.2	+0.2	+0.4

Table 2: **RLAD Results.** Quantitative comparison of RLAD-generated data integrated into DinoV2, RETFound, and SwinV2 across model sizes. Baselines are trained on datasets from section 3.3. Evaluation spans Local, External, and OOD benchmarks, with average performance for External and OOD. Previous state-of-the-art performance (gray) reflects open-source inference or reported results. Performance is the average Dice score for artery and vein. † indicates data leakage during training.

MAE facilitates robust representation learning by reconstructing masked inputs, effectively teaching the model to predict missing portions of an image. WCL, initially designed for depth estimation, employs contrastive learning on small image patches while maintaining local spatial relationships, making it particularly suitable for semantic segmentation tasks. Furthermore, we explore multi-objective pretraining [76–78], by combining MAE and WCL to develop richer representations and improve downstream task performance. The dataset used for pretraining aligns with the one employed to train RLAD.

4.4 Enhancing AV Segmentation with RLAD

The synthetic images generated by RLAD serve as powerful data augmentation tools for vessel segmentation models. By preserving vascular structures while varying other characteristics (e.g., disc or lesions), these images enrich training datasets without requiring additional manual annotations.

Let a vessel segmentation model be denoted as $\mathcal{S}$, trained on real retinal images x_{orig} with ground truth AV annotations y. The segmentation loss combines Dice loss and Binary Cross-Entropy (BCE) where L^{A} and L^{V} specifically represent the loss terms computed over artery and vein, respectively:

$$L_{\text{seg}} = 0.5 \cdot (L^{\text{A}}_{\text{Dice}} + L^{\text{A}}_{\text{BCE}}) + 0.5 \cdot (L^{\text{V}}_{\text{Dice}} + L^{\text{V}}_{\text{BCE}}). \tag{5}$$

The total training objective includes supervised loss on real images and consistency loss on synthetic images:

$$L_{\text{total}} = L_{\text{seg}}(\mathcal{S}(x_{\text{orig}}), y) \; + \; \lambda \cdot L_{\text{seg}}(\mathcal{S}(x_{\text{gen}}), y), \tag{6}$$

where x_{gen} is a synthetic image sharing vascular structure with x_{orig}, and $\lambda > 0$ balances contributions from real and synthetic data. This consistency regularization improves robustness across diverse imaging conditions, enhancing segmentation performance on unseen datasets.

Additional implementation details, including hyperparameters and optimization strategies, are provided in the appendix.

5 Experimental Setup

We address data scarcity in retinal vessel segmentation by evaluating RLAD's ability to generate controllable, realistic fundus images and improve AV segmentation performance. Key evaluations include image realism (section 5.2), segmentation performance across backbones (section 5.3), SOTA comparisons (section 5.4), and ablation studies (section 6). We seek to address three key research questions:

- Can RLAD generate controllable, realistic retinal images?
- Does usage of RLAD-generated data enhance our AV segmentation model?
- How does our model perform compared to SOTA?

5.1 Evaluation Metrics

We evaluate the diffusion model's performance using the Fréchet Distance (FD), which compares the feature distributions of real and generated images. We compute it in the latent space of Inception-v3 (FID) [79] and RETFound [72] (RET-FD), a foundation model pre-trained on 1.6 million retinal images. RETFound likely offers a more accurate representation of retinal image-specific features, while Inception-v3 enables a comparison with previous work.

For AV segmentation, we use the Dice score to measure overlap between predicted and ground truth segmentations, averaged as $(\text{Dice}_A + \text{Dice}_V)/2$. This is complemented by the Intersection over Union (IoU) and centerline Dice (clDice) [80], which emphasizes vessel centerlines. Both Dice and clDice metrics are employed in RLAD ablation studies, with additional IoU and clDice results provided in the appendix. Notably, clDice offers a more nuanced evaluation by balancing sensitivity to both thin and large vessels.

5.2 Evaluation of Realism

We compare the FID scores achieved by RLAD with those of prior works (Table 3), using their publicly available models for image generation or reports their published results when the models were inaccessible. Notably, RLAD demonstrates superior performance by generating more realistic retinal fundus images, as evidenced by lower FID and RET-FD scores.

5.3 Integrating RLAD into Leading Backbones

In Table 2, we present the performance of RLAD-generated data on the AV segmentation task, evaluated using various backbones: $\text{DinoV2}_{\text{small}}$, RETFound, $\text{SwinV2}_{\text{tiny}}$, and $\text{SwinV2}_{\text{large}}$. The results are reported across Local, External, and OOD test sets. For comparison, the first rows include previously published state-of-the-art results under similar settings (i.e., Local, External, and OOD), where available.

RLAD consistently improves performance on External, and OOD test sets, demonstrating its backbone-agnostic advantages and its adaptability to in-domain and out-of-domain pretrained models. For example, integrating RLAD with RETFound yields performance improvements of 6.3%, 7.5%, and 6.8% on AV-WIDE, IOSTAR, and TREND, respectively. Notably, even when applied to the top-performing backbone, $\text{SwinV2}_{\text{large}}$, RLAD provides further performance gains of 0.2% on External and 0.4% in OOD datasets.

5.4 Segmentation performance vs SOTA

$\text{SwinV2}_{\text{large}}$, trained on our newly curated dataset and RLAD-generated data, surpasses previous state-of-the-art models across all Local, External, and OOD datasets, with the exception of RVD (Table 2). As illustrated in Figure 3, it demonstrates superior AV segmentation performance compared to $\text{SwinV2}_{\text{large}}$ trained solely on the UZLF dataset and LUNet, the best performing open-source

Gen Model	Conditioning	FID↓	RET-FD↓
StyleGAN [81]	L	138.0	120.8
StyleGAN2 [82]	Demographics	98.1	116.0
StyleGAN2 [20]†	AV	122.8	-
Pix2PixHD [20]†	AV	86.8	-
RLAD (Our)	AV + L + CD	**30.3**	**79.7**

Table 3: **Realism of Generated Images.** Lower FID and RET-FD on the DRTiD dataset indicate closer alignment with real data, reflecting realism. Notably, RLAD is able to generate controllable and more realistic retinal images. Models† trained and evaluated on private data.

model. Further quantitative and qualitative comparisons are included in the appendix. Moreover, a comprehensive analysis demonstrating the superiority of our model over previous state-of-the-art methods in estimating common vascular parameters is also provided in the appendix.

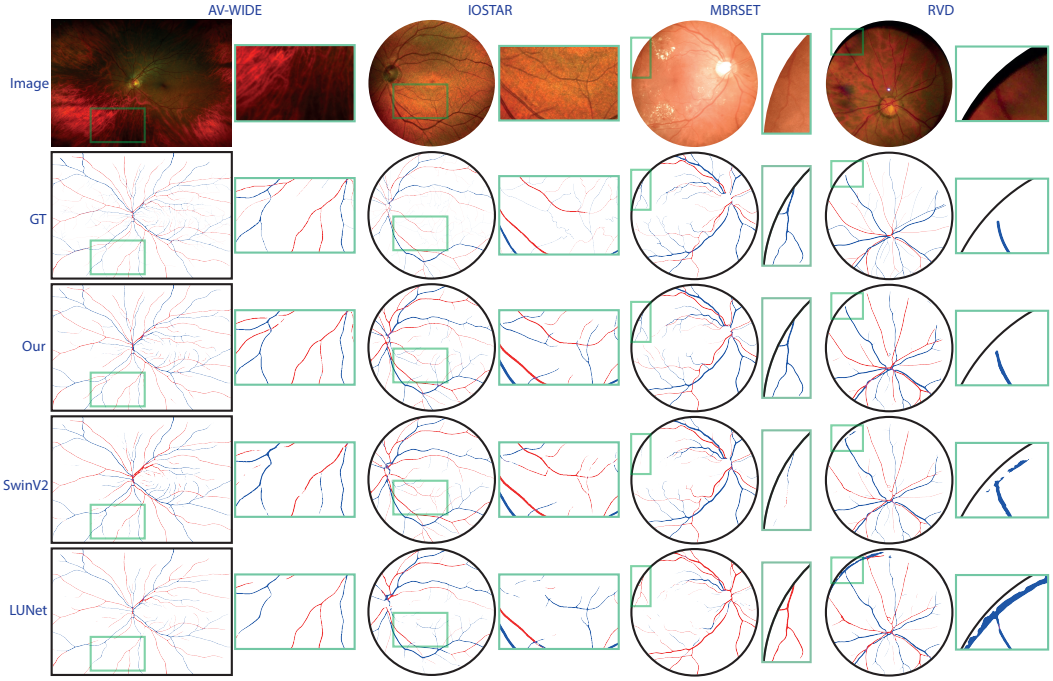

Figure 3: **Qualitative Example on the Segmentation Downstream Task.** Comparing our model's AV segmentation to a SwinV2$_{\text{Large}}$ [24] trained on the UZLF dataset and LUNet [9], a SOTA model, showcasing its superior performance across fundus images from various datasets.

6 Ablation studies

We analyze the effects of RLAD's components, training datasets, and pretraining objectives using SwinV2$_{\text{tiny}}$ as the baseline and Dice score unless stated otherwise.

Training Datasets: Starting with the UZLF dataset, we incrementally added our newly introduced datasets (Table 4). The Local test sets includes optic disc centered images, while External test sets mix optic disc and macula centered images. Adding macula-centered datasets GRAPE and MESSIDOR improved performance across Local, External and OOD test sets. Each dataset addition yielded incremental gains, with final improvements of +1.1%, +4.1%, and +8.3% for Local, External, and OOD, respectively.

Pretraining Objective: We evaluated how pretraining objectives (MAE, WCL, or both) influence our model's performance (see Table 5). Adding MAE or WCL individually improved the OOD Dice score from 68.9% to 69.2% and 69.4%, respectively, while combining them further increased clDice. These findings indicate that combining both strategies enhance model generalization.

Conditioning on multiple layout components: When learning a conditional distribution solely on AV, SwinV2$_{\text{tiny+RLAD}}$ achieved an average Dice score of 70.4% on the OOD datasets. In contrast, conditioning on multiple layout components (AV, CD, and L) improved performance to 70.8%. This highlights the advantage of leveraging a broader range of retinal fundus image features to enhance the learned distribution (see Table 5).

Varying Generated Data Quantity: We explored the impact of varying amounts of RLAD-generated samples: 0.5K (1 per real image), 1.5K (3 per real image), and 7.2K (15 per real image). Increasing generated samples improved the average OOD Dice (Table 6) and clDice (see appendix).

Performance Gains of RLAD Relative to Dataset Size: Figure 4 shows learning curves on OOD datasets for SwinV2$_{\text{tiny}}$ trained with and without RLAD synthetic data. Incorporating RLAD-generated data consistently improves performance across all datasets. For IOSTAR, RVD, DRIVE,

Datasets	Size	Local	External	OOD
UZLF [53]	184	82.1	75.5	60.6
+ GRAPE (Our†)	81	82.6	78.1	65.2
+ MESSIDOR (Our†)	67	82.8	78.9	66.6
+ ENRICH (Our*)	111	83.1	79.2	67.0
+ MAGRABIA (Our†)	69	83.1	79.2	67.2
+ PAPILA (Our†)	78	**83.1**	**79.6**	**68.9**
Δ		+1.0	+4.1	+8.3

Table 4: **Impact of increasing the number of training datasets.** This table shows how adding newly introduced (*) or annotated (†) datasets to the SwinV2$_{\text{tiny}}$ training pipeline impact performance.

PT		FT	Local		External		OOD	
MAE	WCL	Gen	Dice	clDice	Dice	clDice	Dice	clDice
✗	✗	✗	83.1	83.6	79.6	80.7	68.9	68.8
✓	✗	✗	83.1	83.6	79.6	80.8	69.4	69.2
✗	✓	✗	83.2	83.6	79.7	80.8	69.2	69.1
✓	✓	✗	83.2	83.6	79.6	80.8	69.4	69.3
✓	✓	AV	83.3	83.7	79.9	81.1	70.4	70.5
✓	✓	AV + CD + L	**83.3**	**83.7**	**79.9**	**81.1**	**70.8**	**71.1**
Δ			+0.2	+0.1	+0.3	+0.4	+1.9	+2.3

Table 5: **Pretraining Objective and Generation Method.** The top section shows baseline performance on our dataset, the middle highlights the impact of pretraining objectives, and the bottom examines AV conditioning versus AV + CD + L, with notable OOD improvements using AV + CD + L.

# Gen	AV-WIDE	IOSTAR	DRIVE	RVD	TREND	MBRSET	OOD
0.5K	69.2	69.9	77.2	45.8	76.9	75.9	70.4
1.5K	69.5	70.5	77.1	46.4	76.9	76.0	70.6
7.2K	69.5	71.3	77.1	46.3	77.1	76.2	**70.8**

Table 6: **Quantity of Generated Data.** We evaluate the impact of increasing RLAD's generated data on performance, reporting Dice scores for each OOD dataset and their average performance.

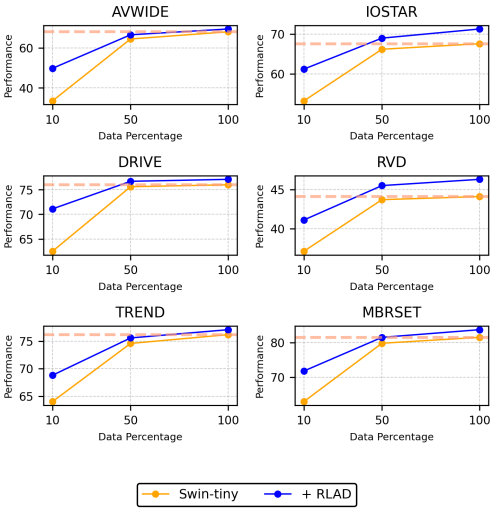

Figure 4: **RLAD Performance vs. Training Data Size.** The figure illustrates the learning curve of the SwinV2$_{\text{tiny}}$ [24] baseline on OOD datasets, demonstrating enhanced performance with RLAD-generated data. The data percentage reflects both real and generated samples, maintaining a 1:15 ratio (real:generated).

and MBRSET, the model trained with synthetic data outperformed the baseline while using less than 50% of the baseline's training data. The largest gains occurred in data-scarce scenarios, highlighting RLAD's effectiveness in enhancing performance.

7 Conclusion

This work presents RLAD, a novel diffusion-based framework designed to generate realistic and controllable retinal fundus images by conditioning on multiple layout components extracted from real-world data. Beyond image generation, RLAD proves to be a valuable tool for advancing downstream tasks. By incorporating the synthetic data generated by RLAD, we significantly enhance the training datasets for AV segmentation tasks, resulting in notable performance improvements across various visual backbones. This capability is particularly impactful in data-scarce scenarios, where access to comprehensive datasets is limited. Our findings highlight the potential of RLAD to drive innovation in medical imaging applications and improve segmentation outcomes. Future research could explore its application to other imaging modalities and investigate optimization strategies to further enhance its adaptability and scalability.

Limitations and Societal Impact: While RLAD improves generalization in retinal vessel segmentation, its effectiveness may be constrained by the quality of the generated images and the diversity of the training data. The approach may not fully generalize to imaging modalities or populations not represented in the training set. We demonstrated that the proposed framework may enhance clinical decision support for retinal image analysis, but care must be taken to avoid over-reliance on synthetic data and to monitor for biases that could affect underrepresented groups. Misapplication to non-target populations or imaging modalities could lead to incorrect diagnoses.

References

[1] Navve Wasserman, Noam Rotstein, Roy Ganz, and Ron Kimmel. Paint by inpaint: Learning to add image objects by removing them first, 2024. URL http://arxiv.org/abs/2404.18212.

[2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 36, 2024. URL https://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

[3] Yinheng Li, Rogerio Bonatti, Sara Abdali, Justin Wagle, and Kazuhito Koishida. Data generation using large language models for text classification: An Empirical case study, 2024. URL http://arxiv.org/abs/2407.12813.

[4] Hugging Face Team. SmolLM-corpus dataset, 2024. URL https://huggingface.co/datasets/HuggingFaceTB/smollm-corpus.

[5] Hugging Face Team. SmolLM models collection, 2024. URL https://huggingface.co/collections/HuggingFaceTB/smollm-models-6695016cad7167254ce15966.

[6] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023. URL http://arxiv.org/abs/2304.10592.

[7] Adrian Galdran, André Anjos, José Dolz, Hadi Chakor, Hervé Lombaert, and Ismail Ben Ayed. State-of-the-art retinal vessel segmentation with minimalistic models. *Scientific Reports*, 12(1):6174, 2022. doi: 10.1038/s41598-022-09675-y. URL https://www.nature.com/articles/s41598-022-09675-y. Publisher: Nature Publishing Group UK London.

[8] Danli Shi, Zhihong Lin, Wei Wang, Zachary Tan, Xianwen Shang, Xueli Zhang, Wei Meng, Zongyuan Ge, and Mingguang He. A deep learning system for fully automated retinal vessel measurement in high throughput image analysis. *Frontiers in Cardiovascular Medicine*, 9:823436, 2022. doi: 10.3389/fcvm.2022.823436. URL https://www.frontiersin.org/articles/10.3389/fcvm.2022.823436/full. Publisher: Frontiers Media SA.

[9] Jonathan Fhima, Jan Van Eijgen, Marie-Isaline Billen Moulin-Romsée, Heloı¨se Brackenier, Hana Kulenovic, Valérie Debeuf, Marie Vangilbergen, Moti Freiman, Ingeborg Stalmans, and Joachim A Behar. LUNet: deep learning for the segmentation of arterioles and venules in high resolution fundus images. *Physiological Measurement*, 45(5):055002, 2024. doi: 10.1088/1361-6579/ad3d28. URL https://iopscience.iop.org/article/10.1088/1361-6579/ad3d28. Publisher: IOP Publishing.

[10] MD Wahiduzzaman Khan, Hongwei Sheng, Hu Zhang, Heming Du, Sen Wang, Minas Coroneo, Farshid Hajati, Sahar Shariflou, Michael Kalloniatis, Jack Phu, and others. RVD: a handheld device-based fundus video dataset for retinal vessel segmentation. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, volume 36, pages 18203–18224, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3a71ee306d6991f2f87dd414e0bdf851-Paper-Datasets_and_Benchmarks.pdf.

[11] Tien Yin Wong, Ronald Klein, Barbara EK Klein, Stacy M Meuer, and Larry D Hubbard. Retinal vessel diameters and their associations with age and blood pressure. *Investigative ophthalmology & visual science*, 44(11):4644–4650, 2003. doi: 10.1167/iovs.03-0079. URL http://iovs.arvojournals.org/article.aspx?doi=10.1167/iovs.03-0079. Publisher: The Association for Research in Vision and Ophthalmology.

[12] Gerald Liew, Jie Jin Wang, Paul Mitchell, and Tien Y Wong. Retinal vascular imaging: a new tool in microvascular disease research. *Circulation: Cardiovascular Imaging*, 1(2):156–161, 2008. doi: 10.1161/CIRCIMAGING.108.784876. Publisher: Am Heart Assoc.

[13] Yu Huang, Carol Y Cheung, Dawei Li, Yih Chung Tham, Bin Sheng, Ching Yu Cheng, Ya Xing Wang, and Tien Yin Wong. AI-integrated ocular imaging for predicting cardiovascular disease: advancements and future outlook. *Eye*, 38(3):464–472, 2024. doi: 10.1038/s41433-023-02724-4. URL https://www.nature.com/articles/s41433-023-02724-4. Publisher: Nature Publishing Group UK London.

[14] Shawn Frost, Yogi Kanagasingam, Hamid Sohrabi, Janardhan Vignarajan, Pierrick Bourgeat, Oliver Salvado, Victor Villemagne, Christopher C Rowe, S Lance Macaulay, Cassandra Szoeke, and others. Retinal vascular biomarkers for early detection and monitoring of Alzheimer's disease. *Translational psychiatry*, 3(2):e233, 2013. Publisher: Nature Publishing Group.

[15] Jonathan Fhima, Jan Van Eijgen, Moti Freiman, Ingeborg Stalmans, and Joachim A Behar. Lirot. ai: a novel platform for crowd-sourcing retinal image segmentations. In *2022 computing in cardiology (CinC)*, volume 498, pages 1–4. IEEE, 2022. doi: 10.22489/CinC.2022.060. URL https://openreview.net/forum?id=fNUvJ9iTrL.

[16] José Ignacio Orlando, João Barbosa Breda, Karel Van Keer, Matthew B Blaschko, Pablo J Blanco, and Carlos A Bulant. Towards a glaucoma risk index based on simulated hemodynamics from fundus images. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 65–73. Springer, 2018. ISBN 978-3-030-00934-2. doi: 10.1007/978-3-030-00934-2_8.

[17] Jonathan Fhima, Jan Van Eijgen, Ingeborg Stalmans, Yevgeniy Men, Moti Freiman, and Joachim A Behar. PVBM: a Python vasculature biomarker toolbox based on retinal blood vessel segmentation. In *European conference on computer vision*, pages 296–312. Springer, 2022. doi: 10.1007/978-3-031-25066-8_15. URL https://link.springer.com/chapter/10.1007/978-3-031-25066-8_15.

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144, 2020. doi: 10.1145/3422622. URL https://dl.acm.org/doi/10.1145/3422622. Publisher: ACM New York, NY, USA.

[19] Diederik P Kingma. Auto-encoding variational bayes, 2013.

[20] Sojung Go, Younghoon Ji, Sang Jun Park, and Soochahn Lee. Generation of structurally realistic retinal fundus images with diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2335–2344, Seattle, WA, USA, 2024. doi: 10.1109/CVPRW63382.2024.00239. URL https://ieeexplore.ieee.org/document/10678363.

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 33:6840–6851, 2020. URL https://papers.nips.cc/paper_files/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

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

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International conference on learning representations (ICLR)*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, Montreal, QC, Canada, 2021. doi: 10.1109/ICCV48922.2021.00986. URL https://ieeexplore.ieee.org/document/9710580/.

[25] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4172–4182, 2023. doi: 10.1109/ICCV51070.2023.00387. URL https://ieeexplore.ieee.org/document/10377858.

[26] R M Gunn. Ophthalmoscopic evidence of (1) arterial changes associated with chronic renal disease, and (2) of increased arterial tension. *Transactions of the Ophthalmological Society of the United Kingdom*, 12: 124–125, 1892.

[27] Harold G Scheie. Evaluation of ophthalmoscopic changes of hypertension and arteriolar sclerosis. *AMA Arch. Ophthalmol.*, 49(2):117–138, 1953.

[28] Norman M Keith. Some differente types of essential hypertension: their course and prognosis. *Am. J. Med. Sci.*, 197:332–343, 1939.

[29] A Richey Sharrett, Larry D Hubbard, Lawton S Cooper, Paul D Sorlie, Rosemary J Brothers, F Javier Nieto, Joan L Pinsky, and Ronald Klein. Retinal arteriolar diameters and elevated blood pressure: the Atherosclerosis Risk in Communities Study. *Am J. Epidemiol.*, 150(3):263–270, 1999.

[30] Nicholas Witt, Tien Y Wong, Alun D Hughes, Nish Chaturvedi, Barbara E Klein, Richard Evans, Mary McNamara, Simon A McG Thom, and Ronald Klein. Abnormalities of retinal microvascular structure and risk of mortality from ischemic heart disease and stroke. *Hypertension*, 47(5):975–981, 2006. ISSN 0194-911X.

[31] Ruben Hemelings, Bart Elen, Ingeborg Stalmans, Karel Van Keer, Patrick De Boever, and Matthew B Blaschko. Artery–vein segmentation in fundus images using a fully convolutional network. *Computerized Medical Imaging and Graphics*, 76:101636, 2019.

[32] Jingfei Hu, Hua Wang, Zhaohui Cao, Guang Wu, Jost B Jonas, Ya Xing Wang, and Jicong Zhang. Automatic artery/vein classification using a vessel-constraint network for multicenter fundus images. *Frontiers in Cell and Developmental Biology*, page 1194, 2021.

[33] Yukun Zhou, Moucheng Xu, Yipeng Hu, Hongxiang Lin, Joseph Jacob, Pearse A Keane, and Daniel C Alexander. Learning to Address Intra-segment Misclassification in Retinal Imaging. In *Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, pages 482–492, 2021.

[34] Yukun Zhou, Siegfried K Wagner, Mark A Chia, An Zhao, Moucheng Xu, Robbert Struyven, Daniel C Alexander, Pearse A Keane, and others. AutoMorph: automated retinal vascular morphology quantification via a deep learning pipeline. *Translational Vision Science & Technology*, 11(7):12, 2022. doi: 10.1167/tvst.11.7.12. URL https://tvst.arvojournals.org/article.aspx?articleid=2783477. Publisher: The Association for Research in Vision and Ophthalmology.

[35] Pedro Costa, Adrian Galdran, Maria Ines Meyer, Meindert Niemeijer, Michael Abràmoff, Ana Maria Mendonça, and Aurélio Campilho. End-to-end adversarial retinal image synthesis. *IEEE Transactions on Medical Imaging*, 37(3):781–791, 2018. doi: 10.1109/TMI.2017.2759102. URL http://ieeexplore.ieee.org/document/8055572/. Publisher: IEEE.

[36] He Zhao, Huiqi Li, Sebastian Maurer-Stroh, and Li Cheng. Synthesizing retinal and neuronal images with generative adversarial nets. *Medical Image Analysis*, 49:14–26, 2018. doi: 10.1016/j.media.2018.07.001. URL https://linkinghub.elsevier.com/retrieve/pii/S1361841518304596. Publisher: Elsevier.

[37] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd international conference on machine learning*, pages 2256–2265, Lille, France, 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

[38] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis, 2020.

[39] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*, pages 1–10, Vancouver BC Canada, 2022. ACM. doi: 10.1145/3528233.3530757. URL https://dl.acm.org/doi/10.1145/3528233.3530757.

[40] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 35:8633–8646, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/39235c56aef13fb05a6adc95eb9d8d66-Abstract-Conference.html.

[41] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D Diffusion, 2022. URL https://arxiv.org/abs/2209.14988.

[42] Elad Levi, Eli Brosh, Mykola Mykhailych, and Meir Perez. DLT: Conditioned layout generation with joint discrete-continuous diffusion layout transformer. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2106–2115, 2023. doi: 10.1109/ICCV51070.2023.00201. URL https://ieeexplore.ieee.org/document/10377598.

[43] Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, and others. Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors. In *Advances in Neural Information Processing Systems*, volume 36, pages 24705–24728, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/4ddab70bf41ffe5d423840644d3357f4-Paper-Conference.pdf.

[44] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 workshop on deep generative models and downstream applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

[45] Kai Jin, Xingru Huang, Jingxing Zhou, Yunxiang Li, Yan Yan, Yibao Sun, Qianni Zhang, Yaqi Wang, and Juan Ye. Fives: A fundus image dataset for artificial Intelligence based vessel segmentation. *Scientific data*, 9(1):475, 2022. doi: 10.1038/s41597-022-01564-3. URL https://www.nature.com/articles/s41597-022-01564-3. Publisher: Nature Publishing Group UK London.

[46] Natasa Popovic, Stela Vujosevic, Miroslav Radunović, Miodrag Radunović, and Tomo Popovic. TREND database: Retinal images of healthy young subjects visualized by a portable digital non-mydriatic fundus camera. *Plos one*, 16(7):e0254918, 2021. doi: 10.1371/journal.pone.0254918. URL https://dx.plos.org/10.1371/journal.pone.0254918. Publisher: Public Library of Science San Francisco, CA USA.

[47] Xiaoling Huang, Xiangyin Kong, Ziyan Shen, Jing Ouyang, Yunxiang Li, Kai Jin, and Juan Ye. GRAPE: A multi-modal dataset of longitudinal follow-up visual field and fundus images for glaucoma management. *Scientific Data*, 10(1):520, 2023. doi: 10.1038/s41597-023-02424-4. URL https://www.nature.com/articles/s41597-023-02424-4. Publisher: Nature Publishing Group UK London.

[48] Laboratoire de Traitement de l'Information Médicale (LaTIM INSERM U650). Messidor-2 dataset (méthodes d'Évaluation de systèmes de segmentation et d'Indexation dédiées à l'Ophthalmologie rétinienne), 2011. URL http://latim.univ-brest.fr/indexfce0.html.

[49] Ahmed Almazroa, Sami Alodhayb, Essameldin Osman, Eslam Ramadan, Mohammed Hummadi, Mohammed Dlaim, Muhannad Alkatee, Kaamran Raahemifar, and Vasudevan Lakshminarayanan. Retinal fundus images for glaucoma analysis: the RIGA dataset. In *Medical imaging 2018: Imaging informatics for healthcare, research, and applications*, volume 10579, page 105790B, 2018. doi: 10.1117/12.2293584. URL https://doi.org/10.1117/12.2293584.

[50] Oleksandr Kovalyk, Juan Morales-Sánchez, Rafael Verdú-Monedero, Inmaculada Sellés-Navarro, Ana Palazón-Cabanes, and José-Luis Sancho-Gómez. PAPILA: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. *Scientific Data*, 9(1):291, 2022. doi: 10.1038/s41597-022-01388-1. Publisher: Nature Publishing Group UK London.

[51] Chenwei Wu, David Restrepo, Luis Filipe Nakayama, Lucas Zago Ribeiro, Zitao Shuai, Nathan Santos Barboza, Maria Luiza Vieira Sousa, Raul Dias Fitterman, Alexandre Durao Alves Pereira, Caio Vinicius Saito Regatieri, and others. MBRSET: A portable retina fundus photos benchmark dataset for clinical and demographic prediction. *medRxiv : the preprint server for health sciences*, pages 2024–07, 2024. Publisher: Cold Spring Harbor Laboratory Press.

[52] Rolando Estrada, Michael J Allingham, Priyatham S Mettu, Scott W Cousins, Carlo Tomasi, and Sina Farsiu. Retinal artery-vein classification via topology estimation. *IEEE transactions on medical imaging*, 34(12):2518–2534, 2015. doi: 10.1109/TMI.2015.2443117. URL https://ieeexplore.ieee.org/document/7120990. Publisher: IEEE.

[53] Jan Van Eijgen, Jonathan Fhima, Marie-Isaline Billen Moulin-Romsée, Joachim A Behar, Eirini Christinaki, and Ingeborg Stalmans. Leuven-haifa high-resolution fundus image dataset for retinal blood vessel segmentation and glaucoma diagnosis. *Scientific Data*, 11(1):257, 2024. doi: 10.1038/s41597-024-03086-6. URL https://www.nature.com/articles/s41597-024-03086-6. Publisher: Nature Publishing Group UK London.

[54] Ling-Ping Cen, Jie Ji, Jian-Wei Lin, Si-Tong Ju, Hong-Jie Lin, Tai-Ping Li, Yun Wang, Jian-Feng Yang, Yu-Fen Liu, Shaoying Tan, and others. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nature Communications*, 12(1):4828, 2021. doi: 10.1038/s41467-021-25138-w. URL https://www.nature.com/articles/s41467-021-25138-w. Publisher: Nature Publishing Group UK London.

[55] Tianyu Li, Yuan Gao, Ke Wang, Shuo Guo, Hao Liu, and Haibo Kang. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences*, 501:511–522, October 2019. doi: 10.1016/j.ins.2019.06.011. URL https://www.sciencedirect.com/science/article/abs/pii/S0020025519305377.

[56] E. Dugas, Jorge Jared, and W. Cukierski. Diabetic retinopathy detection, 2015. URL https://www.kaggle.com/competitions/diabetic-retinopathy-detection.

[57] Muhammad Naseer Bajwa, Gur Amrit Pal Singh, Wolfgang Neumeier, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. G1020: A benchmark retinal fundus image dataset for computer-aided glaucoma detection. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–7. IEEE, 2020. doi: 10.1109/IJCNN48605.2020.9207664.

[58] Prashant Porwal, Sachin Pachade, Rishikesh Kamble, and et al. Indian diabetic retinopathy image dataset (IDRiD), April 2018.

[59] ODIR Team. ODIR dataset: Ocular disease intelligent recognition, 2019. URL https://odir2019.grand-challenge.org/.

[60] Junlin Hou, Jilan Xu, Fan Xiao, Rui-Wei Zhao, Yuejie Zhang, Haidong Zou, Lina Lu, Wenwen Xue, and Rui Feng. Cross-field transformer for diabetic retinopathy grading on two-field fundus images. In *2022 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 985–990. IEEE Computer Society, 2022.

[61] Attila Budai, Rüdiger Bock, Andreas Maier, Joachim Hornegger, and Georg Michelson. Robust vessel segmentation in fundus images. *International Journal of Biomedical Imaging*, 2013:154860, 2013. doi: 10.1155/2013/154860. URL https://doi.org/10.1155/2013/154860.

[62] Meindert Niemeijer, Xiayu Xu, Alina V. Dumitrescu, Priya Gupta, Bram van Ginneken, James C. Folk, and Michael D. Abràmoff. Automated measurement of the arteriolar-to-venular width ratio in digital color fundus photographs. *IEEE Transactions on Medical Imaging*, 30(11):1941–1950, 2011. doi: 10.1109/TMI.2011.2159619. URL https://doi.org/10.1109/TMI.2011.2159619.

[63] Verónica Elisa Castillo Benítez, Ingrid Castro Matto, Julio César Mello Román, José Luis Vázquez Noguera, Miguel García-Torres, Jordan Ayala, Diego P. Pinto-Roa, Pedro E. Gardel-Sotomayor, Jacques Facon, and Sebastian Alberto Grillo. Dataset from fundus images for the study of diabetic retinopathy. *Data in Brief*, 36:107068, 2021. doi: 10.1016/j.dib.2021.107068. URL https://doi.org/10.1016/j.dib.2021.107068.

[64] J. Zhang, B. Dashtbozorg, E. Bekkers, J. P. W. Pluim, R. Duits, and B. M. ter Haar Romeny. Robust retinal vessel segmentation via locally adaptive derivative frames in orientation scores. *IEEE Transactions on Medical Imaging*, 35(12):2631–2644, December 2016. ISSN 0278-0062. doi: 10.1109/TMI.2016.2587062. URL http://ieeexplore.ieee.org/document/7530915/.

[65] Qiang Hu, Michael D Abràmoff, and Mary K Garvin. Automated separation of binary overlapping trees in low-contrast color retinal images. In *Medical image computing and computer-assisted intervention–MICCAI 2013*, volume 16, Part 2 of *Lecture notes in computer science*, pages 436–443, Berlin, Heidelberg, 2013. Springer. doi: 10.1007/978-3-642-40763-5_54. URL https://link.springer.com/chapter/10.1007/978-3-642-40763-5_54.

[66] Jeroen Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, 23(4):501–509, 2004. doi: 10.1109/TMI.2004.825627. URL http://ieeexplore.ieee.org/document/1282003/. Publisher: IEEE.

[67] Y Men, J Fhima, LA Celi, LZ Ribeiro, LF Nakayama, and JA Behar. Deep learning generalization for diabetic retinopathy staging from fundus images. *Physiological Measurement*, 13(1), 2025. doi: 10.1088/1361-6579/ada86a. URL https://doi.org/10.1088/1361-6579/ada86a.

[68] Jonathan Fhima, Jan Van Eijgen, Anat Reiner-Benaim, Lennert Beeckmans, Or Abramovich, Ingeborg Stalmans, and Joachim A Behar. Computerized analysis of the eye vasculature in a mass dataset of digital fundus images: the example of age, sex and primary open-angle glaucoma, 2024. URL https://www.medrxiv.org/content/10.1101/2024.07.21.24310763v1. Publisher: Cold Spring Harbor Laboratory Press.

[69] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, New Orleans, LA, USA, 2022. IEEE. doi: 10.1109/CVPR52688.2022.01042. URL https://ieeexplore.ieee.org/document/9878449/.

[70] Jose Vargas Quiros, Bart Liefers, Karin van Garderen, Jeroen Vermeulen, Eyened Reading Center, Sinergia Consortium, and Caroline Klaver. VascX models: Model ensembles for retinal vascular analysis from color fundus images, 2024. URL http://arxiv.org/abs/2409.16016.

[71] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, and others. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL https://openreview.net/forum?id=a68SUt6zFt.

[72] Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, and others. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023. doi: 10.1038/s41586-023-06555-x. URL https://www.nature.com/articles/s41586-023-06555-x. Publisher: Nature Publishing Group UK London.

[73] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, and others. Swin transformer v2: Scaling up capacity and resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12009–12019, 2022. doi: 10.1109/CVPR52688.2022.01170. URL https://ieeexplore.ieee.org/document/9879380/.

[74] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. doi: 10.1109/CVPR52688.2022.01553.

[75] Rizhao Fan, Matteo Poggi, and Stefano Mattoccia. Contrastive learning for depth prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3226–3237, 2023. doi: 10.1109/CVPRW59228.2023.00325. URL https://ieeexplore.ieee.org/document/10208878.

[76] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

[77] Huihui Yu and Qun Dai. Self-supervised multi-task learning for medical image analysis. *Pattern Recognition*, 150:110327, 2024. doi: 10.1016/j.patcog.2024.110327. URL https://linkinghub.elsevier.com/retrieve/pii/S0031320324000785. Publisher: Elsevier.

[78] Jonathan Fhima, Elad Ben Avraham, Oren Nuriel, Yair Kittenplon, Roy Ganz, Aviad Aberdam, and Ron Litman. TAP-VL: Text layout-aware pre-training for enriched vision-language models, 2024. URL http://arxiv.org/abs/2411.04642.

[79] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, volume 30, pages 6629–6640, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

[80] Suprosanna Shit, Johannes C Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien PW Pluim, Ulrich Bauer, and Bjoern 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)*, pages 16555–16564, Nashville, TN, USA, 2021. doi: 10.1109/CVPR46437.2021.01629. URL https://ieeexplore.ieee.org/document/9578225/.

[81] Benjamin Hou, Amir Alansary, Daniel Rueckert, and Bernhard Kainz. High-fidelity diabetic retina fundus image synthesis from freestyle lesion maps, 2022.

[82] Sarah Müller, Lisa M. Koch, P. A. Lensch, Hendrik, and Philipp Berens. Disentangling representations of retinal images with generative models, 2024.

[83] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th international conference on learning representations, ICLR 2019, new orleans, LA, USA, may 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7. tex.bibsource: dblp computer science bibliography, https://dblp.org tex.timestamp: Thu, 25 Jul 2019 14:26:04 +0200.

[84] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations (ICLR 2017)*, pages 1769–1784, Toulon, France, 2017.

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: The abstract and introduction clearly state the main contributions of the paper, including the introduction of the RLAD framework, its application to retinal vessel segmentation, and the release of a new annotated dataset. These claims are supported by the experimental results and discussion in Sections 5 and 6.

2. **Limitations**

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

   Answer: [Yes]

   Justification: The limitations of the RLAD approach, including potential domain gaps, computational demands, and reliance on accurate annotations, are now discussed in the section 7.

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: [NA]

   Justification: The paper does not include theoretical results.

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: The paper provides open access to both the code of the main experience and the new dataset, with detailed instructions for data preprocessing, model training, and evaluation. All hyperparameters, data splits, and implementation details are described in the main paper and the supplementary material.

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: The code for training and evaluating RLAD, as well as the new REYIA dataset, are provided with the paper.

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, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental settings, including model architectures, optimizer types, learning rates, batch sizes, number of epochs, and data splits, are detailed.

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: Table 2, which support the main claim of the paper is also reporting a 2-sigma error bar.

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: Computer ressources used in the experiments are provided in 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 adheres to the NeurIPS Code of Ethics. All datasets are properly anonymized, patient privacy is respected, and no personally identifiable information is used. An Institutional Review Board (IRB) approval have been received for the newly published data source.

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We discussed the positive and negative societal impact in section 7.

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., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Although the risk for misuse is low, we have implemented the following safeguards: (1) The dataset which is provided (DFI with segmentation masks) is either anonymized or derived from publicly available sources, and thus potential risks of misuse, such as unintended re-identification, are mitigated.
    (2) Data and models will be released under a research-only license prohibiting clinical or commercial use.
    (3) Access to data will require agreement to terms of use, and all users must register and accept usage guidelines.

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: All third-party datasets, models and codebases used are cited in the references.

13. **New assets**

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

    Answer: [Yes]

    Justification: The new REYIA dataset and RLAD code are fully documented, with instructions for use.

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: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

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: [Yes]

    Justification: All images used but one were open access. For the ENRICH dataset IRB was obtained (S60649).

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 rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.