# A General Bézier Tree Encoding Counterfactual Framework for Retinal-Vessel-Mediated Disease Analysis

**Anonymous authors**
Paper under double-blind review

## Abstract

The geometry of the retinal vessel is a key biomarker of vascular diseases, yet clinical evidence remains primarily observational. Disease-label-conditioned generative counterfactuals do not directly expose explicit anatomical parameters for intervention. To address this limitation, we propose the Bézier Tree Encoding Counterfactual Framework (BTECF). BTECF abstracts vascular networks into interconnected cubic-Bézier segments to establish a disease-agnostic representation in which structural topology is explicitly preserved and atomically perturbable. BTECF couples this encoding with a diffusion-based generator to enable parameter-level *do*-interventions on explicit geometric axes (e.g., tortuosity, caliber) while preserving background fundus textures. We validate BTECF on diabetic retinopathy, together with independent cohorts for ischemic stroke and Alzheimer's disease. Counterfactual interventions reveal dose-dependent classifier responses, with weaker pixel-drop responses in the DR and stroke cases and reduced specificity in the AD case. BTECF provides a unified computational framework for probing how disease classifiers respond to explicit vessel-geometry interventions; biological causality requires prospective validation.

## 1 Introduction

The retina provides a unique window for direct, non-invasive optical observation of the central nervous system (CNS) (Zhao et al., 2024). Clinical studies have demonstrated that retinal microvascular alterations are associated with the onset of stroke (Wong et al., 2001; McGeechan et al., 2009; Doubal et al., 2010), cardiovascular events and hypertension risk (Wong et al., 2004; Sandoval-Garcia et al., 2021), as well as Alzheimer's disease and cognitive decline (Frost et al., 2013; Lemmens et al., 2020; Zhang et al., 2022). In particular, morphological changes in the retinal vasculature, including caliber variation, tortuosity, and branching topology, serve as key indicators of systemic disease pathology (Arnould et al., 2023; Yusufu et al., 2025; Doubal et al., 2010).

Retinal vessel segmentation has progressed from handcrafted methods (Hoover et al., 2000; Soares et al., 2006) to convolutional networks (Ronneberger et al., 2015). AutoMorph combines automated retinal analysis with adversarial vessel segmentation based on a U-Net variant (Zhou et al., 2022). Transformer-based medical segmentation is represented by TransUNet (Chen et al., 2024).

Recent approaches predict Diabetic Retinopathy (Liu et al., 2025), Stroke (Khalafi et al., 2025), and Alzheimer's Disease (Jamshidiha et al., 2025) from retinal images. These predictions motivate a complementary query: "If we edit a specific geometric feature $X$ of a given vessel, how would the disease prediction change?"

To implement this query, we introduce Bézier Tree Encoding (Bézier, 1972), a numerical representation of the retinal vascular tree, analogous to the SWC parametric standard for neuronal morphology (Mehta et al., 2023). A $512 \times 512$ pixel mask is represented by 200–500 interconnected cubic-Bézier segments, exposing control points, segment presence, and a radius field for editing. The same encoding supports observational analysis through a 20-dimensional feature vector and generative conditioning through a spatial hint. Building on this encoding, the Bézier Tree Encoding Counterfactual Framework (BTECF; Figure 1) connects these edits to image synthesis and classifier scoring. Relative to pixel-mask conditioning approaches (Feng et al., 2024; Go et al., 2024; Guo et al.,



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

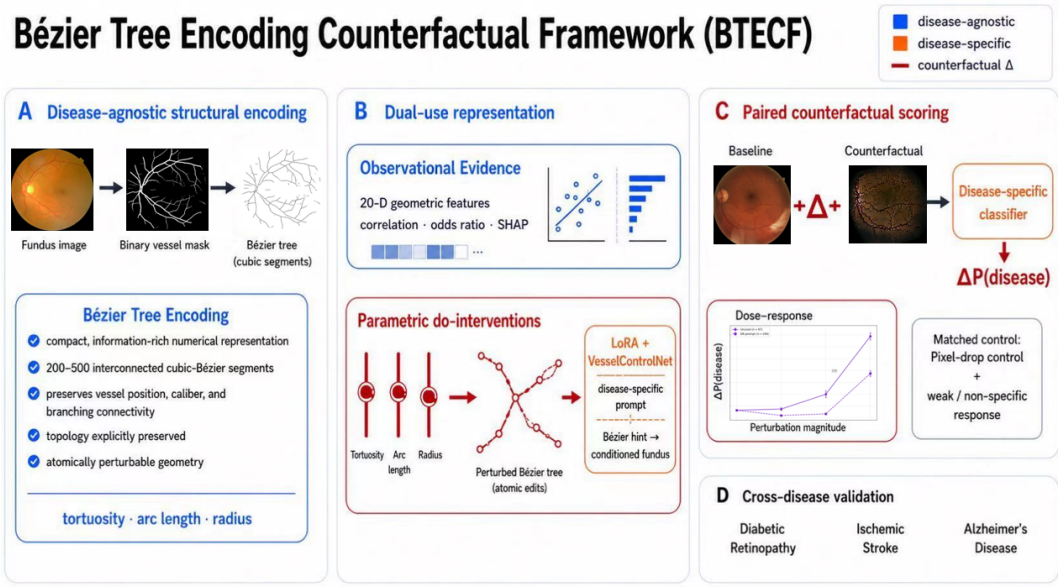


Figure 1: BTECF overview. Color coding: blue = disease-agnostic, orange = disease-specific, red = counterfactual response $\Delta$. **(A)** Cubic-Bézier encoding of the segmented fundus. **(B)** Dual-use representation: 20-D geometric features for observational analysis (top); Bézier hint conditions a LoRA + VesselControlNet generator for explicit geometric edits (bottom). **(C)** Paired counterfactual scoring yields $\Delta P(\text{disease})$; pixel-drop control compares the response with unstructured mask corruption. **(D)** Applications on diabetic retinopathy, ischemic stroke, and Alzheimer's disease.

2024) and classical parametric vessel renderers (Bonaldi et al., 2016; Castro et al., 2020), our contribution is an integrated interface for geometric editing, image generation, and paired classifier-response measurement.

BTECF operates through three modules in Figure 1. The disease-agnostic upstream (Fig. 1A, blue) fits the vessel masks from segmented fundi (Zhou et al., 2022). This encoding is dual-use (Fig. 1B): 20-dimensional geometric features support observational analysis, while a three-channel rasterized hint conditions VesselControlNet (Zhang et al., 2023a) and a LoRA-adapted (Hu et al., 2022) diffusion generator. Within-start paired scoring (Fig. 1C) measures $\Delta P(\text{disease})$ between baseline and edited generations, with pixel-drop as a corruption comparison. The pipeline is instantiated on diabetic retinopathy, ischemic stroke, and Alzheimer's disease (Fig. 1D).

DR is the primary evaluation case. The full set of starts establishes the strongest tortuosity-edit response beyond the selected low-baseline subset, while prompt and backbone comparisons characterize its dependence on evaluation conditions. Stroke and AD extend the protocol to more challenging scoring regimes, including weak discrimination and reduced control specificity. Throughout, "counterfactual" refers to an edit of the computational pipeline: the operators target distinct parameters in the representation, while downstream geometric and image attributes may covary.

**Our contributions**

1. **An explicit geometric editing interface.** BTECF connects automatically extracted cubic-Bézier vessel trees, rasterized diffusion conditioning, and classifier scoring. One representation supports observational features and graded edits targeting shape, segment presence, and caliber.

2. **A paired model-response protocol.** We report within-start prediction changes across all 500 DR starts alongside selected low-baseline subsets. Prompt, backbone, perturbation-strength, and pixel-drop comparisons characterize dependence on evaluation conditions.

3. **Prompt-conditioned and cross-disease analysis.** DR responses are strongest under the tortuosity edit and attenuated under disease-text conditioning. Stroke and AD extend the same protocol to regimes with weaker discrimination and reduced response specificity.

## 2 Related Work

**Observational vessel-geometry studies** Retinal vascular studies characterize associations across DR (Cheung et al., 2008; Klein et al., 2004; Velayutham et al., 2020; Xu et al., 2026) and other diseases using diverse measurement protocols. These representations already extend beyond a few summary indices: Frost et al. (2013) measure 19 parameters including branching geometry, while Yusufu et al. (2025) assess 118 parameters spanning multiple vascular properties. BTECF's distinction is not feature count: its 20-dimensional summary is derived from an editable curve representation that also supplies generation hints. This connects observational measurements to explicitly specified input edits and classifier-response analysis. Mainstream DR observational studies report four characteristic geometric changes: vessel arc length decreases as the vasculature atrophies (Forster et al., 2021; Pramil et al., 2021); fractal dimension declines (Broe et al., 2014; Popovic et al., 2019; Forster et al., 2021; Pramil et al., 2021); vessel tortuosity rises (Sasongko et al., 2011; 2016; Forster et al., 2021; Fathimah et al., 2025; Xu et al., 2026); and vessel caliber follows a biphasic trajectory across DR severity, widening in early DR and narrowing in late stages (Ashraf et al., 2021; Cheung et al., 2008; Klein et al., 2004; Velayutham et al., 2020; Xu et al., 2026). These observations anchor the three perturbation axes (tortuosity, arc length, radius) we select in §3.4.

**Generative and counterfactual fundus imaging** Mask-conditioned deep generators synthesize retinal images from rasterized anatomical maps, including vessel masks and DR lesion masks (Feng et al., 2024; Go et al., 2024; Guo et al., 2024). Go et al. condition fundus diffusion on artery/vein masks and evaluate realism with clinical experts as well as downstream segmentation augmentation. Explicit geometry editing can complement this synthesis paradigm. Classical parametric vessel-tree renderers (Bonaldi et al., 2016; Castro et al., 2020) provide interpretable geometry without a learned fundus texture prior. Deep structural causal models (Pawlowski et al., 2020) target medical imaging more broadly. The fundus counterfactual method of Ilanchezian et al. (2025) guides diffusion with disease-classifier gradients and evaluates realism and clinical meaningfulness with domain experts; vessel-related changes are less prominent in its reported results. CauDR (Wei et al., 2024) uses do-operations on frequency-domain features for DR domain generalization, while counterfactual contrastive learning (Roschewitz et al., 2025) varies acquisition-domain factors for representation robustness. Quadratic-Bézier curves have also been used for procedural vessel synthesis as segmentation pretraining augmentation (Niu et al., 2023). BTECF specifies edits on an extracted vessel tree before generation and measures the resulting classifier response.

## 3 Method

### 3.1 Framework Overview

BTECF is organized as six modules: vessel segmentation (Zhou et al., 2022), Bézier Tree Encoding parameterization, a LoRA (Hu et al., 2022) domain adapter on Stable Diffusion 2.1 (Rombach et al., 2022), a VesselControlNet (Zhang et al., 2023a) conditioned on the Bézier hint, a downstream disease-specific classifier, and a within-start paired counterfactual scoring protocol. Vessel segmentation, Bézier parameterization, and the scoring protocol are disease-agnostic and transfer with zero retraining; LoRA, ControlNet, and classifier are instantiated per disease. The Bézier parametric pipeline yields both a 20-dimensional geometric feature vector for the observational analysis and a three-channel ControlNet hint for the generative and counterfactual analyses. Figure 1 shows the full pipeline; per-module deployment cost is in Appendix A.

### 3.2 Bézier Parametric Hint

Cubic-Bézier curves (Bézier, 1972) are a standard parametric basis in computer graphics for smooth curve representation. Given a segmented vessel mask, we split the skeleton at branch nodes and fit each resulting polyline piecewise into 200–500 cubic-Bézier segments per fundus, retaining

shared branch endpoints and approximating each segment. This parametric representation is then rasterized into a three-channel ControlNet hint: Channel 0 carries the distance-transform radius field of the raw mask, Channel 1 renders the Bézier curves at variable radius read from Channel 0, and Channel 2 is a Gaussian-smoothed version of Channel 1. This split provides separate axis-specific editing operations (§3.4), although the resulting image features may covary downstream. Per-segment equations, rasterization parameters, the empirical Channel-0 field check, and construction constraints are in Appendix B.1. Figure 2 visualizes the Bézier fit and the resulting three-channel hint on a representative fundus.

### 3.3 Training and Inference

We finetune Stable Diffusion 2.1 (Rombach et al., 2022) with a LoRA adapter (Hu et al., 2022) (rank 32, 36k steps) and `noise_offset = 0.1`, motivated by prior observations that offset noise improves diffusion models' ability to represent global brightness shifts. Empirically, in our fundus finetuning runs, it reduced gray/blue color drift. A VesselControlNet (Zhang et al., 2023a) checkpoint is trained on our Bézier hint and used by counterfactual intervention. For discriminative evaluation, we fine-tune a BiomedCLIP encoder (Zhang et al., 2023b) on the target disease labels, then freeze it for feature extraction. The downstream classifier is trained from scratch on the deduplicated DR training split: EfficientNet-B2 serves as the primary scorer for the counterfactual intervention, ResNet-50 provides the cross-backbone sensitivity check, and a ViT-B/16 trained at the same scale underperforms at this training budget and is treated as diagnostic rather than confirmatory. Per-backbone AUC and sensitivity figures are in Appendix B.4.

At inference we use classifier-free guidance (Ho & Salimans, 2022) with weight $w = 7.5$,

$$\epsilon_{\text{final}} = \epsilon_{\text{unc}} + w \cdot (\epsilon_{\text{cond}} - \epsilon_{\text{unc}}), \tag{1}$$

where the unconditional branch corresponds to an empty text string and is exactly the branch we analyse under the Uncond condition (§4.4.1). For the counterfactual intervention, the same starting grade-0 fundi are generated under two conditions: (A) DR prompt and (B) Uncond, with all perturbation configurations identical across conditions, allowing comparison of responses under the two prompt conditions. The reported effect is the within-start paired difference

$$\Delta_i \;=\; P(\text{disease} \mid \text{pert}, \text{start}_i) \;-\; P(\text{disease} \mid \text{baseline}, \text{start}_i), \tag{2}$$

aggregated over the $n$ valid starts as $\text{mean} \pm \text{SEM}$; an additive generator bias shared by baseline and perturbed generations cancels in $\Delta$, while perturbation-dependent bias may remain. 95% confidence intervals use the one-sample $t$-distribution and $p$-values test $\Delta = 0$. Per-backbone hyperparameters, the prompt-pairing protocol, the empirical color root-cause analysis, and inference hyperparameters (DDIM steps, ControlNet residual strength, LoRA weights, latent offset) are in Appendix B.2 and B.5.

### 3.4 Counterfactual Perturbation

The counterfactual intervention perturbs three Bézier feature axes (tortuosity, arc length, and radius), selected on four grounds: (i) observational analysis (SHAP top-5 features, §4.2); (ii) availability of explicit editing operations; (iii) literature motivation across diseases (Forster et al., 2021; Frost et al., 2013; Lemmens et al., 2020; Sandoval-Garcia et al., 2021; Doubal et al., 2010; Wong et al., 2001; Klein et al., 2004; Cheung et al., 2008; Velayutham et al., 2020; Xu et al., 2026; Wong et al., 2004); and (iv) dose-response quantifiability (graded strengths per axis). Per-axis justification is in Appendix B.3; the same three-axis selection applies to the cross-disease cases without re-selection.

Each perturbation family uses a distinct operation on the parametric representation. The *tortuosity family* at multipliers $\alpha \in \{1, 2, 4\}$ perpendicularly displaces the interior control points $P_1, P_2$ by an amount proportional to the segment's chord length. For a cubic Bézier segment with chord vector $\vec{v} = P_3 - P_0$ of length $L = \|\vec{v}\|$ and unit perpendicular $\hat{n} = (-v_y, v_x)/L$, with a per-segment random sign $s \in \{-1, +1\}$ drawn uniformly,

$$P_1' = P_1 + s \cdot \gamma\alpha L \cdot \hat{n}, \qquad P_2' = P_2 - s \cdot \gamma\alpha L \cdot \hat{n}, \tag{3}$$

where $\gamma = 0.15$ is the base displacement coefficient. Fixing endpoints $P_0, P_3$ preserves chord length and encoded endpoint connections; the antisymmetric $P_1, P_2$ assignment produces an S-shaped local curvature within each segment. The $\alpha = 1$ multiplier is therefore non-zero displacement; the

zero-displacement reference is the baseline configuration that bypasses this function entirely. The *arc-drop family* at strengths $10\%, 20\%, 30\%$ randomly deletes a fraction of Bézier segments, leaving the remaining segments intact. The *radius family* at scaling factors $0.85, 0.70, 0.55$ multiplies the distance-transform field by a global factor, reducing caliber uniformly. A *pixel-drop control* uses the same nominal deletion percentages and re-runs skeletonization and Bézier fitting after removing mask pixels uniformly at random. It provides an unstructured mask-corruption reference rather than a deformation-matched control for magnitude, spatial frequency, or perceptual distance. Shape edits can also change arc length and curvature, while segment deletion can change connectivity.

### 3.5 Evaluation Protocols

**Observational analysis** For each fundus in the DR cohort we automatically extract the 20-dimensional Bézier feature vector and apply three complementary statistical analyses: Spearman rank correlation captures the marginal monotonic trend; logistic regression yields per-SD and quintile (Q5/Q1) odds ratios summarizing standardized and extreme-quintile associations; TreeSHAP importance on a trained XGBoost model also reports a per-feature *feature-versus-SHAP correlation* (`feat-SHAP corr`), defined as the Pearson correlation between each feature's raw value and its tree-conditional SHAP contribution. A sign disagreement between Spearman $\rho$ and `feat-SHAP corr` identifies a difference between marginal association and conditional model attribution, motivating examination of non-linearity, correlated features, and interactions. The sign comparison alone does not select among these explanations or identify a U-shaped risk curve.

**Discriminative benchmark** An XGBoost benchmark compares predictive performance using each feature representation. We compare the 20-dimensional Bézier geometric features with frozen BiomedCLIP features, both at native dimensionality and in a matched 20-dimensional PCA regime.

**Counterfactual intervention** We draw 500 grade-0 fundi from the held-out test split and apply 13 perturbation configurations (one baseline plus three doses each of arc-drop, radius scaling, tortuosity, and the pixel-drop control), giving 6,500 ControlNet generations per prompt condition. All images are scored with the EfficientNet-B2 classifier (primary); a ResNet-50 sensitivity check is reported on the original subset $n = 50$. We report two complementary estimators: an *all-start mean response* (reported as ATE) on all $n = 500$ starts, and a *strict reverse-counterfactual subset* of starts with baseline DR probability below $0.3$ that pass a baseline intensity/color filter (full thresholds in Appendix B.5). The population ATE covers all starts; the strict subset focuses on disease-emergence responses from low-probability baselines that pass this baseline filter. The latter reduces ceiling effects but estimates a response in a selected population. The Strict-subset yields are $n = 97$ (Uncond) and $n = 190$ (DR-prompt).

## 4 Experiments

### 4.1 Data and Setup

**DR primary case** We use the Eyepacs, Aptos, Messidor Diabetic Retinopathy dataset on Kaggle, which curates four publicly available DR datasets into a single $600 \times 600$ resized collection (EyePACS (Dugas et al., 2015), APTOS-2019 (Karthik et al., 2019), APTOS Gaussian Filtered, and Messidor-2 (Decencière et al., 2014)). The redistribution applies manual augmentation to expand the dataset by approximately $55\%$, yielding $143{,}669$ images. The grades follow the 5-step ICDR clinical scale: $0 =$ no DR, $1 =$ mild NPDR, $2 =$ moderate NPDR, $3 =$ severe NPDR, $4 =$ proliferative DR. We use binary labels throughout: grade 0 as normal and grades 1–4 as DR. We apply the recorded base-ID deduplication, removing 2,954 redundant rows to obtain the final splits of $114{,}209/13{,}269/13{,}237$ (train / val / test). The val + test union ($n = 26{,}506$) is the cohort of observational and discriminative analysis. Detailed deduplication procedure in Appendix B.6 and data license in Appendix I.

**Cross-disease cases** Two independent UK Biobank (Sudlow et al., 2015) cohorts with disjoint patient groups: an ischemic-stroke cohort ($n = 11{,}361$ participants; $3{,}409$ evaluation, $12\%$ positive) and an Alzheimer's disease cohort ($n = 1{,}877$ participants; $550$ evaluation, $7.5\%$ positive). Cohort design and disease-specific retraining are discussed in §4.5 and §3.1.

### 4.2 Observational Findings

The SHAP top-5 features on the DR cohort (Table 1) recover the geometric directions reported in mainstream DR observational and histopathological literature (§2): arc length and radius decrease with disease, tortuosity rises, and branching density shows marginal–conditional sign disagreement. The complete 20-feature analysis is in Appendix C.2 and a detailed pathology analysis is in Appendix H.

Table 1: SHAP top-5 features on the DR cohort. Metrics defined in §3.5. Significance: $***$ $p < 0.001$, $* p < 0.05$.

| Rank | Feature | Spearman $\rho$ | feat-SHAP corr | OR/SD | Q5/Q1 OR | SHAP abs | Direction |
|---|---|---|---|---|---|---|---|
| 1 | `total_arc_length` | $-0.358$*** | $-0.94$ | 0.55 | 0.20 | 0.456 | $\downarrow$ DR |
| 2 | `branching_density` | $+0.014$* | $-0.55$ | 1.21 | 1.09 | 0.364 | SHAP non-linear |
| 3 | `mean_radius` | $-0.084$*** | $-0.90$ | 0.85 | 0.68 | 0.285 | $\downarrow$ DR |
| 4 | `mean_chord_length` | $-0.079$*** | $-0.84$ | 0.80 | 0.67 | 0.231 | $\downarrow$ DR |
| 5 | `mean_tortuosity` | $+0.086$*** | $+0.17$ | 1.00 | 1.60 | 0.160 | $\uparrow$ DR |

### 4.3 Discriminative Results

Using only the 20-dimensional Bézier geometric features yields test AUC $0.706$, covering $61\%$ of the random-to-BiomedCLIP attainable gap; matching dimensionality at 20 gives AUC $0.71$ (Bézier) and $0.80$ (BiomedCLIP PCA) (Table 2). These results show that the intervened geometric parameters carry disease-discriminative information, motivating their use as explicit controls for probing classifier behavior.

Table 2: XGBoost discriminative comparison on the DR cohort.

| Feature set | Dim | Test AUC | 5-fold CV AUC |
|---|---|---|---|
| Bézier geometric (this work) | 20 | 0.7058 | $0.7146 \pm 0.0037$ |
| BiomedCLIP frozen features | 512 | 0.8358 | $0.8587 \pm 0.0031$ |
| BiomedCLIP PCA to 20-d (EVR $= 85\%$) | 20 | 0.8028 | $0.8128 \pm 0.0030$ |

### 4.4 Parameter-Level Counterfactual Intervention

The complete results on the DR case follow (§4.4.1–§4.4.3); cross-disease applications are in §4.5.

#### 4.4.1 Primary Result

Empty-prompt generation (the unconditioned CFG branch) measures the response to vessel-geometry interventions without disease-text conditioning. Under Uncond condition in the strongest tortuosity dose (`tortuosity_4x`), the EfficientNet-B2 classifier produces a population-level ATE of $\Delta = +0.432$ (95% CI $[+0.399, +0.464]$, $p < 10^{-10}$, $n = 500$); in the strict reverse-counterfactual subset ($n = 97$), the mean probability of DR increases from $0.064$ at baseline to $0.689$ after intervention ($\Delta = +0.625$, $p < 10^{-10}$). The pixel-drop control has $\sim 25\times$ lower mean post-intervention DR probability than the strongest tortuosity dose; its negative mean change is a corruption response, not evidence of a null effect. Under the DR-prompt condition, the same classifier produces a population-level ATE of $\Delta = +0.079$ (95% CI $[+0.036, +0.122]$, $p = 3.7 \times 10^{-4}$, $n = 500$); in the strict reverse-counterfactual subset ($n = 190$), the mean probability of DR increases from $0.064$ at baseline to $0.375$ after intervention ($\Delta = +0.311$, $p < 10^{-10}$). Relative to Uncond, the DR-prompt ATE is $\sim 5.5\times$ smaller at the population level and $\sim 2.0\times$ smaller in the respective strict subsets; the latter comparison also reflects prompt-specific selection. This attenuation is discussed in §5. The complete landscape of 13-configurations $\Delta$ (Table 3) shows the largest response for the strongest tortuosity edit; per-perturbation analysis and the cross-prompt response curves are in Appendix D.

#### 4.4.2 Backbone and Prompt Sensitivity

The strongest tortuosity effect (`tortuosity_4x`) is positive and statistically significant in all four cells of the $2 \times 2$ backbone $\times$ prompt design (Table 4); EfficientNet-B2 cells use strict subsets drawn

Table 3: Full 13-configuration cross-condition $\Delta$ table on the DR primary case (paired within-start differences vs. baseline). (A) DR-prompt at $n = 190$; (B) Uncond at $n = 97$; both at $w = 7.5$, EfficientNet-B2 scorer, strict reverse-counterfactual subset.

| Perturbation | (A) DR-prompt $\Delta$ | (B) Uncond $\Delta$ |
|---|---|---|
| `tortuosity_4x` (strong) | $+0.311$ ($p<10^{-10}$) | $+0.625$ ($p<10^{-10}$) |
| `tortuosity_2x` (mid) | $-0.031$ | $+0.136$ |
| `tortuosity_1x` (weak, 15% chord) | $-0.044$ | $+0.011$ |
| `arc_drop_30` | $+0.002$ | $-0.011$ |
| `arc_drop_20` | $+0.025$ | $-0.009$ |
| `arc_drop_10` | $+0.017$ | $-0.014$ |
| `radius_x0.55` (strong thin) | $-0.059$ | $-0.020$ |
| `radius_x0.70` (mid thin) | $-0.058$ | $-0.009$ |
| `radius_x0.85` (weak thin) | $-0.058$ | $-0.026$ |
| `pixdrop_30` (control) | $-0.044$ | $-0.036$ |
| `pixdrop_20` | $-0.031$ | $-0.037$ |
| `pixdrop_10` | $-0.016$ | $-0.010$ |
| baseline | 0 (by def.) | 0 |

from $n = 500$ starts, and ResNet-50 cells use subsets drawn from $n = 50$ pilot starts. This establishes a positive aggregate response across the tested backbone–prompt conditions. Because cohort sizes and selected starts differ, the comparison concerns aggregate direction rather than image-level agreement. ViT-B/16 trained on the same scale yields only one qualifying start and is excluded.

Table 4: Counterfactual response $\Delta$ on `tortuosity_4x` across the $2 \times 2$ backbone $\times$ prompt design. All four cells significant at $p < 5 \times 10^{-3}$.

| Backbone | DR-prompt $\Delta$ | Uncond $\Delta$ |
|---|---|---|
| EfficientNet-B2 | $+0.311$ | $+0.625$ |
| ResNet-50 | $+0.294$ | $+0.377$ |

#### 4.4.3 Guidance-Weight Ablation

We sweep the CFG guidance weight $w \in \{3, 5, 7.5, 10\}$ on the $n = 50$ pilot subset under the DR-prompt condition (Table 5). DR-prompt $\Delta$ rises monotonically across $w \geq 5$, corresponding to 15%, 54%, and 72% of the Uncond reference $\Delta = +0.625$ (§4.4.1) at $w = 5, 7.5, 10$. This pattern is consistent with prompt and hint affecting overlapping classifier-relevant visual content (§5): the observed DR-prompt response approaches the Uncond reference as $w$ rises. Retained starts differ across settings; the table's "ceiling" denotes this empirical reference, not an established upper bound.

Table 5: Guidance-weight ablation on `tortuosity_4x` ($n = 50$ pilot, DR-prompt). DR-prompt $\Delta$ partially recovers toward the Uncond ceiling as $w$ rises, but never crosses it. Uncond ceiling $\Delta = +0.625$ from the $n = 500$ main strict subset (§4.4.1); $w$ has no formal effect under `prompt_mode=uncond` since both CFG branches encode the empty string.

| $w$ | DR-prompt $\Delta$ ($n$, $p$) | DR-prompt $\Delta$ / Uncond ceiling |
|---|---|---|
| 3.0 | 0.150 ($n$=14, $p$=0.14, n.s.) | 0.24 (n.s.) |
| 5.0 | 0.095 ($n$=14, $p$=0.05) | 0.15 |
| 7.5 | 0.340 ($n$=17, $p$=$3.3\times10^{-3}$) | 0.54 |
| 10.0 | 0.450 ($n$=13, $p$=$5.3\times10^{-5}$) | 0.72 |

### 4.5 Cross-Disease Applications: Stroke and Alzheimer's Disease

We apply BTECF to two UK Biobank cohorts spanning the vascular–neurodegenerative spectrum: ischemic stroke and Alzheimer's disease. The three-disease summary is in Appendix E (Table 9).

**Ischemic stroke** Spearman analysis on the 20-dimensional Bézier features yields nine significant features ($p < 0.05$), six of which align directionally with prior observations (Wong et al., 2001; McGeechan et al., 2009; Doubal et al., 2010). Applying the same parameter-level intervention (`tortuosity_4x`) on a stroke-retrained EfficientNet-B2 classifier reproduces a monotonic dose-response: mean stroke probability rises from a baseline of $0.001$ to $0.002$ at `tortuosity_1x`, $0.018$ at `tortuosity_2x`, and $0.338$ at `tortuosity_4x` ($\Delta = +0.337$; post-intervention mean $95\%$ CI $[+0.19, +0.49]$, $n = 17$, $p < 10^{-3}$). The `pixdrop_30` mean is $0.002$, approximately $200\times$ lower than `tortuosity_4x`, although this probability comparison is not a paired null-effect test. The monotonic intervention response demonstrates portability of the model-behavior probe to a challenging scorer (AUC $0.534$); disease-mechanism interpretation requires stronger discrimination. Full tables and cohort-design discussion are in Appendix E.1.

**Alzheimer's disease** The discriminative ranking inverts relative to DR (Bézier test AUC $0.646 >$ BiomedCLIP $0.454$), showing stronger discrimination by geometric features than by this image-feature baseline. Univariate Spearman is power-limited at this cohort scale, but `std_curvature` reaches significance with the largest Q5/Q1 odds ratio across the three diseases ($3.84$), alongside prior evidence of retinal vascular differences in AD (Frost et al., 2013). Because the measurements differ, the prior work supplies disease context rather than a directional replication. The tortuosity intervention yields a positive response ($\Delta = +0.19$ on `tortuosity_4x`, $n = 30$). The higher `pixdrop_10` mean ($0.323$ vs. $0.299$), together with scorer AUC $0.624$ and sensitivity $0.50$, identifies reduced response specificity in this AD application (Appendix E.2).

## 5 Discussion

BTECF connects an interpretable editing interface to a learned image prior and paired model-response measurement. Across all 500 DR starts, the strongest tortuosity operation produces a substantial response, and the $2 \times 2$ backbone–prompt analysis preserves its positive aggregate direction. Attenuation under DR prompting establishes prompt condition as an important evaluation variable and is consistent with overlap between prompt- and hint-induced visual content. The cross-disease applications further show that the same intervention interface can reveal graded scorer behavior under different disease settings, while scorer quality and control specificity determine how far that behavior supports disease-level interpretation.

**Limitations** The measured response belongs to the full edit–render–generate–classify pipeline. The next validation step is to remeasure target and non-target vessel geometry in generated images and quantify background and lesion preservation, complementing the present intensity/color screen, Grad-CAM analysis, and pixel-drop comparison. Deformation-matched controls, direct mask/skeleton baselines, and blinded expert realism ratings would distinguish target geometry from accompanying changes and test any Bézier-specific fidelity gain. The strict subsets trade coverage for reduced ceiling effects, and AD interpretation depends on a weaker scorer and control specificity (Appendix E.2); complete patient-identity separation and fold-specific preprocessing should also be verified. The L2 conditioning interface (Appendix G) makes this validation program operationally accessible before extending the results to biological causality.

**Broader Impacts** The interface supports research into disease-classifier sensitivity and failure modes. Generated-image responses should not guide diagnosis or treatment without independent clinical validation. The same capability could facilitate adversarial classifier probing, and visually plausible outputs or large probability shifts could be mistaken for biological evidence.

## 6 Conclusion

BTECF links editable vessel trees, diffusion conditioning, and paired scoring in a unified interface for probing retinal disease classifiers. The full DR results establish substantial responses to the strongest tortuosity edit and reveal systematic dependence on text conditioning. Cross-backbone and cross-disease analyses show how the same interface exposes both positive aggregate responses and disease-specific variation. These results support computational testing of geometry-related hypotheses and define the rendered-image validation needed to isolate geometric effects more precisely.

### AI USE STATEMENT

Generative AI tools were used to aid and polish the writing, for retrieval and discovery including finding related work, and to critique the manuscript and check consistency between claims and reported evidence. The authors reviewed all AI-assisted work and verified the scientific claims, citations, experimental design, analyses, and final manuscript.

### ETHICS STATEMENT

This work analyzes de-identified retinal images from public or controlled-access datasets and does not recruit or intervene on human participants. The generated counterfactuals are computational probes of model behavior and should not be interpreted as clinical evidence without prospective validation.

### REPRODUCIBILITY STATEMENT

The main text reports the data splits, leakage controls, model families, perturbation axes, and paired counterfactual estimators. The appendices provide the reported training and inference settings, filter thresholds, feature tables, and cross-disease deployment recipe. The Kaggle EyePACS/APTOS/Messidor redistribution (Appendix I) is used for diabetic retinopathy and UK Biobank cohorts (Appendix I) are used for stroke and Alzheimer's disease, subject to their respective access terms. Code, model weights, and evaluation scripts will be publicly released upon acceptance. The reported computation is approximately 100 GPU-hours per disease on RTX-5090-class hardware.

## REFERENCES

Lloyd Paul Aiello, Robert L Avery, Paul G Arrigg, Bruce A Keyt, Henry D Jampel, Sabera T Shah, Louis R Pasquale, Hagen Thieme, Mami A Iwamoto, John E Park, et al. Vascular endothelial growth factor in ocular fluid of patients with diabetic retinopathy and other retinal disorders. *New England Journal of Medicine*, 331(22):1480–1487, 1994.

David A. Antonetti, Ronald Klein, and Thomas W. Gardner. Diabetic retinopathy. *New England Journal of Medicine*, 366(13):1227–1239, 2012. doi: 10.1056/NEJMra1005073. URL `https://www.nejm.org/doi/full/10.1056/NEJMra1005073`.

Louis Arnould, Fabrice Meriaudeau, Charles Guenancia, Clement Germanese, Cécile Delcourt, Ryo Kawasaki, Carol Y Cheung, Catherine Creuzot-Garcher, and Andrzej Grzybowski. Using artificial intelligence to analyse the retinal vascular network: the future of cardiovascular risk assessment based on oculomics? a narrative review. *Ophthalmology and therapy*, 12(2):657–674, 2023.

Mohamed Ashraf, Siamak Shokrollahi, Alex U Pisig, Konstantina Sampani, Omar Abdelal, Jerry D Cavallerano, Gavin Robertson, Alan Fleming, Jano van Hemert, Cloyd M Pitoc, et al. Retinal vascular caliber association with nonperfusion and diabetic retinopathy severity depends on vascular caliber measurement location. *Ophthalmology Retina*, 5(6):571–579, 2021.

Pierre Bézier. Numerical control-mathematics and applications. *Translated by AR Forrest*, 1972.

Lorenza Bonaldi, Elisa Menti, Lucia Ballerini, Alfredo Ruggeri, and Emanuele Trucco. Automatic generation of synthetic retinal fundus images: vascular network. *Procedia Computer Science*, 90: 54–60, 2016.

Rebecca Broe, Malin L Rasmussen, Ulrik Frydkjaer-Olsen, Birthe S Olsen, Henrik B Mortensen, Tunde Peto, and Jakob Grauslund. Retinal vascular fractals predict long-term microvascular complications in type 1 diabetes mellitus: the danish cohort of pediatric diabetes 1987 (dcpd1987). *Diabetologia*, 57(10):2215–2221, 2014.

Dario Lo Castro, Domenico Tegolo, and Cesare Valenti. A visual framework to create photorealistic retinal vessels for diagnosis purposes. *Journal of Biomedical Informatics*, 108:103490, 2020.

Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, Matthew P. Lungren, Shaoting Zhang, Lei Xing, Le Lu, Alan Yuille, and Yuyin Zhou. Transunet: Rethinking the u-net architecture design for medical image

segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024. doi: 10.1016/j.media.2024.103280.

Ning Cheung, Sophie L Rogers, Kim C Donaghue, Alicia J Jenkins, Gabriella Tikellis, and Tien Yin Wong. Retinal arteriolar dilation predicts retinopathy in adolescents with type 1 diabetes. *Diabetes care*, 31(9):1842–1846, 2008.

David G Cogan, Daniel Toussaint, and Toichiro Kuwabara. Retinal vascular patterns: Iv. diabetic retinopathy. *Archives of ophthalmology*, 66(3):366–378, 1961.

TM Curtis, TA Gardiner, and AW Stitt. Microvascular lesions of diabetic retinopathy: clues towards understanding pathogenesis? *Eye*, 23(7):1496–1508, 2009.

Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, John-Richard Ordóñez-Varela, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, pp. 231–234, 2014.

FN Doubal, TJ MacGillivray, N Patton, B Dhillon, MS Dennis, and JM Wardlaw. Fractal analysis of retinal vessels suggests that a distinct vasculopathy causes lacunar stroke. *Neurology*, 74(14): 1102–1107, 2010.

Emma Dugas, Jared, Jorge, and Will Cukierski. Diabetic retinopathy detection. `https://kaggle.com/competitions/diabetic-retinopathy-detection`, 2015. Kaggle.

Fariztah Sukainah Nur Fathimah, Sauli Ari Widjaja, Wimbo Sasono, Ima Yustiarini, Muhammad Firmansjah, Ady Dwi Prakosa, Aulia Kezia Mulyazhara, and Soebagijo Adi Soelistijo. Retinal vessel tortuosity and fractal dimension in diabetic retinopathy. *International Journal of Retina and Vitreous*, 11(1):64, 2025.

Xiaoyi Feng, Minqing Zhang, Mengxian He, Mengdi Gao, Hao Wei, and Wu Yuan. Diversified and structure-realistic fundus image synthesis for diabetic retinopathy lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 77–86. Springer, 2024.

Rachel B Forster, Emmanuel Sandoval Garcia, Anniek J Sluiman, Sheila M Grecian, Stela McLachlan, Tom J MacGillivray, Mark WJ Strachan, Jackie F Price, and Edinburgh Type 2 Diabetes Study (ET2DS) investigators. Retinal venular tortuosity and fractal dimension predict incident retinopathy in adults with type 2 diabetes: the edinburgh type 2 diabetes study. *Diabetologia*, 64 (5):1103–1112, 2021.

Shaun Frost, Yogesan Kanagasingam, Hamid R. Sohrabi, Jothy Vignarajan, Pierrick Bourgeat, Olivier Salvado, Victor Villemagne, Christopher C. Rowe, S. Lance Macaulay, Cassandra Szoeke, Kathryn A. Ellis, David Ames, Colin L. Masters, Stephanie R. Rainey-Smith, and Ralph N. Martins. Retinal vascular biomarkers for early detection and monitoring of alzheimer's disease. *Translational Psychiatry*, 3:e233, 2013. doi: 10.1038/tp.2012.150.

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

Xiaoxin Guo, Xiang Li, Qifeng Lin, Guangyu Li, Xiaoying Hu, and Songtian Che. Controllable fundus image generation based on conditional generative adversarial networks with mask guidance. *Multimedia Tools and Applications*, 83(15):46065–46085, 2024.

Hans-Peter Hammes, Jihong Lin, Oliver Renner, Moshe Shani, Andrea Lundqvist, Christer Betsholtz, Michael Brownlee, and Urban Deutsch. Pericytes and the pathogenesis of diabetic retinopathy. *Diabetes*, 51(10):3107–3112, 2002.

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

Adam Hoover, Valentina Kouznetsova, and Michael Goldbaum. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical Imaging*, 19(3):203–210, 2000. doi: 10.1109/42.845178.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

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

Saeed Jamshidiha, Alireza Rezaee, Farshid Hajati, Mojtaba Golzan, and Raymond Chiong. An explainable transformer model for alzheimer's disease detection using retinal imaging. *Scientific Reports*, 15:26773, 2025. doi: 10.1038/s41598-025-12498-2.

Karthik, Maggie, and Sohier Dane. Aptos 2019 blindness detection. `https://kaggle.com/competitions/aptos2019-blindness-detection`, 2019. Kaggle.

Parsa Khalafi, Soroush Morsali, Sana Hamidi, Hamidreza Ashayeri, Navid Sobhi, Siamak Pedrammehr, and Ali Jafarizadeh. Artificial intelligence in stroke risk assessment and management via retinal imaging. *Frontiers in Computational Neuroscience*, 19:1490603, 2025. doi: 10.3389/fncom.2025.1490603.

Ronald Klein, Barbara EK Klein, Scot E Moss, Tien Y Wong, Larry Hubbard, Karen J Cruickshanks, and Mari Palta. The relation of retinal vessel caliber to the incidence and progressionof diabetic retinopathy: Xix: The wisconsin epidemiologic study of diabetic retinopathy. *Archives of ophthalmology*, 122(1):76–83, 2004.

Sophie Lemmens, Annelies Devulder, Karel Van Keer, Joris Bierkens, Patrick De Boever, and Ingeborg Stalmans. Systematic review on fractal dimension of the retinal vasculature in neurodegeneration and stroke: assessment of a potential biomarker. *Frontiers in Neuroscience*, 14:16, 2020. doi: 10.3389/fnins.2020.00016.

Yuanyuan Liu, Dazhi Yao, Yongwen Ma, Hua Wang, Jinming Wang, Xuefeng Bai, Guang Zeng, and Yuejuan Liu. Stmf-drnet: A multi-branch fine-grained classification model for diabetic retinopathy using swin-transformerv2. *Biomedical Signal Processing and Control*, 103:107352, 2025. doi: 10.1016/j.bspc.2024.107352.

Kevin McGeechan, Gerald Liew, Petra Macaskill, Les Irwig, Ronald Klein, Barbara EK Klein, Jie Jin Wang, Paul Mitchell, Johannes R Vingerling, Paulus TVM De Jong, et al. Prediction of incident stroke events based on retinal vessel caliber: a systematic review and individual-participant meta-analysis. *American journal of epidemiology*, 170(11):1323–1332, 2009.

Ketan Mehta, Bengt Ljungquist, James Ogden, Sumit Nanda, Ruben G Ascoli, Lydia Ng, and Giorgio A Ascoli. Online conversion of reconstructed neural morphologies into standardized swc format. *Nature Communications*, 14(1):7429, 2023.

Luyu Niu, Xiaojuan Chen, and Song Guo. Bézier curves meet deep learning: A novel pretraining method for improved and generalized retinal vessel segmentation. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2159–2164. IEEE, 2023.

Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. *Advances in neural information processing systems*, 33:857–869, 2020.

Natasa Popovic, Mirko Lipovac, Miroslav Radunovic, Jurgi Ugarte, Erik Isusquiza, Andoni Beristain, Ramón Moreno, Nerea Aranjuelo, and Tomo Popovic. Fractal characterization of retinal microvascular network morphology during diabetic retinopathy progression. *Microcirculation*, 26 (4):e12531, 2019.

Varsha Pramil, Emily S Levine, and Nadia K Waheed. Macular vessel density in diabetic retinopathy patients: how can we accurately measure and what can it tell us? *Clinical Ophthalmology*, pp. 1517–1527, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *arXiv preprint arXiv:1505.04597*, 2015. doi: 10.48550/arXiv.1505.04597.

Mélanie Roschewitz, Fabio De Sousa Ribeiro, Tian Xia, Galvin Khara, and Ben Glocker. Robust image representations with counterfactual contrastive learning. *Medical Image Analysis*, 105: 103668, 2025.

Sayon Roy and Dongjoon Kim. Retinal capillary basement membrane thickening: Role in the pathogenesis of diabetic retinopathy. *Progress in Retinal and Eye Research*, 82:100903, 2021.

Emmanuel Sandoval-Garcia, Stela McLachlan, Anna H Price, Thomas J MacGillivray, Mark WJ Strachan, James F Wilson, and Jackie F Price. Retinal arteriolar tortuosity and fractal dimension are associated with long-term cardiovascular outcomes in people with type 2 diabetes. *Diabetologia*, 64(10):2215–2227, 2021.

MB Sasongko, TY Wong, TT Nguyen, CY Cheung, JE Shaw, and JJ Wang. Retinal vascular tortuosity in persons with diabetes and diabetic retinopathy. *Diabetologia*, 54(9):2409–2416, 2011.

Muhammad Bayu Sasongko, Tien Yin Wong, Thanh T Nguyen, Carol Y Cheung, Jonathan E Shaw, Ryo Kawasaki, Ecosse L Lamoureux, and Jie Jin Wang. Retinal vessel tortuosity and its relation to traditional and novel vascular risk markers in persons with diabetes. *Current eye research*, 41(4): 551–557, 2016.

João V. B. Soares, Jorge J. G. Leandro, Roberto M. Cesar Jr., Herbert F. Jelinek, and Michael J. Cree. Retinal vessel segmentation using the 2-d gabor wavelet and supervised classification. *IEEE Transactions on Medical Imaging*, 25(9):1214–1222, 2006. doi: 10.1109/TMI.2006.879967.

Alan W Stitt, Timothy M Curtis, Mei Chen, Reinhold J Medina, Gareth J McKay, Alicia Jenkins, Thomas A Gardiner, Timothy J Lyons, Hans-Peter Hammes, Rafael Simo, et al. The progress in understanding and treatment of diabetic retinopathy. *Progress in retinal and eye research*, 51: 156–186, 2016.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):e1001779, 2015. doi: 10.1371/journal.pmed.1001779.

Vallimayil Velayutham, Maria E Craig, Gerald Liew, Tien Y Wong, Alicia J Jenkins, Paul Z Benitez-Aguirre, and Kim C Donaghue. Extended-zone retinal vascular caliber and risk of diabetic retinopathy in adolescents with type 1 diabetes. *Ophthalmology Retina*, 4(12):1151–1157, 2020.

Hao Wei, Peilun Shi, Juzheng Miao, Mingqin Zhang, Guitao Bai, Jianing Qiu, Furui Liu, and Wu Yuan. Caudr: A causality-inspired domain generalization framework for fundus-based diabetic retinopathy grading. *Computers in Biology and Medicine*, 175:108459, 2024.

Tien Yin Wong, Ronald Klein, David J Couper, Lawton S Cooper, Eyal Shahar, Larry D Hubbard, Marion R Wofford, and A Richey Sharrett. Retinal microvascular abnormalities and incident stroke: the atherosclerosis risk in communities study. *The Lancet*, 358(9288):1134–1140, 2001.

Tien Yin Wong, Ronald Klein, A Richey Sharrett, Bruce B Duncan, David J Couper, Barbara EK Klein, Larry D Hubbard, F Javier Nieto, and Atherosclerosis Risk in Communities Study. Retinal arteriolar diameter and risk for hypertension. *Annals of internal medicine*, 140(4):248–255, 2004.

Yi Xu, Yuting Wu, Saiguang Ling, Zhou Dong, Xin Ke, Lina Lu, Zheng Ye, Jianling Song, and Haidong Zou. Quantitative evaluation of retinal vascular morphology based on the human visual bionic mechanism for the evaluation of diabetic retinopathy onset. *Microvascular Research*, pp. 104901, 2026.

Mayinuer Yusufu, David S Friedman, Mengtian Kang, Ambhruni Padhye, Xianwen Shang, Lei Zhang, Danli Shi, and Mingguang He. Retinal vascular fingerprints predict incident stroke: findings from the uk biobank cohort study. *Heart*, 111(7):306–313, 2025.

Jie Zhang, Lei Shi, and Yong Shen. The retina: A window in which to view the pathogenesis of alzheimer's disease. *Ageing Research Reviews*, 77:101590, 2022. doi: 10.1016/j.arr.2022.101590.

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

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023b.

B. Zhao, Y. Li, Z. Fan, and et al. Eye-brain connections revealed by multimodal retinal and brain imaging genetics. *Nature Communications*, 15:6064, 2024. doi: 10.1038/s41467-024-50309-w.

Yukun Zhou, Siegfried K Wagner, Mark A Chia, An Zhao, Moucheng Xu, Robbert Struyven, Daniel C Alexander, Pearse A Keane, et al. Automorph: automated retinal vascular morphology quantification via a deep learning pipeline. *Translational vision science & technology*, 11(7):12–12, 2022.

# Technical Appendices and Supplementary Material

## A Framework Modules and Cross-Disease Deployment Recipe

The six modules of BTECF are summarized in Table 6. Modules 1, 2, and 6 (vessel segmentation, Bézier parameterization, counterfactual scoring) are disease-agnostic and transfer with zero retraining: they reuse the released code and weights without modification. Modules 3, 4, and 5 (LoRA, ControlNet, classifier) are disease-specific and require retraining on the new disease's fundus cohort. LoRA and ControlNet must be co-trained in the same batch with matched `noise_offset = 0.1`; mixing an old LoRA with a freshly trained ControlNet produces persistent grayscale or blue-white color-drift artifacts at inference. Geometric feature extraction uses the disease-agnostic modules without retraining; generated-image intervention and scoring use the full six-module pipeline.

Table 6: The six modules of BTECF, grouped by cross-disease transferability. Modules 1, 2, and 6 transfer with zero retraining; modules 3, 4, and 5 require per-disease retraining.

| # | Module | Role | Training-domain dependency | Cross-disease transferability |
|---|---|---|---|---|
| 1 | AutoMorph vessel segmentation (Zhou et al., 2022) | Fundus to binary vessel mask | General vessel segmentation | Disease-agnostic, zero retrain; OOD risk on new cameras |
| 2 | Bézier parameterization | Mask to control points, geometric features, three-channel raster hint | No learned parameters | Disease-agnostic, zero retrain |
| 3 | LoRA domain adapter on SD 2.1 | Adapts the diffusion backbone to the cohort visual style | Target-cohort fundus images (no labels required) | Per-cohort retrain when camera or acquisition shifts |
| 4 | VesselControlNet (Bézier-conditioned) | Bézier hint to disease-conditioned fundus | Mask, label, and fundus | Per-disease retrain; the core counterfactual probe |
| 5 | Downstream disease classifier | Scores counterfactually generated fundi | Fundus and label pairs | Per-disease retrain |
| 6 | Counterfactual scoring protocol | Within-start paired $\Delta$ and statistics | No learned parameters | Disease-agnostic |

## B Implementation Details

### B.1 Bézier Hint Pipeline

Given a binary mask $M \in \{0,1\}^{H\times W}$, the hint is constructed in three steps.

**Skeletonization** Morphological skeletonization produces a 1-pixel-wide topological skeleton, discarding caliber while preserving connectivity.

**Bézier fitting** The skeleton is first broken at degree-1 endpoints and degree-$\geq 3$ branch nodes, producing a set of ordered open polylines (typically 10–30 per fundus). Each polyline is then chopped into overlapping chunks of 30 ordered skeleton points (4-point overlap between adjacent chunks for endpoint continuity), and each chunk is fit by least squares to a cubic Bézier with arc-length parameterization within the chunk:

$$B(t) = (1-t)^3 P_0 + 3(1-t)^2 t P_1 + 3(1-t) t^2 P_2 + t^3 P_3, \qquad t \in [0,1], \tag{4}$$

with $(P_0, P_3)$ clamped to the chunk endpoints and $(P_1, P_2)$ the interior control points fitted by least squares. A typical fundus produces between 200 and 500 such cubic segments. Per-segment fitting examples are shown in the bottom two rows of Figure 2.

**Fusion rendering into three channels (Figure 2)** Channel 0 is the Euclidean distance-transform radius field of the raw mask, normalized to $[0, 1]$. Channel 1 is the Bézier variable-radius render: each Bézier segment is sampled adaptively at twice the polyline chord length (with a floor of 20 samples per segment, yielding sub-pixel coverage at $512 \times 512$), and each sample point is dilated by the local radius read from the distance-transform field. Channel 2 is a Gaussian-smoothed version of Channel 1 with kernel size $7 \times 7$ and standard deviation $\sigma = 2.0$. The three channels are stacked into an RGB hint of resolution $512 \times 512$ and rescaled to $[-1, 1]$ before being fed to ControlNet.

**Channel 0 field stability** Tortuosity and arc-drop operate on control points or segment presence rather than intentionally rescaling Channel 0. Empirically, the mean of Channel 0 differs between baseline and tortuosity-amplified hints by less than $0.022$ on the $[0, 1]$ scale. This check establishes stability of the field itself. Because moving a curve changes the coordinates at which local radii are read, rendered caliber and post-generation caliber require separate measurements. Figure 3 illustrates the field comparison on a representative grade-0 fundus.

**Construction constraints** The tortuosity perturbation preserves endpoints $P_0$ and $P_3$, retaining the encoded branch connections, while rasterization can still introduce new crossings or sharp local bends. The displacement coefficient $\gamma = 0.15$ scales displacement with chord length; its `tortuosity_4x` setting has not been calibrated to a clinical tortuosity distribution. Channel 1 draws local radii from Channel 0 along the perturbed curves, merges overlapping rendered segments, and Channel 2 applies Gaussian smoothing (kernel $7 \times 7$, $\sigma = 2.0$). These construction constraints define the operator but do not by themselves validate non-target geometry or membership in the training distribution.

### B.2 Training Details

**LoRA finetuning** We start from the Stable Diffusion 2.1 base and finetune the UNet with a LoRA adapter of rank 32, trained for 36,000 steps with AdamW (learning rate $1 \times 10^{-4}$), batch size 8, image size $512 \times 512$. The adapter is applied to the attention and cross-attention projection weights of all transformer blocks. We enable `noise_offset = 0.1`: a per-channel spatially-constant offset added to the Gaussian noise during training. In our fundus finetuning runs, omitting this offset made the UNet more prone to gray or bluish color drift rather than the expected red-orange-brown fundus tone. Co-training the LoRA `noise_offset` with the ControlNet target in the same batch is required to maintain this calibration; an old LoRA paired with a freshly trained ControlNet reintroduces the drift. The same coupling rule applies under cross-disease retraining.

**VesselControlNet** Following the original ControlNet paradigm (Zhang et al., 2023a), we duplicate the SD 2.1 UNet encoder, attach zero-convolution residuals (convolution layers initialized to zero so the conditioning signal starts from the exact pretrained UNet and gradually leaks in residuals during training), and inject these residuals back into the corresponding UNet decoder blocks. The checkpoint takes the parametric Bézier hint described above as its conditioning input and is used by the counterfactual intervention.

**Prompt pairing during training (DR case)** For each training sample we draw a text prompt uniformly at random from three paraphrases of the sample's class:

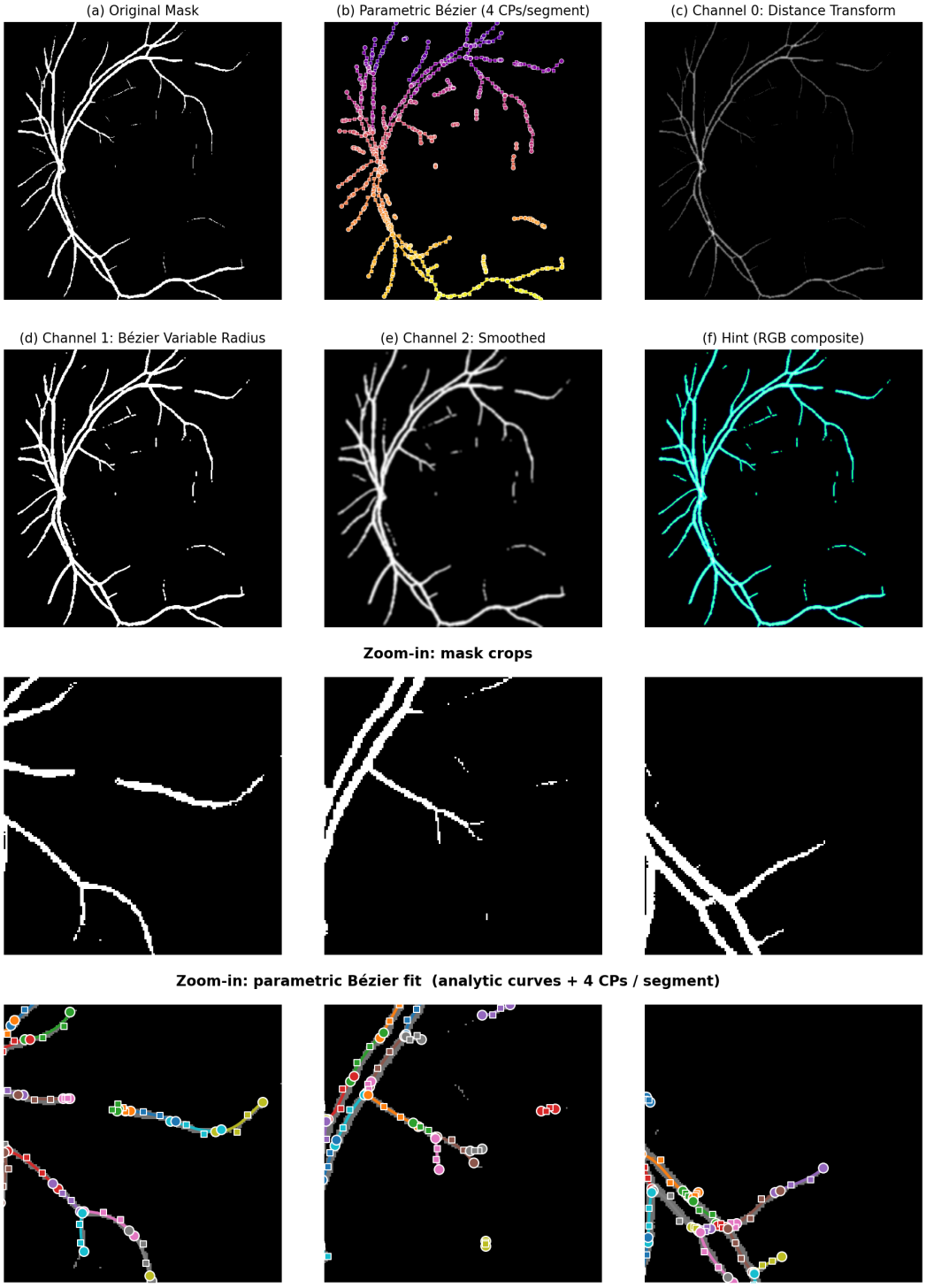


Figure 2: Decomposition of the three-channel Bézier hint derived from the binary vessel mask. **Top section, panels (a)–(f):** Channel 0 is the distance-transform radius field, Channel 1 is the Bézier variable-radius render, and Channel 2 is the Gaussian-smoothed version of Channel 1. The three channels are stacked into the final RGB hint passed to ControlNet. **Bottom two rows:** zoom-in examples cropped from the same source mask, illustrating per-segment fitting. The mask-crop row shows three local regions of the binary mask; the parametric-fit row overlays the cubic-Bézier reconstruction on the same crops, with each color denoting one fitted segment. Within a single color, the four nearest markers (two circles for endpoints $P_0, P_3$ and two squares for interior control points $P_1, P_2$) form the four control points of one minimum-unit cubic Bézier.

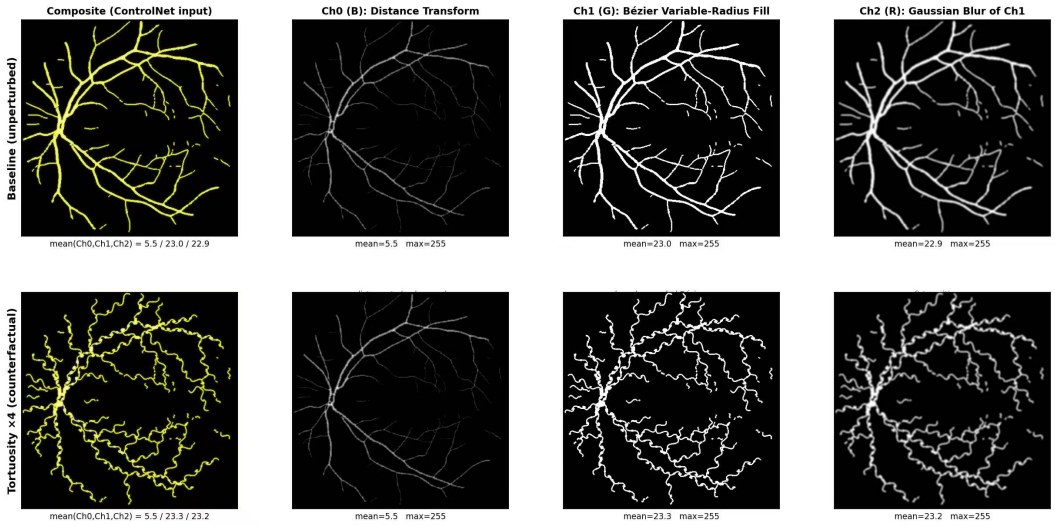


Figure 3: Three-channel Bézier hint decomposition on a representative grade-0 fundus. Top row: baseline (unperturbed). Bottom row: `tortuosity_4x` counterfactual. Channel 0 (distance-transform radius field) has identical mean (5.5 on the 0–255 scale) across both rows, illustrating equal field means in this example, not equal rendered widths; Channels 1 and 2 visibly distort with the tortuosity perturbation.

- Grade 0 (normal):
  - `"a normal fundus photograph without diabetic retinopathy"`
  - `"a fundus image of a healthy retina with no lesions"`
  - `"a retinal photograph showing no signs of diabetic retinopathy"`
- Grades 1–4 (DR):
  - `"a fundus photograph with diabetic retinopathy showing microaneurysms and hemorrhages"`
  - `"a fundus image showing signs of diabetic retinopathy with retinal lesions"`
  - `"a retinal photograph with diabetic retinopathy featuring exudates and hemorrhages"`

The same prompt family supplies prompted inference. The current record specifies the family but not whether each baseline–edit pair reuses the exact paraphrase.

**BiomedCLIP encoder for discriminative features** The discriminative-evaluation features come from a BiomedCLIP encoder (Zhang et al., 2023b) (PubMedBERT + ViT-B/16) fine-tuned on disease labels of the target cohort and then frozen for feature extraction. Under cross-disease transfer, fine-tuning uses the target-disease labels before the encoder is frozen. Reproducing the discriminative benchmark additionally requires the fine-tuning sample membership, PCA fitting split, and fold-specific preprocessing boundaries, which are not recorded in the current report.

### B.3 Perturbation-Axis Selection

The four selection grounds in §3.4 each constrain the candidate axis set; their intersection picks tortuosity, arc length, and radius from the 20-dimensional Bézier feature space. We expand each ground in turn below.

*Observational analysis.* The three perturbed axes all sit within the SHAP top-5 features of the deduplicated DR cohort (Appendix C.2, Table 8): #1 `arc_length`, #3 `radius`, #5 `tortuosity`.

The two skipped top-5 features (#2 `branching_density` and #4 `mean_chord_length`) are both not selected as separate operations for the reasons below.

*Explicit editing operations.* The interface provides separate operations targeting shape, segment completeness, and thickness; these need not change only one derived feature. `tortuosity` displaces $P_1, P_2$ while fixing endpoints, `arc_drop` deletes fitted segments, and `radius_x` scales `dist_map`. Curve displacement can change arc length, curvature, and sampled widths; deletion can change branching and connectivity. Consequently, features in the 20-dimensional set can covary. We do not add `branching_density` as a separate operator because branch removal also deletes segments. Likewise, `mean_chord_length`, `mean_curvature`, `n_segments`, and `min/max_radius` are derived summaries, not additional independent controls in the present interface.

*Literature motivation across diseases.* For tortuosity, Forster et al. (2021) (T2D, $n = 718$, venular tortuosity to incident DR, OR = 1.51) provides DR-specific observational motivation. Frost et al. (2013) instead report lower venular tortuosity in AD; their vessel-specific curvature measure is not interchangeable with our pooled `std_curvature` feature or the control-point edit. Lemmens et al. (2020) review fractal dimension in neurodegeneration and stroke, supporting complexity-related motivation rather than a common positive tortuosity direction. Sandoval-Garcia et al. (2021) further links arteriolar tortuosity and fractal dimension to long-term cardiovascular outcomes in type 2 diabetes. For `arc_drop` as a fractal-dimension proxy, Forster et al. (2021) (venular FD OR = 0.75), Sandoval-Garcia et al. (2021) (arteriolar FD HR = 0.73), Doubal et al. (2010) (decreased FD to lacunar stroke), and the systematic review by Lemmens et al. (2020) all support reduced retinal vascular complexity as a neurovascular disease marker. For `radius_x`, Klein et al. (2004), Cheung et al. (2008), Velayutham et al. (2020), and Xu et al. (2026) report early dilation in DR; Wong et al. (2004) and Wong et al. (2001) report arteriolar narrowing for hypertension and incident stroke. The opposite directions across diseases motivate why the intervention effect must be scored by a disease-specific classifier.

*Dose-response quantifiability.* Each axis admits three graded perturbation strengths (`tort 1x / 2x / 4x`; `arc_drop` doses $10/20/30\%$; `radius_x 0.85 / 0.70 / 0.55`), enabling explicit dose-response analyses.

### B.4 Downstream Classifier Hyperparameters

Three backbones are trained with identical hyperparameters: 15 epochs, batch 64, AdamW, learning rate $3 \times 10^{-4}$, cosine schedule, seed 42, $384 \times 384$ ImageNet-normalized input. EfficientNet-B2 @ 384 (test AUC 0.9612, sens 0.8858) is the primary scorer; ResNet-50 (AUC 0.9545, sens 0.8320) is the cross-backbone sensitivity check; ViT-B/16 (AUC 0.8295, sens 0.5553) underperforms at the current training scale, and we treat its counterfactual-intervention output as diagnostic rather than confirmatory. The stroke and AD EfficientNet-B2 heads are retrained from scratch on their respective cohorts with the same hyperparameter family; numbers are in §4.5.

### B.5 Inference Hyperparameters and Filter Thresholds

A single fixed set of sampling hyperparameters is used across all counterfactual intervention experiments to standardize the sampling setup for within-start comparisons; the guidance-weight ablation (§4.4.3) sweeps CFG weight $w$ as the sole exception. Table 7 summarizes the values.

Table 7: Inference hyperparameters used uniformly across the counterfactual intervention experiments.

| Parameter | Value | Role |
|---|---|---|
| DDIM steps | 50 | Sampling schedule |
| CFG `guidance_scale` | 7.5 | Text-conditioning strength |
| ControlNet `cn_scale` | 1.0 | Residual strength |
| `mid_scale` | 0.5 | Mid-block residual strength |
| `latent_offset` | 0.3 | Initial latent bias toward real-fundus latent mean |
| LoRA weights | v2 (`noise_offset = 0.1`) | Color-calibrated adapter |

**Baseline intensity/color screening** The baseline screen referenced in §3.5 targets gross intensity and color failures rather than anatomical fidelity. A start is retained only if its baseline generation satisfies all three conditions: (i) mean pixel intensity in $[50, 170]$ on the 0–255 scale, rejecting near-black collapses (mean $< 50$) and saturated bright-white collapses (mean $> 170$); (ii) pixel standard deviation $> 25$, rejecting monochromatic collapses where the entire image converges to a single tone; and (iii) red-to-green channel ratio $> 1.3$, enforcing the red-orange-brown chromaticity of healthy fundi and rejecting gray or bluish drift. On the $n = 500$ DR cohort, 350 of 500 starts pass the intensity/color filter; further restricting to baseline DR probability $< 0.3$ yields $n = 190$ under DR-prompt and $n = 97$ under Uncond. Post-intervention vessel fidelity, background preservation, and lesion consistency require separate measurements.

### B.6 Dataset Deduplication

The documented deduplication uses two base-ID operations:

- **Train/test base-ID exclusion.** From the train split, we remove 1,032 rows whose `base_id` appears in the test split, eliminating shared base IDs across the train/test boundary.
- **In-split augmentation collapse.** The manual augmentation of the redistribution produces multiple synthetic copies that share the same `base_id`. Within val and test, we collapse these into a single representative per base ID (958 rows removed from val and 964 from test), removing duplicate augmented copies within each evaluation split.

Together the two operations remove $1{,}032 + 958 + 964 = 2{,}954$ rows from the original 143,669-image redistribution, yielding the final splits $114{,}209/13{,}269/13{,}237$ (train / val / test) reported in §4.1. This establishes train/test base-ID separation and removes repeated augmented copies within each evaluation split. The recorded base IDs do not document train/val or val/test patient exclusion, bilateral-image linkage, or every cross-source preprocessing variant, so those aspects of patient-level independence require separate verification.

## C Extended Empirical Analyses

### C.1 Generator Bias Controls and the Counterfactual Probe

The paired design and pixel-drop comparison provide two complementary checks on generator bias in the counterfactual measurement. Their scope is defined below.

**Pilot pixel-drop comparison** In the earlier pilot analysis, `pixdrop_30` yields mean post-intervention DR probabilities of 0.004 (DR-prompt) / 0.054 (Uncond), 16 to 100 times lower than `tortuosity_4x`. These pilot ratios compare mean probabilities and are separate from the main-cohort paired estimates. They show a substantially larger response to the tested tortuosity operation than to mask-pixel deletion, while a paired difference-of-effects test would quantify this contrast directly.

**Style versus content** The within-start paired $\Delta$ design compares outputs for the same starting fundus, generator, and prompt condition under baseline and edited hints, cancelling additive bias shared by the two generations. The current record does not specify exact prompt-string or latent-randomness reuse within each pair, and perturbation-dependent tone, texture, or lesion changes may remain in $\Delta$. The resulting estimate therefore captures the pipeline response under the reported sampling protocol, complemented by pixel-drop as an alternative perturbation (Appendix D).

**Pixel-drop contrast** In the DR-prompt pilot, the mean-probability contrast among Baseline (0.097), `tortuosity_4x` (0.437), and `pixdrop_30` (0.004) supports a larger response to the tested tortuosity operation than to pixel-drop. Attribution beyond this tested contrast requires direct measurement of the generated geometry and accompanying image changes.

### C.2 Observational Analysis: Full 20-Feature Table

Table 8 reports the complete 20-dimensional Bézier observational analysis on the DR cohort ($n = 26{,}506$). Here, "SHAP non-linear" flags opposite signs of marginal correlation and conditional attribution; it is not a test establishing a non-monotonic risk curve.

Table 8: Full 20-feature Bézier observational analysis on the DR cohort, ordered by SHAP importance. Metrics defined in §3.5; *SHAP non-linear* flags features failing the Spearman–SHAP sign-disagreement criterion. Significance: $***\ p < 0.001$, $*\ p < 0.05$.

| Rank | Feature | Spearman $\rho$ | feat-SHAP corr | OR/SD | Q5/Q1 OR | SHAP abs | Direction |
|---|---|---|---|---|---|---|---|
| 1 | `total_arc_length` | $-0.358$*** | $-0.94$ | 0.55 | 0.20 | 0.456 | ↓ DR |
| 2 | `branching_density` | $+0.014$* | $-0.55$ | 1.21 | 1.09 | 0.364 | SHAP non-linear |
| 3 | `mean_radius` | $-0.084$*** | $-0.90$ | 0.85 | 0.68 | 0.285 | ↓ DR |
| 4 | `mean_chord_length` | $-0.079$*** | $-0.84$ | 0.80 | 0.67 | 0.231 | ↓ DR |
| 5 | `mean_tortuosity` | $+0.086$*** | $+0.17$ | 1.00 | 1.60 | 0.160 | ↑ DR |
| 6 | `mean_segment_length` | $-0.077$*** | $-0.81$ | 0.80 | 0.67 | 0.141 | ↓ DR |
| 7 | `radius_cv` | $+0.007$ n.s. | $+0.74$ | 0.94 | 1.29 | 0.135 | ↑ DR |
| 8 | `std_radius` | $-0.028$*** | $+0.30$ | 0.91 | 1.04 | 0.102 | SHAP non-linear |
| 9 | `std_chord_length` | $+0.087$*** | $-0.49$ | 0.99 | 1.53 | 0.071 | SHAP non-linear |
| 10 | `std_tortuosity` | $+0.078$*** | $-0.51$ | 1.17 | 1.55 | 0.063 | SHAP non-linear |
| 11 | `max_radius` | $-0.120$*** | $-0.68$ | 0.84 | 0.62 | 0.062 | ↓ DR |
| 12 | `thick_vessel_ratio` | $-0.033$*** | $+0.20$ | 1.06 | 0.96 | 0.050 | SHAP non-linear |
| 13 | `mean_curvature` | $+0.091$*** | $-0.64$ | 1.22 | 1.65 | 0.049 | SHAP non-linear |
| 14 | `std_segment_length` | $+0.087$*** | $+0.35$ | 0.99 | 1.55 | 0.046 | ↑ DR |
| 15 | `n_segments` | $-0.338$*** | $-0.57$ | 0.58 | 0.23 | 0.035 | ↓ DR |
| 16 | `coverage_ratio` | $-0.315$*** | $+0.39$ | 0.59 | 0.25 | 0.034 | SHAP non-linear |
| 17 | `std_curvature` | $+0.102$*** | $-0.18$ | 1.21 | 1.72 | 0.028 | SHAP non-linear |
| 18 | `max_tortuosity` | $-0.026$*** | $+0.37$ | 1.08 | 0.98 | 0.026 | SHAP non-linear |
| 19 | `max_curvature` | $-0.043$*** | $+0.29$ | 1.04 | 1.01 | 0.025 | SHAP non-linear |
| 20 | `min_radius` | $+0.123$*** | $+0.83$ | 1.17 | **2.55** | 0.005 | ↑ DR |

## D Full Counterfactual Intervention Causal Table

Figure 4 visualizes mean DR probability across all 13 perturbation configurations under DR-prompt and Uncond; Table 3 reports the corresponding paired $\Delta$ values. Five patterns recur. (i) The strongest tortuosity dose yields the largest effect under both prompt conditions, with the Uncond response $2.0\times$ stronger than the DR-prompt response, visible as the steeply rising tortuosity curve in Figure 4. (ii) The mid-dose tortuosity response differs between the two prompt conditions. (iii) Arc-drop produces small mean changes at all three doses relative to the strongest tortuosity edit. (iv) Radius scaling yields a small consistent negative $\Delta$ under both conditions. (v) Pixel-drop yields negative mean paired changes at all three doses and mean post-intervention DR probabilities $\sim 18\times$ / $25\times$ lower than `tortuosity_4x` under DR-prompt/Uncond, respectively. These probability ratios complement, rather than replace, the paired changes and become large as control probabilities approach zero. Figure 4 shows the `pixdrop` trace saturating near zero throughout.

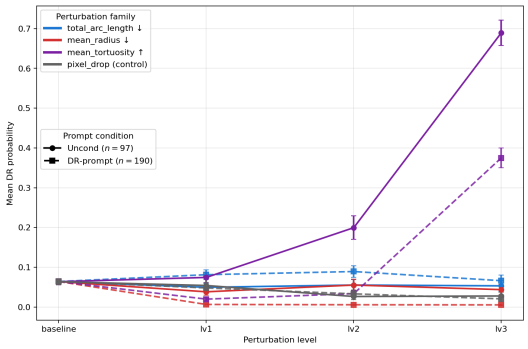


Figure 4: Classifier-response curves: mean DR probability across all 13 perturbation configurations under DR-prompt and Uncond. Mean probabilities increase across the three nonzero tortuosity doses; `pixdrop` remains low. Numerical values for all configurations are in Table 3.

**Statistical consistency check for pixel-drop** For `pixdrop_30` under DR prompting, $n = 190$, mean baseline probability $0.064$, and mean paired change $-0.044$ imply a mean edited probability of $0.020$. Let $b_i$ and $q_i$ denote paired baseline and edited probabilities and $d_i = q_i - b_i$. Since $0 \le b_i < 0.3$ and $0 \le q_i \le 1$, we have $d_i^2 \le q_i + 0.3b_i$ and $\mathrm{SE}(\bar{d})^2 \le (\bar{q} + 0.3\bar{b} - \bar{d}^2)/(n-1)$. Allowing each reported mean to vary by half a unit in its last decimal place gives a conservative unadjusted two-sided $t$-test bound $p < 0.003$. Conditional on the stated common subset and test, this bound establishes that the negative control response should not be summarized as nonsignificant. Exact empirical $p$-values and paired tortuosity-minus-control confidence intervals require the paired scores.

# E Cross-Disease Summary

Table 9 compares the framework's three diseases (DR primary case, ischemic stroke, Alzheimer's disease) on cohort scale, observational top signals, discriminative AUC across feature sets, and the strongest counterfactual effect. The reported effects follow DR > stroke > AD when using DR Uncond and disease-prompt results for the other cohorts; differing conditions preclude attributing this ordering to cohort scale alone. The AD cohort additionally inverts the discriminative ranking (Bézier > CLIP), showing that geometric features outperform this image-feature baseline in the AD cohort. The control-to-tortuosity ratios compare post-intervention mean probabilities within each reported condition, not paired changes.

Table 9: Cross-disease summary across the three diseases. Image-level AUC refers to a disease-specific EfficientNet-B2 classifier trained end-to-end. The counterfactual row reports $\Delta$ on `tortuosity_4x` on the strict reverse-counterfactual subset (baseline disease probability $< 0.3$).

| | DR (primary) | Ischemic Stroke | Alzheimer's Disease |
|---|---|---|---|
| *Cohort and prevalence* | | | |
| Eval cohort $n$ | 26,506 | 3,409 | 550 |
| Positive prevalence | $\sim$49% (grades 1–4) | 12% | 7.5% |
| *Observational analysis ($p < 0.05$ Spearman)* | | | |
| Significant features ($/20$) | 19 | 9 | 1 |
| Top Q5/Q1 OR (feature) | **2.55** (`min_radius`) | 0.62 (`n_segments`) | **3.84** (`std_curvature`) |
| *Discriminative test AUC* | | | |
| Bézier (20-d) | 0.706 | 0.536 | **0.646** |
| BiomedCLIP (512-d) | **0.836** | 0.528 | 0.454 |
| Image-level (EffB2) | **0.961** | 0.534 | 0.624 |
| *Counterfactual $\Delta$ on `tortuosity_4x`* | | | |
| $\Delta$ (Uncond) | $+$**0.625** | — | — |
| $\Delta$ (DR-/disease-prompt) | $+0.311$ | $+0.337$ | $+0.191$ |
| Strict-subset $n$ | 97 / 190 | 17 | 30 |
| Pixdrop control / Tort. ratio | $\sim 1/18$–$1/25$ | $\sim 1/200$ | $\sim 1/1$ |

## E.1 Stroke Full Tables

**Cohort design and discriminative context** The stroke cohort combines past, concurrent, and post-imaging events under a single ICD-coded "stroke history" label, making the task broader than incident-only prediction and potentially reducing single-image discriminative signal. XGBoost yields Bézier (20-d) test AUC $0.536$ and BiomedCLIP frozen (512-d) $0.528$; a single-view stroke CNN reaches AUC $0.534$, sensitivity $0.205$, and specificity $0.845$. The intervention response therefore demonstrates pipeline behavior under a difficult scoring regime; stronger discrimination is needed for a stroke-mechanism interpretation.

**Comparison with the most relevant UK Biobank stroke benchmark** Yusufu et al. (2025) report an adjusted AUC of $0.752$ on UK Biobank ($n = 45{,}161$, 749 incident strokes) using 118 retinal vascular parameters (calibre, density, tortuosity, branching angle, complexity) extracted by the Retina-based Microvascular Health Assessment System (RMHAS) together with traditional risk factors, vs. $0.739$ from risk factors alone. Their incident-only observational design (Cox regression

with restricted cubic splines) provides a strong population-risk benchmark, while BTECF addresses the complementary question of how a trained classifier responds to explicit geometry edits. The incident-only vs. mixed-temporal design, along with cohort composition, inputs, and representation quality, prevents a direct AUC comparison.

**Counterfactual responses and observational context** Despite weak discrimination, parameter-level intervention on the strongest tortuosity dose (`tortuosity_4x`) shows a monotonic classifier response (Table 11); Spearman regression on the 20-dimensional Bézier feature set yields nine features at $p < 0.05$ (Table 10), all in the down-stroke direction, six of which align directionally with caliber-narrowing (Wong et al., 2001) and fractal-dimension reduction (Doubal et al., 2010).

Table 10: Stroke observational analysis: 20-dimensional Bézier feature association on the $n = 3{,}409$ evaluation cohort. Significance flags use the convention $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. The nine features in the upper block are significant at $p < 0.05$; the lower block is non-significant.

| Feature | Spearman $\rho$ | OR/SD (95% CI) | Q5/Q1 OR |
|---|---|---|---|
| `n_segments` | −0.065 *** | 0.819 (0.74, 0.91) | 0.62 |
| `total_arc_length` | −0.062 *** | 0.831 (0.75, 0.92) | 0.66 |
| `coverage_ratio` | −0.058 *** | 0.839 (0.76, 0.93) | 0.68 |
| `max_radius` | −0.050 ** | 0.859 (0.78, 0.95) | 0.62 |
| `std_radius` | −0.040 * | 0.885 (0.80, 0.98) | 0.79 |
| `radius_cv` | −0.039 * | 0.892 (0.81, 0.98) | 0.71 |
| `mean_radius` | −0.038 * | 0.885 (0.81, 0.97) | 0.84 |
| `max_curvature` | −0.038 * | 0.883 (0.78, 1.00) | 0.97 |
| `thick_vessel_ratio` | −0.035 * | 0.904 (0.81, 1.01) | 0.82 |
| `std_curvature` | −0.033 n.s. | 0.888 (0.80, 0.98) | 0.79 |
| `mean_curvature` | −0.033 n.s. | 0.895 (0.79, 1.02) | 0.72 |
| `mean_segment_length` | +0.032 n.s. | 0.942 (0.86, 1.04) | 1.28 |
| `mean_chord_length` | +0.031 n.s. | 0.942 (0.86, 1.04) | 1.23 |
| `min_radius` | −0.023 n.s. | 0.909 (0.83, 1.00) | 0.94 |
| `max_tortuosity` | −0.017 n.s. | 0.938 (0.84, 1.05) | 0.93 |
| `std_chord_length` | −0.009 n.s. | 0.918 (0.84, 1.01) | 0.90 |
| `std_segment_length` | −0.008 n.s. | 0.919 (0.84, 1.01) | 0.92 |
| `std_tortuosity` | −0.008 n.s. | 0.979 (0.87, 1.10) | 1.07 |
| `mean_tortuosity` | +0.002 n.s. | 0.917 (0.84, 1.00) | 0.92 |
| `branching_density` | −0.001 n.s. | 1.007 (0.91, 1.12) | 0.95 |

Table 11: Stroke counterfactual intervention: 13-configuration causal table on the $n = 17$ strict reverse-counterfactual subset (baseline stroke probability $< 0.3$). Mean ± SEM and 95% CI on stroke probability under the EfficientNet-B2 stroke head. Configurations sorted by mean post-intervention probability.

| Configuration | Mean | SEM | 95% CI lower | 95% CI upper |
|---|---|---|---|---|
| baseline | 0.0009 | 0.0003 | 0.0002 | 0.0016 |
| `pixdrop_30` | 0.0017 | 0.0014 | −0.0013 | 0.0047 |
| `tortuosity_1x` | 0.0019 | 0.0008 | 0.0002 | 0.0036 |
| `arc_drop_10` | 0.0027 | 0.0016 | −0.0007 | 0.0061 |
| `radius_x0.85` | 0.0028 | 0.0016 | −0.0006 | 0.0062 |
| `arc_drop_20` | 0.0030 | 0.0015 | −0.0002 | 0.0062 |
| `pixdrop_10` | 0.0053 | 0.0035 | −0.0021 | 0.0127 |
| `pixdrop_20` | 0.0065 | 0.0052 | −0.0045 | 0.0175 |
| `radius_x0.70` | 0.0115 | 0.0096 | −0.0089 | 0.0319 |
| `radius_x0.55` | 0.0163 | 0.0146 | −0.0146 | 0.0472 |
| `tortuosity_2x` | 0.0178 | 0.0087 | −0.0006 | 0.0362 |
| `arc_drop_30` | 0.0278 | 0.0173 | −0.0089 | 0.0645 |
| `tortuosity_4x` | **0.3383** | 0.0716 | 0.1865 | 0.4901 |

### E.2 ALZHEIMER'S DISEASE FULL TABLES

**Cohort, discriminative-ranking inversion, and AD classifier** The Alzheimer's disease (AD) cohort comprises 550 unique participants with patient-level deduplication (eval split), of whom 41 (7.5%) carry an AD diagnosis label; training and val/test splits follow the protocol used for DR and stroke. XGBoost on the evaluation cohort yields Bézier (20-d) test AUC 0.646, BiomedCLIP frozen (512-d) 0.454. The Bézier > CLIP ordering inverts that of DR, showing stronger discrimination by geometric features than by this image-feature baseline. The single-view AD CNN used as the counterfactual scorer (EfficientNet-B2, weighted cross-entropy, 30 epochs) achieves test AUC 0.624, sensitivity 0.50, specificity 0.677.

**Observational anchor** Spearman regression on the 20-dimensional Bézier feature set yields `std_curvature` significant at $p < 0.05$ ($\rho = +0.086$, OR/SD $= 1.41$ [1.04, 1.92]) with the largest Q5/Q1 odds ratio across the three diseases (3.84, 95% CI $[1.30, 11.37]$). This result complements prior evidence of retinal vascular differences in AD from Frost et al. (2013); their venular tortuosity measure and our pooled `std_curvature` capture different geometric quantities and need not share direction.

**Counterfactual intervention** On the $n = 30$ strict reverse-counterfactual subset (baseline AD probability $< 0.3$), parameter-level intervention on `tortuosity_4x` produces the highest mean response within the tortuosity family (0.299 vs. baseline 0.109, $\Delta = +0.19$), reproducing the positive tortuosity direction observed on DR and stroke. The 13-configuration response table is in Table 12; the specificity caveat tied to the AD classifier is discussed in §4.5.

Table 12: AD counterfactual intervention: 13-configuration causal table on the $n = 30$ strict reverse-counterfactual subset (baseline AD probability $< 0.3$). Mean and SEM on AD probability under the AD-retrained EfficientNet-B2 head; two-sided 95% CI computed using Student's $t$ distribution with $df = 29$ ($t_{0.025} = 2.045$). Configurations sorted by mean post-intervention probability.

| Configuration | Mean | SEM | 95% CI lower | 95% CI upper |
|---|---|---|---|---|
| baseline | 0.109 | 0.021 | 0.066 | 0.152 |
| `pixdrop_30` | 0.169 | 0.046 | 0.075 | 0.263 |
| `arc_drop_10` | 0.182 | 0.034 | 0.112 | 0.252 |
| `radius_x0.85` | 0.202 | 0.044 | 0.112 | 0.292 |
| `tortuosity_2x` | 0.202 | 0.040 | 0.120 | 0.284 |
| `pixdrop_20` | 0.205 | 0.050 | 0.103 | 0.307 |
| `tortuosity_1x` | 0.207 | 0.044 | 0.117 | 0.297 |
| `radius_x0.70` | 0.252 | 0.055 | 0.140 | 0.364 |
| `radius_x0.55` | 0.272 | 0.055 | 0.160 | 0.384 |
| `tortuosity_4x` | **0.299** | 0.065 | 0.166 | 0.432 |
| `arc_drop_30` | 0.318 | 0.059 | 0.197 | 0.439 |
| `pixdrop_10` | 0.323 | 0.061 | 0.198 | 0.448 |
| `arc_drop_20` | 0.333 | 0.054 | 0.223 | 0.443 |

## F COUNTERFACTUAL EXAMPLES AND GRAD-CAM SALIENCY

Figure 5 shows three examples selected for large EfficientNet-B2 responses to the strongest tortuosity edit: baseline DR probability $< 0.17$ increases to $> 0.93$. These cases visualize pronounced classifier sensitivity and the accompanying changes in color, texture, and vessel boundaries; they are selected response examples rather than a representative image-quality sample. Figure 6 shows the same images with Grad-CAM overlays under both scoring backbones.

The Grad-CAM comparison (Figure 6) reveals substantial backbone dependence. EfficientNet-B2 post-intervention probabilities are 0.973, 0.932, and 1.000; the corresponding ResNet-50 values are 0.013, 0.773, and 0.004. The overlays localize broad vessel-adjacent regions while showing that high-confidence responses are not consistent across the two backbones for these examples. Together with the pixel-drop contrasts in §C.1, they characterize model sensitivity but do not exclusively attribute it to target geometry rather than coincident texture or boundary changes.

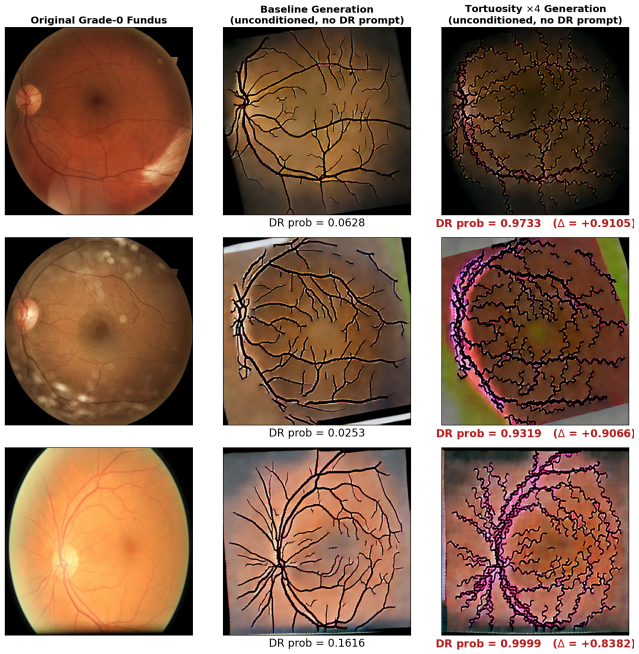


Figure 5: Three examples selected for large EfficientNet-B2 responses. Under this scorer, each example shifts from baseline DR probability $< 0.17$ to $> 0.93$ under `tortuosity_4x` intervention. Each row shows (i) the original grade-0 fundus, (ii) the baseline ControlNet generation, (iii) the `tortuosity_4x` generation.

## G Conditioning Taxonomy (L0–L3)

Table 13 summarizes the four levels and their representative work. L0 has parametric control but no photorealistic texture. L1 has photorealistic texture but no clean parametric decoupling. L2 bridges these representations: control is parametric in the upstream representation while the generator side is unchanged: ControlNet receives a $(512, 512, 3)$ pixel tensor identical in format to a binary mask. L3 is left to future work: a parameter-native ControlNet that takes the Bézier control-point sequence directly (e.g. as a token sequence attended by the diffusion UNet) would remove the rasterization step. Whether this improves control or fidelity, and whether any benefit transfers across diseases, remains to be tested.

Table 13: Four-level taxonomy of conditioning for fundus vessel synthesis. This work occupies the L2 bridge.

| Level | Upstream control | Downstream generator | Representative work |
|---|---|---|---|
| L0 | Classical renderers (ASM, Catmull-Rom) | Non-generative / interactive GUI | Bonaldi 2016 (Bonaldi et al., 2016); Castro 2020 (Castro et al., 2020) |
| L1 | Binary vessel mask (pixel) | Deep generative (GAN / Diffusion) | Guo2024 (Guo et al., 2024); Feng 2024 (Feng et al., 2024); Go 2024 (Go et al., 2024) |
| L2 ⋆ | Cubic Bézier control points to 3-channel hint | SD 2.1 + LoRA + ControlNet | This work (BTECF) |
| L3 | Bézier parameters as condition embeddings | Re-trained ControlNet / diffusion | Future work |

## H DR Pathology Context: Three Mechanistic Hypotheses

This appendix uses three literature-based histopathological chains to interpret the SHAP top-5 directions of Table 1 (§4.2): `total_arc_length` (↓ DR), `branching_density` (*SHAP non-linear* with weak positive Spearman), `mean_radius` (↓ DR), `mean_chord_length` (↓ DR), and `mean_tortuosity` (↑ DR). The chains share intermediate nodes (pericyte loss, capillary closure, local retinal hypoxia) and provide mechanistic hypotheses for the negative associations in `total_arc_length`, `mean_radius`, and `mean_chord_length`, the positive association in `mean_tortuosity`, and the more complex `branching_density` pattern. Because the

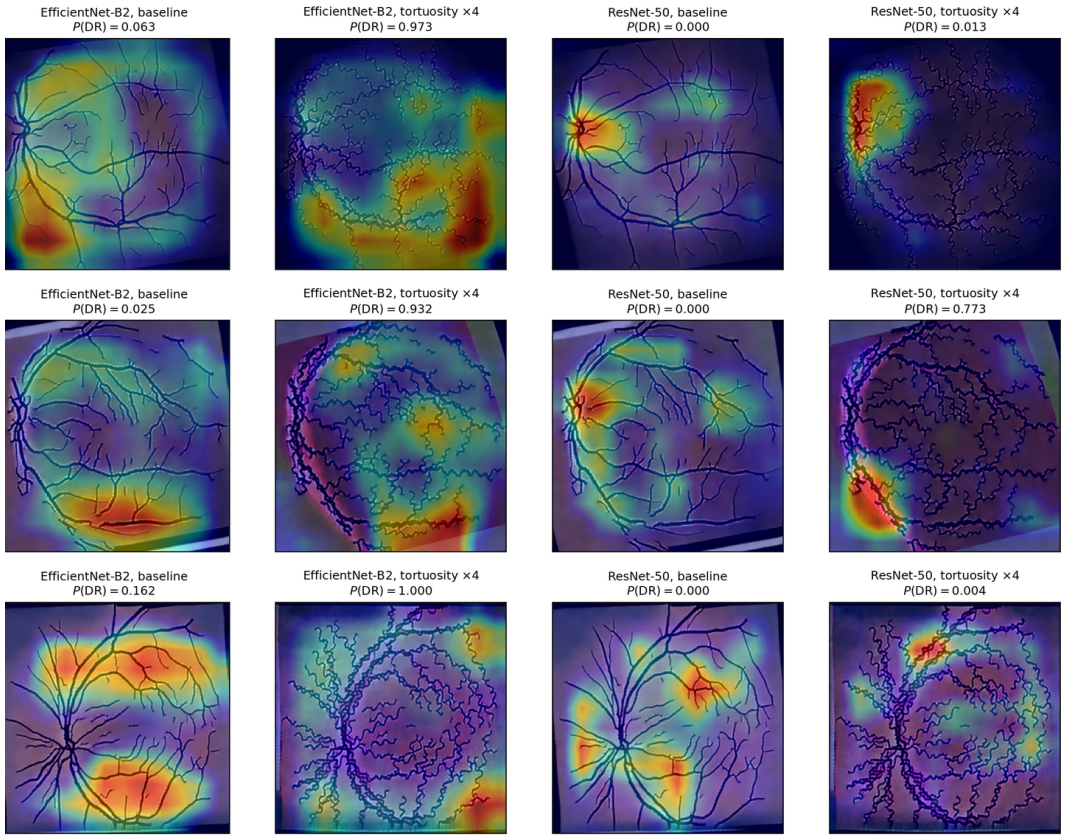


Figure 6: Grad-CAM saliency overlays on the three selected counterfactual examples (rows) under EfficientNet-B2 and ResNet-50 (columns), comparing baseline against `tortuosity_4x`. Heatmaps use the standard Grad-CAM `jet` colormap with $\alpha = 0.40$ overlay so that vessel topology stays legible underneath. Predicted DR probability $P(\mathrm{DR})$ is annotated per panel. Responses and saliency patterns differ across backbones; these qualitative overlays do not isolate vessel geometry from accompanying image changes.

analysis is binary, these hypotheses organize the observed directions rather than identify stagewise trajectories.

### H.1 Chain 1: Vasoregression and Branch Loss

The first chain begins at the inner blood–retinal barrier. Hyperglycemia drives selective pericyte loss, which removes mechanical and signaling support from the capillary endothelium and produces endothelial-cell dysfunction. The dysfunctional endothelium is the substrate of two co-occurring lesions: focal microaneurysm formation and segmental capillary closure. Capillary closure produces local retinal ischemia, which drives *vasoregression*, the active deletion of the mature capillary network together with atrophy of the immediately upstream arterioles. The vessels deleted by vasoregression are predominantly the small distal branches that once filled the periphery of the AutoMorph segmentation skeleton; once deleted, they are no longer detected as fitted Bézier segments (Cogan et al., 1961; Hammes et al., 2002; Curtis et al., 2009; Stitt et al., 2016; Forster et al., 2021; Pramil et al., 2021; Antonetti et al., 2012; Broe et al., 2014; Popovic et al., 2019). `total_arc_length`, `branching_density`, and `mean_chord_length` could decrease when regression dominates detectable vascular growth, offering a possible explanation for the negative associations at ranks 1 and 4 and a subtractive contribution to rank 2.

### H.2 Chain 2: Caliber Biphasic Trajectory

The second chain offers an interpretation of `mean_radius` (rank 3) through a biphasic time-course in which an early arteriolar dilation is overtaken by a late arteriolar narrowing. In early DR, two parallel processes widen the arteriolar lumen. First, impaired retinal autoregulation under sustained hyperglycemia produces tonic hyperperfusion. Second, mild local hypoxia and high glucose drive endothelial release of nitric oxide (NO); NO relaxes the arteriolar smooth muscle and produces functional vasodilation. Both arms push `mean_radius` upward in the early phase. In late DR, two further independent paths narrow the same vessels. *Path A*: chronic oxidative stress thickens the vascular basement membrane (BM); BM thickening directly thickens the wall, reduces the effective diameter of the lumen, and stiffens the vessel. *Path B*: cumulative downstream capillary closure (the late-stage end of Chain 1) removes the distal perfusion demand and the upstream arterioles atrophy in response. The two late-phase paths jointly produce a net reduction of `mean_radius` (Stitt et al., 2016; Roy & Kim, 2021; Antonetti et al., 2012; Curtis et al., 2009; Ashraf et al., 2021; Cheung et al., 2008; Klein et al., 2004; Velayutham et al., 2020; Xu et al., 2026). The observed negative association is compatible with the late-phase arm of this trajectory. Resolving the full stagewise course requires stage-stratified data beyond the present binary analysis (Appendix C.2).

### H.3 Chain 3: Tortuosity Biphasic Trajectory and Late-Stage Neovessel Coupling

The third chain offers a hypothesis for `mean_tortuosity` (rank 5) and the differing association and attribution signs for `branching_density` (rank 2) carrying a weak positive Spearman ($\rho = +0.014$, $p < 0.05$) together with a strongly negative tree-conditional `feat-SHAP corr` ($-0.55$). Tortuosity rises in both the early and the late phase of DR, but through different mechanisms.

**Early-mid phase** Four contributors push `mean_tortuosity` upward. (i) Pericyte loss, the same starting node as Chain 1, removes mural support from the endothelium, so the vessel wall is more compliant under pulsatile load. (ii) High axial shear stress from the early hyperperfusion arm of Chain 2 drives elongation and buckling of compliant segments. (iii) Impaired retinal autoregulation amplifies the pressure pulse delivered to the affected segments. (iv) Early basement-membrane remodeling alters the passive elastic properties (Roy & Kim, 2021). The four contributors all act in the early-mid disease window and form the early upward arm of the tortuosity time-course.

**Late phase** The end of Chain 1, that is, widespread vasoregression and cumulative capillary closure, produces local retinal hypoxia. Hypoxia sharply elevates vascular endothelial growth factor (VEGF), and VEGF drives neovessel proliferation. The new vessels are typically thin and short, and may be missed by segmentation or not resolved as distinct fitted Bézier segments (Aiello et al., 1994; Stitt et al., 2016; Sasongko et al., 2011; 2016; Forster et al., 2021; Xu et al., 2026; Fathimah et al.,

2025). This suggests one possible source of differing associations across the SHAP top-5 features. `total_arc_length` and `mean_chord_length`, which sum or average over segments that are actually fitted, could decrease if regression of the captured mature network outweighs detectable neovessel growth. However, `branching_density`, being more sensitive to the junction-level cues produced by neovessel sprouts at their basal anastomoses, could retain an additive signal. Furthermore, the characteristic geometry of neovascularization, with its tight loops and abrupt directional changes, reinforces `mean_tortuosity`. Together, these mechanisms provide a coherent hypothesis for the positive tortuosity association; separating their relative contributions requires stage-resolved analysis.

**Summary mapping to SHAP top-5** Chain 1 links vasoregression to the negative associations at ranks 1 and 4 (`total_arc_length`, `mean_chord_length`) and to a subtractive influence on rank 2 (`branching_density`). Chain 2 provides biological context for rank 3 (`mean_radius`), while Chain 3 connects neovascular remodeling to rank 5 (`mean_tortuosity`) and branching density. These literature-based hypotheses map the observed feature directions onto established DR processes; correlated features and model interactions remain alternative explanations for differences between marginal associations and conditional SHAP attributions.

# I Datasets, Code, and Licenses

## I.1 Datasets

**Eyepacs, Aptos, Messidor Diabetic Retinopathy** `https://www.kaggle.com/datasets/ascanipek/eyepacs-aptos-messidor-diabetic-retinopathy`. This dataset does not claim ownership of the original images. It is provided as a unified, resized, augmented, and preprocessed version for research and educational purposes.

**EyePACS** `https://www.kaggle.com/competitions/diabetic-retinopathy-detection`. License/terms: Custom Kaggle Competition Rules for the Diabetic Retinopathy Detection competition; dataset redistribution is not permitted.

**APTOS-2019** `https://www.kaggle.com/competitions/aptos2019-blindness-detection/rules#7-competition-data`. License/terms: Custom Kaggle Competition Rules; non-commercial academic/research and educational use only; redistribution is prohibited.

**APTOS Gaussian Filtered** `https://www.kaggle.com/datasets/sovitrath/diabetic-retinopathy-224x224-gaussian-filtered`. CC0: Public Domain.

**Messidor-2** `https://www.adcis.net/en/third-party/messidor2/`. License/terms: Custom research and educational use license; redistribution and unauthorized commercial use are prohibited. Acknowledgment: Kindly provided by the Messidor program partners (`https://www.adcis.net/en/third-party/messidor/`).

**UK Biobank** (`https://www.ukbiobank.ac.uk/`). Managed-access biomedical resource; usage is governed by the UK Biobank Access Procedures.

## I.2 Code and Models

**AutoMorph** Source code for this pipeline is available from (`https://github.com/rmaphoh/AutoMorph`). Apache License Version 2.0, January 2004

**Stable Diffusion 2.1** (`https://huggingface.co/sd2-community/stable-diffusion-2-1`). Openrail++ license.

**LoRA adapter** (`https://github.com/microsoft/LoRA`). MIT license.

**ControlNet** (`https://github.com/lllyasviel/ControlNet?tab=readme-ov-file`). Apache License Version 2.0, January 2004.

**BiomedCLIP** (https://huggingface.co/microsoft/BiomedCLIP-PubMedBERT_256-vit_base_patch16_224) MIT license.