SEGMENTATION AS EDITING: UNIFIED MEDICAL OPTICAL IMAGE SEGMENTATION WITH DIFFERENTIABLE EDIT-SPACE GEOMETRY ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical optical imaging is widely used in clinical diagnosis and pathological assessment. Precise segmentation of clinically meaningful anatomical structures and pathological regions enables quantitative assessment for clinical decision-making and research. However, this field still faces at least two challenges: the lack of a unified segmentation dataset dedicated to diverse optical modalities and the need for a unified segmentation model that achieves competitive performance across multiple optical imaging modalities. To this end, we construct OptiMedEdit-37K, a segmentation-centered medical optical image editing dataset covering 7 medical optical imaging modalities and 18 anatomical and pathological targets. Meanwhile, we develop ***Segmentation as Editing*** as a unified paradigm for medical optical image segmentation. Through staged fine-tuning on OptiMedEdit-37K under this paradigm, we first establish a **Uni**fied **Med**ical **Opti**cal image segmentation model (UniMedOpti). To retain its shared capabilities while better modeling target-specific geometric structures, we then augment UniMedOpti with ***G**eometry-aware **R**esidual **A**daptation through **D**ifferentiable **E**dit-space Alignment* (GRADE). It freezes the shared backbone and trains low-rank residuals using geometric losses on soft masks differentiably recovered from generated edits. UniMedOpti achieves competitive segmentation performance across multiple medical optical imaging modalities and datasets. The experimental results demonstrate that *Segmentation as Editing* combines cross-modal generality with competitive accuracy.

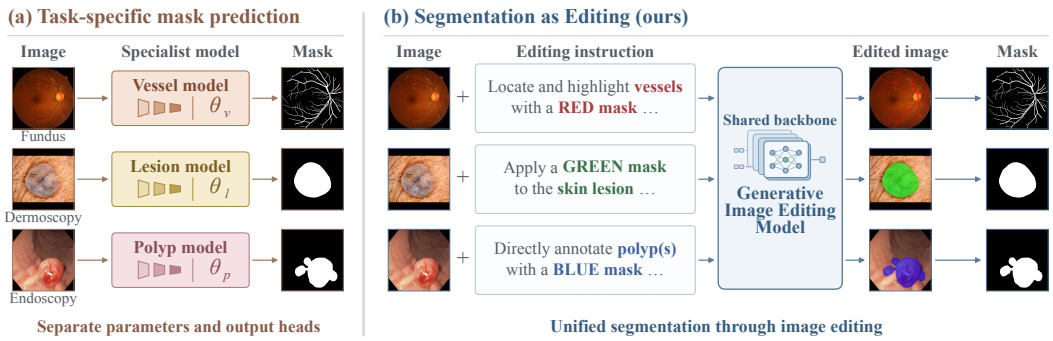

Figure 1: **From task-specific model to Segmentation as Editing. (a)** Traditional methods struggle to unify medical image segmentation across modalities. **(b)** Segmentation as Editing uses a generative image editing model to generate instruction-guided edits, from which masks are recovered.

1 INTRODUCTION

Medical optical imaging comprises modalities that acquire tissue morphology and pathology through light–tissue interactions (DICOM Standards Committee, 2026; National Institute of Biomedical Imaging and Bioengineering, 2025). Accurate segmentation converts the location, extent, and organization of lesions and anatomical structures into reproducible measurements for diagnosis, treatment planning, and biomedical research (Ma et al., 2024; Caicedo et al., 2019; Cheung

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.

et al., 2021; Go et al., 2025; Hardo et al., 2024). Despite their shared optical basis, these modalities differ substantially in visual appearance, spatial scale, and target geometry, spanning cellular structures, lesions, and branching vascular networks (Zhao et al., 2025; Kuş & Aydin, 2024). We focus on this imaging family to investigate a central challenge in unified segmentation: accommodating such heterogeneity within a single generative image editing model while retaining the geometric precision required by each target.

Task-specific segmentation models achieve strong performance through dedicated training and configuration (Ronneberger et al., 2015; Isensee et al., 2021), but separately trained specialists require distinct parameters and task interfaces across modalities (Figure 1(a)). Promptable frameworks broaden this coverage (Ma et al., 2024; Zhao et al., 2025).

Multimodal generative models offer a potential common interface. In natural images, shared architectures support text-guided generation, localized editing, and predictions expressed as images (Wu et al., 2025a; Geng et al., 2024). Yet public medical optical segmentation datasets usually provide source images and masks (Jin et al., 2022; Codella et al., 2019; Ali et al., 2023; Gamper et al., 2019), not the target instructions and source-preserving edits needed for native editing supervision (Lai et al., 2026). We therefore construct OptiMedEdit-37K with source–instruction–target edit triplets for staged adaptation. Given a source image and target instruction, the model produces a source-preserving edit that marks the requested structure, from which the evaluation mask is recovered deterministically (Figure 1(b)).

Segmenting topologically complex structures such as vascular networks poses a particular challenge: thin, branching structures demand precise localization and connectivity preservation (Jin et al., 2022; Shit et al., 2021; Bai et al., 2026; Johny et al., 2026). We introduce GRADE to specialize the shared model through additive low-rank residuals (Hu et al., 2022), activated only by instructions targeting vascular structures. Differentiable edit-space alignment backpropagates structural errors from the generated edit to these residuals, adding vascular modeling capacity while preserving the shared non-vessel path. Together, these components realize *Segmentation as Editing*. Our contributions are threefold:

(1) We construct OptiMedEdit-37K as a reusable data foundation for training and evaluating generative segmentation models. It unifies segmentation supervision across optical modalities and targets within a common image editing format.

(2) We develop UniMedOpti, a unified medical optical segmentation model under the *Segmentation as Editing* paradigm, and introduce GRADE to enhance geometric and topological modeling of tubular structures. We design a three-stage training strategy comprising instruction alignment, cross-modal adaptation, and geometry-aware refinement.

(3) UniMedOpti achieves competitive segmentation performance across 7 medical optical imaging modalities and diverse targets, with 86.17% Dice and 80.54% IoU on the full OptiMedEdit-37K test set. These results support the feasibility of unifying diverse optical segmentation tasks within a single generative model.

2 Related Work

2.1 Specialist and Unified Medical Segmentation

U-Net and nnU-Net provide strong task-specific baselines (Ronneberger et al., 2015; Isensee et al., 2021). Optical specialists include PraNet for endoscopic polyps and HoVer-Net for histological nuclei (Fan et al., 2020; Graham et al., 2019), while vessel models exploit iterative refinement and geometry-adaptive convolutions (Li et al., 2020; Qi et al., 2023). Unifying these tasks requires accommodating variation in appearance and geometry while retaining segmentation accuracy. Cellpose and CellSAM generalize across cellular imaging domains (Stringer et al., 2021; Marks et al., 2025). MedSAM and BiomedParse broaden coverage through spatial and textual prompts, respectively (Ma et al., 2024; Zhao et al., 2025), while UniverSeg uses support image–mask pairs without task-specific fine-tuning (Butoi et al., 2023). We study a shared instruction-guided image editing interface across optical modalities.

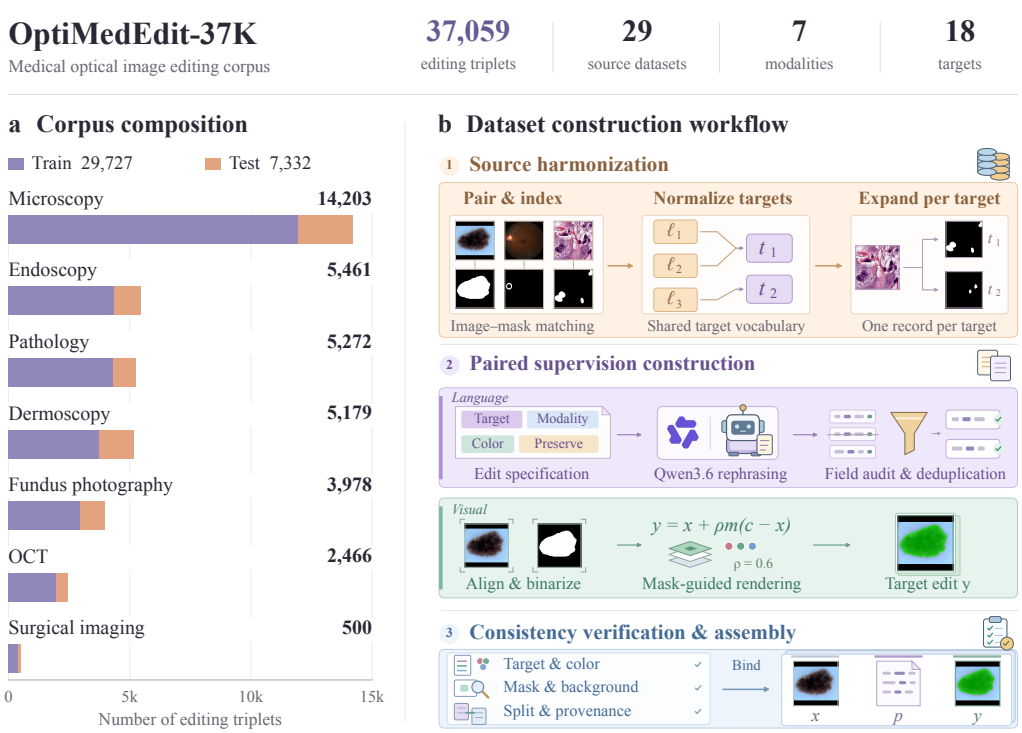

Figure 2: **OptiMedEdit-37K: coverage and editing-triplet construction.** (a) Training and test triplets across seven modalities, with totals indicated by the labels. (b) Source harmonization pairs images with annotations, normalizes targets, and forms target-specific records. Language and visual branches construct instructions and deterministic target edits. Consistency checks bind $(\mathbf{x}, p, \mathbf{y})$ with provenance. Image examples are from ISIC 2018, REFUGE, and PanNuke.

2.2 Generative Models and Medical Image Editing

Painter and SegGPT express visual tasks in image space, while InstructDiffusion unifies segmentation and generation as instruction-conditioned pixel prediction (Wang et al., 2023a;b; Geng et al., 2024). HealthGPT combines medical comprehension and generation through heterogeneous low-rank adaptation (Lin et al., 2025), and UniMedVL includes promptable mask generation (Ning et al., 2026). MieDB-100k includes source-preserving mask overlays and analytic recovery as Perception tasks within a broader medical editing benchmark (Lai et al., 2026). We build on this formulation to unify optical segmentation and add geometric supervision.

clDice and cbDice supervise vascular connectivity and geometry in mask prediction (Shit et al., 2021; Shi et al., 2024). GRADE routes this supervision through differentiable mask recovery and the frozen decoder to target-activated residuals, retaining the native editing interface at inference.

3 OptiMedEdit-37K

To train a shared segmentation model through image editing, we construct OptiMedEdit-37K, which links target-specific instructions to edits grounded in existing segmentation annotations.

3.1 Corpus Organization

Each triplet $(\mathbf{x}, p, \mathbf{y})$ contains a source image and target edit $\mathbf{x}, \mathbf{y} \in [0,1]^{H\times W\times 3}$, and an editing instruction p. We map source labels to shared targets and form one record per annotated target. The binary mask $\mathbf{m} \in \{0,1\}^{H\times W}$, original source label, and sample identity are retained for supervision and evaluation. The corpus contains 29,727 training and 7,332 test triplets from 29 datasets, detailed in Appendix A.

3.2 Paired Supervision Construction

Figure 2(b) summarizes the language and visual branches. Canonical instructions specify the target, modality, rendering color, and requirement to preserve the background.

For newly constructed triplets, Qwen3.6-35B-A3B (Qwen Team, 2026) varies instruction wording and syntax while preserving these fields. Field checks retain variants with consistent editing semantics, and duplicate strings are removed.

The visual branch renders source-preserving target edits from aligned binary annotations through alpha compositing (Porter & Duff, 1984; Lai et al., 2026):

$$\mathbf{y}_i = \mathbf{x}_i + \rho m_i \mathbf{d}_i, \qquad \mathbf{d}_i = \mathbf{c} - \mathbf{x}_i, \tag{1}$$

where i indexes spatial pixels, $\mathbf{x}_i, \mathbf{y}_i, \mathbf{c} \in [0,1]^3$ are RGB vectors, $m_i \in \{0,1\}$ is a scalar mask value, and $\mathbf{d}_i \in \mathbb{R}^3$ is the prescribed edit direction. The instruction p specifies a red, green, or blue $\mathbf{c}$, with overlay opacity $\rho = 0.6$. This operation preserves the source appearance beneath the overlay and leaves all pixels outside the annotated target unchanged. Before assembly, instruction–edit pairs are checked for target/color agreement, mask alignment, background preservation, and consistency of split and provenance metadata.

For every nonzero edit direction $\mathbf{d}_i$, this relation admits an analytic scalar inverse. Section 3.3 uses its ridge-stabilized form, which remains defined for all pixels. This deterministic relation yields source-preserving, directly auditable editing triplets at scale. Further details of the dataset construction process are provided in Appendix A.

3.3 Deterministic Mask Recovery

At inference, the model produces a predicted target edit $\widehat{\mathbf{y}}$ rather than a standalone segmentation mask. Equation (1) expresses the RGB edit as a scalar displacement along $\mathbf{d}_i$. We compute the unconstrained edit strength estimate $\widetilde{a}_i$ using a ridge-stabilized form of alpha-deblending (Lai et al., 2026), with $\varepsilon = 10^{-6}$ for numerical stability:

$$\widetilde{a}_i = \frac{(\widehat{\mathbf{y}}_i - \mathbf{x}_i)^{\mathsf{T}} \mathbf{d}_i}{\|\mathbf{d}_i\|_2^2 + \varepsilon}, \qquad \widehat{a}_i = \mathrm{Clip}_{[0,1]}(\widetilde{a}_i). \tag{2}$$

The clipped strength $\widehat{a}_i$ estimates ρm_i. Using the indicator function $\mathbf{1}[\cdot]$, we obtain the binary mask at a threshold τ fixed before test evaluation:

$$\widehat{m}_i = \mathbf{1}[\widehat{a}_i > \tau]\,. \tag{3}$$

The mask is recovered analytically from the generated edit, without additional trainable parameters. Derivations and evaluation details are in Appendices B.4 and D.1.

4 Method

UniMedOpti unifies medical optical segmentation through instruction-guided image editing and analytic mask recovery (Section 3.3). Figure 3 summarizes training: Stages I and II align instruction conditioning and adapt the complete MMDiT across modalities (Section 4.1, left panel). Stage III freezes this backbone and refines vascular geometry through target-activated residuals (Section 4.2) supervised by differentiable edit-space alignment (Section 4.3, right panel).

4.1 Learning a Shared Generative Segmentation Backbone

We instantiate the framework with Qwen-Image-Edit (Wu et al., 2025a). The frozen multimodal encoder H (the VLM in Figure 3) converts the source image $\mathbf{x}$ and instruction p into semantic features $\mathbf{h} = H(\mathbf{x}, p)$. The MLP connector maps these features into the MMDiT conditioning stream. The frozen VAE encoder E produces source image latents $\mathbf{z}_x = E(\mathbf{x})$ for appearance conditioning and target edit latents $\mathbf{z}_y = E(\mathbf{y})$ for training. The frozen decoder D maps predicted latents back to RGB images. We denote the MMDiT, including its connector, by $f_{\boldsymbol{\theta}}$.

We retain the backbone's Flow Matching formulation (Lipman et al., 2023; Wu et al., 2025a). Let L be the dimension of the vectorized target latents, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_L)$ Gaussian noise, and $\sigma \in [0,1]$

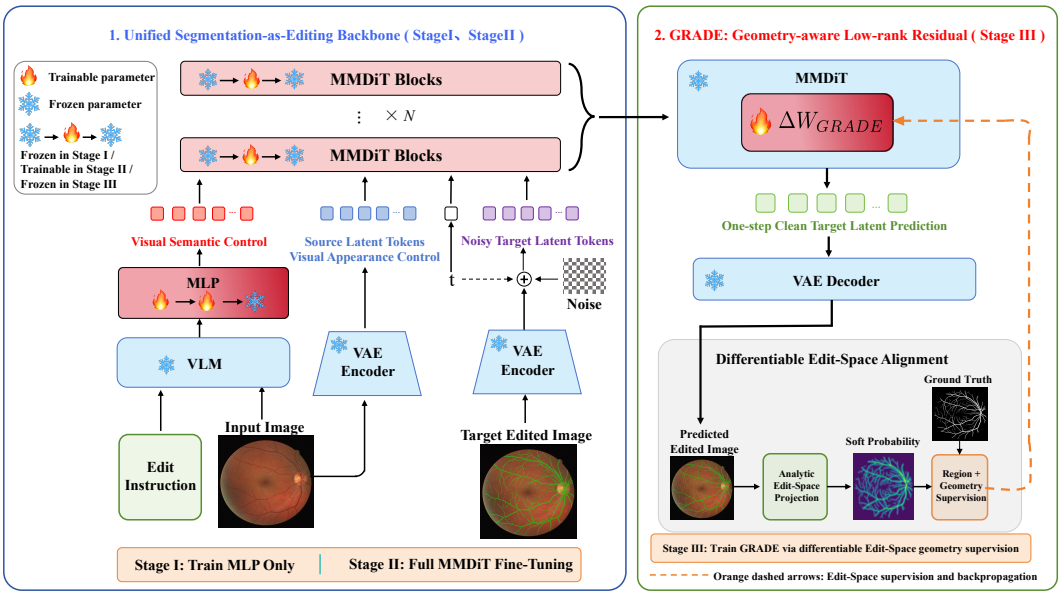

Figure 3: **Overview of the proposed training framework.** Left: semantic features and source image latents condition target edit generation. Stages I and II train the MLP connector and then the complete MMDiT (Section 4.1). Right: Stage III freezes the shared backbone and trains target-activated low-rank residuals (Section 4.2). Predicted target latents are decoded into edits, which are projected to obtain region and geometry supervision (Section 4.3). Orange dashed arrows show the auxiliary supervision and gradient path.

the sampled noise level. For the noised target latent $\mathbf{z}_\sigma$, the target velocity is $\mathbf{v}^\star = \boldsymbol{\epsilon} - \mathbf{z}_y$, and the predicted velocity is $\widehat{\mathbf{v}} = f_{\boldsymbol{\theta}}(\mathbf{z}_\sigma, \mathbf{z}_x, \mathbf{h}, \sigma)$:

$$\mathbf{z}_\sigma = (1-\sigma)\mathbf{z}_y + \sigma\boldsymbol{\epsilon}, \qquad \mathcal{L}_{\text{flow}} = \mathbb{E}\big[w(\sigma)\|\widehat{\mathbf{v}} - \mathbf{v}^\star\|_2^2\big]\,. \tag{4}$$

Here, $w(\sigma)$ is the backbone's timestep weight. The expectation covers training triplets, noise levels, and noise realizations.

Stage I: connector alignment. We train only the MLP connector to adapt the medical instruction interface, keeping the remaining MMDiT parameters fixed. **Stage II: shared backbone adaptation.** Starting from Stage I, we fine-tune the complete MMDiT, including the connector, on all training triplets. Both stages use Eq. (4) with H, E, and D frozen. Optimization details are given in Appendix C.

4.2 Target-Activated Geometry Residuals

The shared backbone provides cross-modal coverage, but velocity regression alone does not explicitly prioritize the connectivity of thin, branching structures. **Stage III: geometry-aware residual adaptation.** We freeze the Stage II parameters $\boldsymbol{\theta}$ and add trainable residuals to selected attention, feed-forward, and modulation projections. Following the low-rank parameterization of LoRA (Hu et al., 2022), each adapted projection ℓ, with input and output widths d_{in} and d_{out}, uses

$$\mathbf{W}_\ell^{\text{GRADE}}(p) = \mathbf{W}_\ell + g_{\text{ves}}(p)\Delta\mathbf{W}_\ell, \qquad \Delta\mathbf{W}_\ell = \frac{\alpha_\ell}{r}\mathbf{B}_\ell\mathbf{A}_\ell. \tag{5}$$

Here, $\mathbf{W}_\ell \in \mathbb{R}^{d_{\text{out}}\times d_{\text{in}}}$ is frozen, $\mathbf{A}_\ell \in \mathbb{R}^{r\times d_{\text{in}}}$ and $\mathbf{B}_\ell \in \mathbb{R}^{d_{\text{out}}\times r}$ are trainable, r is the rank bound, and α_ℓ is a fixed scale. The scalar gate $g_{\text{ves}}(p) \in \{0,1\}$ is one when the normalized target in p is a vascular structure and zero otherwise. We write $\boldsymbol{\phi}$ for the vector containing all entries of the residual matrices $\mathbf{A}_\ell$ and $\mathbf{B}_\ell$, and train only $\boldsymbol{\phi}$ on the vessel subset.

For a non-vessel instruction, the gate disables every residual, so the model evaluates exactly the Stage II velocity function for the same inputs. This preserves a shared parameter path for the remaining targets. Zero-initializing $\mathbf{B}_\ell$ also preserves the Stage II function at the start of adaptation,

as in standard LoRA. The explicit functional identity and parameterization details are provided in Appendix B.3.

4.3 Differentiable Edit-Space Alignment

The right panel of Figure 3 traces geometric supervision from the generated edit to the residuals defined in Section 4.2.

Recovering the current edit. During Stage III, $\widehat{\mathbf{v}}$ is predicted by the residual-augmented model $f_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{z}_\sigma, \mathbf{z}_x, \mathbf{h}, \sigma; p)$, where p also specifies the gate in Eq. (5). The linear path in Eq. (4) yields the estimated clean target latents $\widehat{\mathbf{z}}_y$ and decoded edit $\widehat{\mathbf{y}}$:

$$\widehat{\mathbf{z}}_y = \mathbf{z}_\sigma - \sigma\widehat{\mathbf{v}}, \qquad \widehat{\mathbf{y}} = D(\widehat{\mathbf{z}}_y). \tag{6}$$

This algebraic reconstruction reuses a single MMDiT evaluation and exposes geometry in the native RGB output through the frozen, differentiable decoder. Appendix B.2 gives the endpoint error identity.

A differentiable link to target geometry. During training, we reuse the unclipped strength $\widetilde{a}_i$ from Eq. (2) to provide differentiable geometric supervision. After normalization by the rendering opacity ρ, centering at b, and division by the temperature $T > 0$, the strength becomes a logit ℓ_i. A sigmoid maps this logit to a foreground probability $\widehat{q}_i \in (0, 1)$, preserving a differentiable path through mask recovery. The mapping and its derivatives are detailed in Appendix B.4.

Joint generative and geometric supervision. Let $\boldsymbol{\ell} = (\ell_i)$ and $\widehat{\mathbf{q}} = (\widehat{q}_i)$ denote the logit and probability maps, and let $\mathbf{m}$ be the reference mask. We apply binary cross-entropy (BCE), soft Dice, and centerline boundary Dice (cbDice) (Shi et al., 2024) to these recovered predictions:

$$\begin{aligned} \ell_{\text{aux}} &= \omega_{\text{B}}\ell_{\text{BCE}}(\boldsymbol{\ell}, \mathbf{m}) + \omega_{\text{D}}\ell_{\text{Dice}}(\widehat{\mathbf{q}}, \mathbf{m}) + \omega_{\text{C}}\ell_{\text{cbDice}}(\widehat{\mathbf{q}}, \mathbf{m}), \\ \mathcal{L}_{\text{GRADE}} &= \mathbb{E}\big[w(\sigma)\left(\|\widehat{\mathbf{v}} - \mathbf{v}^\star\|_2^2 + \lambda_{\text{GRADE}}\ell_{\text{aux}}\right)\big]. \end{aligned} \tag{7}$$

The nonnegative weights $\omega_{\text{B}}, \omega_{\text{D}}, \omega_{\text{C}}$ sum to one, and λ_{GRADE} controls geometric supervision relative to velocity regression. BCE uses logits for numerical stability, soft Dice measures region overlap, and cbDice adds centerline and boundary sensitivity. Both objectives use the same sampled state, velocity prediction, and timestep weight. The loss definitions, piecewise-differentiable cbDice implementation, and coefficient values are in Appendices B.5 and C.

The orange dashed arrows in Figure 3 represent the resulting gradient route. For $N = HW$ pixels, stack RGB vectors into $\mathbb{R}^{3N}$ and scalar maps into $\mathbb{R}^N$. The projection matrix $\mathbf{P}_{\mathbf{x},\mathbf{c}} \in \mathbb{R}^{N\times 3N}$ has diagonal row blocks $\mathbf{d}_i^\top/(\|\mathbf{d}_i\|_2^2 + \varepsilon)$. With K residual parameters, the decoder Jacobian is $\mathbf{J}_D \in \mathbb{R}^{3N\times L}$ and the velocity Jacobian with respect to $\boldsymbol{\phi}$ is $\mathbf{J}_{f,\boldsymbol{\phi}} \in \mathbb{R}^{L\times K}$. On a differentiable branch of the implemented losses, the chain rule gives

$$\nabla_{\boldsymbol{\phi}}\ell_{\text{aux}} = -\frac{\sigma}{\rho T}\mathbf{J}_{f,\boldsymbol{\phi}}^\top\mathbf{J}_D^\top\mathbf{P}_{\mathbf{x},\mathbf{c}}^\top\nabla_{\boldsymbol{\ell}}\ell_{\text{aux}}. \tag{8}$$

The product maps pixelwise loss gradients in $\mathbb{R}^N$ to gradients with respect to residual parameters in $\mathbb{R}^K$. Because mask recovery has no trainable parameters, geometric supervision must act on the generated edit and its underlying velocity prediction, with parameter updates confined to the target-activated residuals. Derivations and sensitivity bounds are provided in Appendix B.

5 Experiments

We evaluate cross-modal segmentation, dataset utility, the Segmentation as Editing paradigm, staged training, and the geometric contribution of GRADE. Our UniMedOpti uses Qwen-Image-Edit as its primary backbone. We test cross-backbone transfer with DeepGen 1.0, a unified multimodal model for image generation and editing (Wang et al., 2026).

Evaluation. We evaluate on the held-out subsets summarized in Table A.1. Reference selection follows the adopted partition and target definition. For official source splits, we prioritize methods reporting results on the same split. For benchmark-derived partitions, we prioritize results reported under the corresponding benchmark protocol and retain its published baselines when comparable

Table 1: **Segmentation performance across 7 medical optical imaging modalities.** To assess cross-modal competitiveness, we compare UniMedOpti with published references selected according to the evaluation splits and segmentation targets used for each dataset. Benchmark baselines are used when no comparable follow-up results are available. **Bold** marks the higher tabulated value, and a dash denotes an unavailable value.

Modality	Dataset	Published reference			Ours	
		Method	Dice↑	IoU↑	**Dice↑**	**IoU↑**
Fundus photography	FIVES	SuperVessel (Qiu et al., 2024)	**92.10**	**85.36**	89.13	81.26
	REFUGE-Disc	GCN (Tian et al., 2020)	**97.76**	**95.64**	95.52	92.97
	REFUGE-Cup	GCN (Tian et al., 2020)	95.58	91.60	**96.17**	**93.26**
Dermoscopy	ISIC 2016	CFG-MambaNet (Ren et al., 2026)	94.46	89.83	**94.82**	**92.05**
	ISIC 2018	TESL-Net (Iqbal et al., 2024)	**94.22**	**90.56**	93.19	88.24
Microscopy	DeepBacs	MedSegBench (Kuş & Aydin, 2024)	93.10	87.00	**95.58**	**91.73**
	BriFiSeg	MedSegBench (Kuş & Aydin, 2024)	84.00	73.80	**89.72**	**81.46**
	NuSeT	MedSegBench (Kuş & Aydin, 2024)	95.10	91.00	**96.14**	**93.43**
	YeaZ	MedSegBench (Kuş & Aydin, 2024)	95.40	91.40	**96.07**	**93.00**
Pathology	GlaS	ODS-SAM (Huang et al., 2025)	92.74	**87.01**	**92.76**	86.63
Endoscopy	Kvasir-SEG	Citrus-V (Wang et al., 2025)	90.91	—	**93.10**	89.00
OCT	CystoidFluid	MedSegBench (Kuş & Aydin, 2024)	85.70	76.50	**88.97**	**82.16**
Surgical imaging	RoboTool	MedSegBench (Kuş & Aydin, 2024)	88.20	79.80	**93.79**	**89.30**

follow-up results cannot be established. Higher scores obtained under different custom splits are not treated as directly comparable. Published references retain study-specific preprocessing, training, and metric aggregation, so Table 1 provides context for cross-modal competitiveness without establishing a controlled ranking. For space, it presents a subset spanning all 7 modalities, while aggregate results on the full test set are reported in Figure 4. For our model, Dice and IoU are computed for each test image–target pair and then averaged. Controlled comparisons share the evaluation split, valid region, and inference settings. Reference selection details and metric definitions are provided in Appendix D, and optimization settings in Appendix C.

5.1 Segmentation Across Modalities and Targets

Table 1 compares our unified model with published references across seven medical optical imaging modalities. The shared image editing paradigm maintains competitive performance across datasets and targets. For branching tubular structures, Figure 5a,b shows that GRADE improves overlap over Flow-only adaptation on the FIVES and DRIVE subsets of our dataset. Figure 6 further examines FIVES centerlines, showing improved structural alignment and coverage.

5.2 Dataset Utility and Segmentation as Editing

The role of OptiMedEdit-37K. Figure 4a compares multiple pretrained general-purpose image editing models with UniMedOpti trained on OptiMedEdit-37K. These baselines receive no fine-tuning on our dataset and achieve limited segmentation accuracy. Adapting the strongest baseline, Qwen-Image-Edit, on this corpus with our training strategy improves Dice by 58.57 points. This contrast highlights the dataset's value for adapting general-purpose image editing models to medical optical segmentation. Baseline configurations are provided in Appendix C.

Editing versus Direct Mask Prediction. Figure 4b compares source-preserving editing with direct binary mask generation under the same backbone, training data, split, initialization, and total training epochs. The 20.63-point Dice gain supports Segmentation as Editing. With total training epochs also

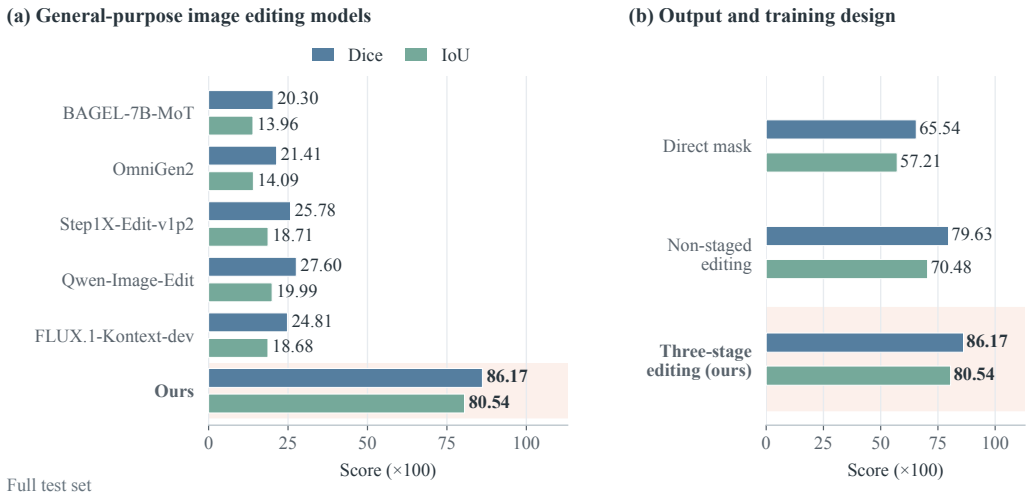

Figure 4: **Adapting image editing models with OptiMedEdit-37K.** (a) Five pretrained models without fine-tuning on OptiMedEdit-37K and our model trained on it. (b) Direct-mask prediction, non-staged editing, and the complete three-stage editing model on the full test set.

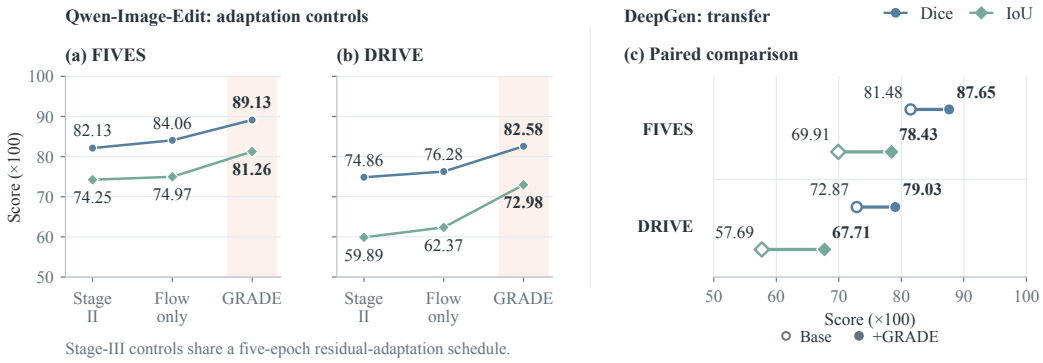

Figure 5: **Geometry adaptation and cross-backbone transfer.** (a,b) Qwen on FIVES and DRIVE: Stage II, Flow-only residual adaptation, and GRADE. (c) DeepGen on FIVES and DRIVE without (open markers) and with (filled markers) GRADE.

matched, three-stage editing gains 6.54 Dice points over non-staged editing, indicating that staged optimization better adapts the unified model's generative image editing capabilities to segmentation.

5.3 Contribution of Geometric Supervision

Figure 5a,b compares Stage II, Flow-only residuals, and GRADE to isolate geometric supervision. Residual variants share the frozen backbone, initialization, vessel training set, capacity, and five-epoch schedule. GRADE gains 5.07 Dice points on FIVES and 6.30 on DRIVE over Flow-only adaptation, with consistent IoU gains, supporting geometric supervision beyond Flow-only training.

5.4 Backbone Transfer and Geometric Supervision

Figure 5c tests transfer to DeepGen under matched inference and evaluation settings. GRADE improves Dice by approximately 6.2 points on both FIVES and DRIVE, with consistent IoU gains.

Figure 6 examines region losses and cbDice on FIVES using Dice, IoU, clDice (Shit et al., 2021), and centerline recall (CRecall) (Appendix D.3). Region losses improve overlap and centerline metrics, and adding cbDice further improves all four, including 4.73 points in clDice.

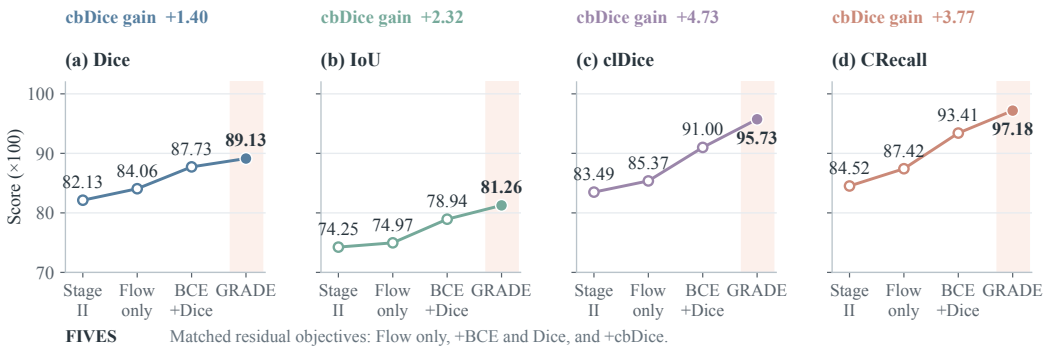

Figure 6: **Decomposing geometry supervision on FIVES.** Stage II is the shared backbone. Matched residual variants add region losses (BCE+Dice) and then cbDice to Flow-only training. Annotations show the additional gain from cbDice.

5.5 Qualitative Comparison

Figure 7 presents best-case examples across three modalities. Some pretrained models extend overlays beyond targets or alter surrounding tissue. UniMedOpti better follows vessel branches and object boundaries while preserving non-target content in these examples.

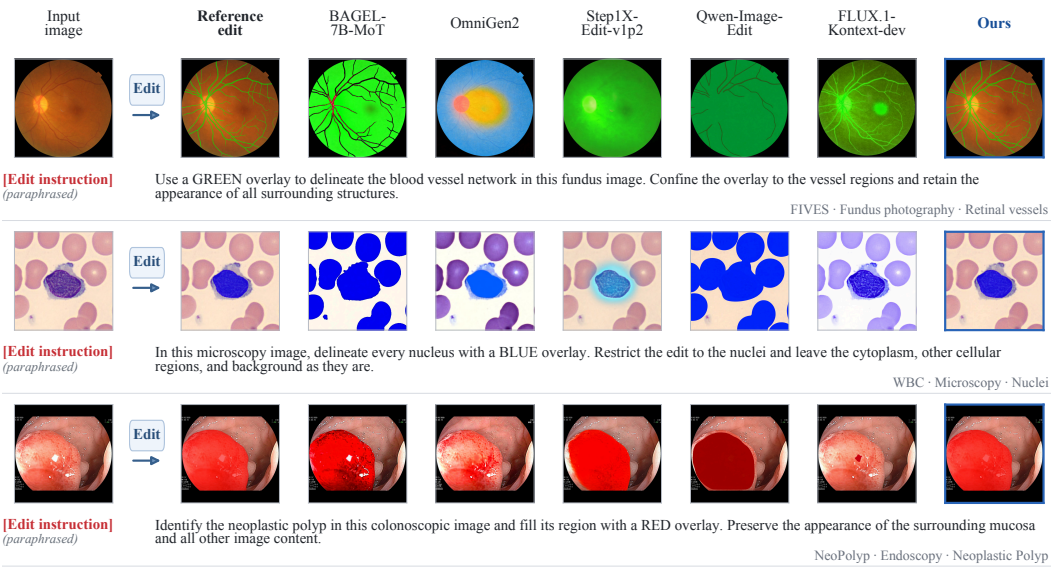

Figure 7: **Qualitative segmentation examples.** More comparisons in Appendix E.

6 Conclusion

We presented *Segmentation as Editing* for unified medical optical segmentation. OptiMedEdit-37K and staged training establish a shared image editing backbone, while GRADE adds geometry supervision through differentiable mask recovery and target-activated residuals. Experiments demonstrate competitive performance across modalities, improved vascular geometry beyond continued generative training, and transfer to a second backbone. These findings support a common editing interface that combines cross-modal segmentation with target-specific geometric learning. Generalization to additional datasets and imaging mechanisms remains an important direction for further study.

AI USE STATEMENT

AI tools were used to assist with manuscript preparation and language editing. The authors reviewed all AI-assisted content and take full responsibility for the final manuscript.

REFERENCES

Sharib Ali, Debesh Jha, Noha Ghatwary, Stefano Realdon, Renato Cannizzaro, Osama E. Salem, Dominique Lamarque, Christian Daul, Michael A. Riegler, Kim V. Anonsen, Andreas Petlund, Pål Halvorsen, Jens Rittscher, Thomas de Lange, and James E. East. A multi-centre polyp detection and segmentation dataset for generalisability assessment. *Scientific Data*, 10:75, 2023.

Pooya Ashtari, Shahryar Noei, Fateme Nateghi Haredasht, Jonathan H. Chen, Giuseppe Jurman, Aleksandra Pizurica, and Sabine Van Huffel. Deconver: A deconvolutional network for medical image segmentation. arXiv preprint arXiv:2504.00302, 2025.

Haixia Bai, Fuquan Wu, Yushuai Zhou, Shijing Wu, Ailing Sui, Zizhao Wu, and Zhiqing Chen. nnLoGoNet: A hybrid local–global network for retinal vessel segmentation with skeleton recall loss. *Scientific Reports*, 16:19297, 2026.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742, 2025.

Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. UniverSeg: Universal medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21438–21451, 2023.

Juan C. Caicedo, Allen Goodman, Kyle W. Karhohs, Beth A. Cimini, Jeanelle Ackerman, Marzieh Haghighi, CherKeng Heng, Tim Becker, Minh Doan, Claire McQuin, Mohammad Rohban, Shantanu Singh, and Anne E. Carpenter. Nucleus segmentation across imaging experiments: The 2018 data science bowl. *Nature Methods*, 16(12):1247–1253, 2019.

Carol Y. Cheung, Dejiang Xu, Ching-Yu Cheng, Charumathi Sabanayagam, Yih-Chung Tham, Marco Yu, Tyler Hyungtaek Rim, Chew Yian Chai, Bamini Gopinath, Paul Mitchell, Richie Poulton, Terrie E. Moffitt, Avshalom Caspi, Jason C. Yam, Clement C. Tham, Jost B. Jonas, Ya Xing Wang, Su Jeong Song, Louise M. Burrell, Omar Farouque, Ling Jun Li, Gavin Tan, Daniel S. W. Ting, Wynne Hsu, Mong Li Lee, and Tien Y. Wong. A deep-learning system for the assessment of cardiovascular disease risk via the measurement of retinal-vessel calibre. *Nature Biomedical Engineering*, 5(6):498–508, 2021.

Noel C. F. Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen W. Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration, 2019.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683, 2025.

DICOM Standards Committee. DICOM PS3.3 2026c: Information object definitions, section a.32 visible light image IODs, 2026.

Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. PraNet: Parallel reverse attention network for polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pp. 263–273, 2020.

Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benet, Ali Khuram, and Nasir Rajpoot. PanNuke: An open pan-cancer histology dataset for nuclei instance segmentation and classification. In *European Congress on Digital Pathology*, pp. 11–19, 2019.

Mathieu Gendarme, Annika Lambert, and Bachir El Debs. BriFiSeg: A deep learning-based method for semantic and instance segmentation of nuclei in brightfield images. arXiv preprint arXiv:2211.03072, 2022.

Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Houqiang Li, Han Hu, Dong Chen, and Baining Guo. InstructDiffusion: A generalist modeling interface for vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12709–12720, 2024.

Sojung Go, Jaemin Chae, Uichan Kim, Jongsoo Lim, Jooyoung Kim, Stephen Hogg, Emanuele Trucco, Sang Jun Park, and Soochahn Lee. Improving retinal vessel assessment precision by integrating deep learning with interactive editing and graphical modeling. *Scientific Reports*, 15: 41552, 2025.

Simon Graham, Quoc Dang Vu, Shan E. Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. HoVer-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, 2019.

Georgeos Hardo, Ruizhe Li, and Somenath Bakshi. Quantitative microbiology with widefield microscopy: Navigating optical artefacts for accurate interpretations. *npj Imaging*, 2:26, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Chao Huang, Qianyi Li, Jie Wen, and Bob Zhang. Omni-dimensional state space model-driven SAM for pixel-level anomaly detection. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 1152–1160, 2025.

Shahzaib Iqbal, Muhammad Zeeshan, Mehwish Mehmood, Tariq M. Khan, and Imran Razzak. TESL-Net: A transformer-enhanced CNN for accurate skin lesion segmentation. In *2024 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 313–320, 2024.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18:203–211, 2021.

Kai Jin, Xingru Huang, Jingxing Zhou, Yunxiang Li, Yan Yan, Yibao Sun, Qianni Zhang, Yaqi Wang, and Juan Ye. FIVES: A fundus image dataset for artificial intelligence based vessel segmentation. *Scientific Data*, 9:475, 2022.

Dinoy Johny, V. Chandan, Shubham Loni, A. Devika Gireesh, Vasarla Saimahesh, Kalpana George, Manju Anup, S. S. Anup, and P. B. Jayaraj. Structure-aware learnable multi-scale attention for retinal vessel analysis. *Scientific Reports*, 2026.

Zeki Kuş and Musa Aydin. MedSegBench: A comprehensive benchmark for medical image segmentation in diverse data modalities. *Scientific Data*, 11:1283, 2024.

Yongfan Lai, Wen Qian, Bo Liu, Hongyan Li, Hao Luo, Fan Wang, Bohan Zhuang, and Shenda Hong. MieDB-100k: A comprehensive dataset for medical image editing, 2026.

Liangzhi Li, Manisha Verma, Yuta Nakashima, Hajime Nagahara, and Ryo Kawasaki. IterNet: Retinal image segmentation utilizing structural redundancy in vessel networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3656–3665, 2020.

Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, et al. HealthGPT: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37975–37995. PMLR, 2025.

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

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1X-Edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761, 2025.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15:654, 2024.

Markus Marks, Uriah Israel, Rohit Dilip, Qilin Li, Changhua Yu, Emily Laubscher, Ahamed Iqbal, Elora Pradhan, Ada Ates, Martin Abt, Caitlin Brown, Edward Pao, Shenyi Li, Alexander Pearson-Goulart, Pietro Perona, Georgia Gkioxari, Ross Barnowski, Yisong Yue, and David Van Valen. CellSAM: A foundation model for cell segmentation. *Nature Methods*, 22(12):2585–2593, 2025.

National Institute of Biomedical Imaging and Bioengineering. Optical imaging, 2025.

Junzhi Ning, Wei Li, Cheng Tang, Jiashi Lin, Chenglong Ma, Chaoyang Zhang, Jiyao Liu, Ying Chen, Shujian Gao, Yuandong Pu, Huihui Xu, Chenhui Gou, Ziyan Huang, Yi Xin, Qi Qin, Diping Song, Bin Fu, Guang Yang, Yuanfeng Ji, Tianbin Li, Yanzhou Su, Jin Ye, Shixiang Tang, Zhongying Deng, Lihao Liu, Ming Hu, and Junjun He. UniMedVL: Unifying medical multimodal understanding and generation through observation–knowledge–analysis. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Thomas Porter and Tom Duff. Compositing digital images. In *Proceedings of the 11th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 253–259. ACM, 1984.

Yaolei Qi, Yuting He, Xiaoming Qi, Yuan Zhang, and Guanyu Yang. Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6070–6079, 2023.

Zhongxi Qiu, Yan Hu, Xiaoshan Chen, Dan Zeng, Qingyong Hu, and Jiang Liu. Rethinking dual-stream super-resolution semantic learning in medical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):451–464, 2024.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, 2026.

Guoqiang Ren, Zhen Chen, Pengxiang Su, Da Li, Xiaoping Yang, Di Gai, Xin Wei, Weifeng Xu, Hongping Chen, Xiaoguang Zhao, Xiaofei Wang, Pengfei Liu, Honghua Ye, and Yanfeng Ma. CFG-MambaNet: Contextual and frequency-guided Mamba network for medical image segmentation. *npj Digital Medicine*, 9:202, 2026.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, 2015.

Pengcheng Shi, Jiesi Hu, Yanwu Yang, Zilve Gao, Wei Liu, and Ting Ma. Centerline boundary dice loss for vascular segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15008 of *Lecture Notes in Computer Science*, pp. 46–56. Springer Nature Switzerland, 2024.

Suprosanna Shit, Johannes C. Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien P. W. Pluim, Ulrich Bauer, and Bjoern H. Menze. clDice: A novel topology-preserving loss function for tubular structure segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16560–16569, 2021.

Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: A generalist algorithm for cellular segmentation. *Nature Methods*, 18(1):100–106, 2021.

Zhiqiang Tian, Yaoyue Zheng, Xiaojian Li, Shaoyi Du, and Xiayu Xu. Graph convolutional network based optic disc and cup segmentation on fundus images. *Biomedical Optics Express*, 11(6): 3043–3057, 2020.

Dianyi Wang, Ruihang Li, Feng Han, Chaofan Ma, Wei Song, Siyuan Wang, Yibin Wang, Yi Xin, Hongjian Liu, Zhixiong Zhang, et al. DeepGen 1.0: A lightweight unified multimodal model for advancing image generation and editing. arXiv preprint arXiv:2602.12205, 2026.

Guoxin Wang, Jun Zhao, Xinyi Liu, Yanbo Liu, Xuyang Cao, Chao Li, Zhuoyun Liu, Qintian Sun, Fangru Zhou, Haoqiang Xing, and Zhenhong Yang. Citrus-V: Advancing medical foundation models with unified medical image grounding for clinical reasoning. arXiv preprint arXiv:2509.19090, 2025.

Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6830–6839, 2023a.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. SegGPT: Towards segmenting everything in context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1130–1140, 2023b.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-Image technical report, 2025a.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. OmniGen2: Towards instruction-aligned multimodal generation. arXiv preprint arXiv:2506.18871, 2025b.

Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Sid Kiblawi, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature Methods*, 22(1): 166–176, 2025.

Yaoyue Zheng, Xuetao Zhang, Xiayu Xu, Zhiqiang Tian, and Shaoyi Du. Deep level set method for optic disc and cup segmentation on fundus images. *Biomedical Optics Express*, 12(11):6969–6983, 2021.

A OPTIMEDEDIT-37K: DATA CONSTRUCTION DETAILS

This appendix expands the construction procedure in Section 3 and Figure 2(b). We detail source-level organization, annotation semantics, constrained instruction rephrasing, and checks on the stored editing targets. The compositing rule is given in Eq. (1). Its inverse and evaluation use are treated separately in Appendices B.4 and D.1.

A.1 SOURCE ORGANIZATION AND PROVENANCE

Record organization. The final corpus contains 37,059 editing samples from 29 datasets across seven modalities, with 29,727 training and 7,332 test samples. We organize the source data into target-specific editing records. Each image–instruction–edit triplet is associated with its source identity, imaging modality, target label, and partition assignment. Annotation-derived records additionally retain the original image and mask references, linking the editing supervision to its source annotation. Table A.1 summarizes the source datasets, training/test counts, segmentation targets, and recorded licenses.

Multi-target expansion and partition consistency. An image with several annotated targets yields one record per target, with the source-image identity and partition shared across these records. The reported counts describe the final assembled records, which need not correspond one-to-one to unique source images. Each record retains its assigned training or test subset throughout construction. Target expansion and instruction rephrasing preserve the underlying annotation and partition assignment.

A.2 TARGET CANONICALIZATION AND ANNOTATION HANDLING

Target vocabulary. Source labels are mapped to shared target names, with their original wording and modality retained in the annotation-derived metadata. For example, a dermoscopic *lesion* maps to *Skin Lesion*, while *neoplastic cells* and *inflammatory cells* remain distinct targets. Vascular targets use the common category *Vessel*. The source annotation defines the spatial support, and modality context disambiguates broader names such as *nuclei* and *organism*.

Mask preparation. Each target mask defines a binary support, with nonzero values treated as foreground. Instances of the requested class share this support, while other classes form separate records. We preserve the source mask convention without merging classes or filling holes. Figure A1 illustrates this organization for anatomical regions and cell categories. Source images are decoded as RGB and masks as single-channel arrays. When their dimensions differ, nearest-neighbor interpolation aligns the mask to the source grid before binarization. The aligned support is shared across RGB channels, and any subsequent training-time geometric transform must be applied consistently to the source, edit, and supervision mask.

A.3 CONSTRAINED INSTRUCTION REPHRASING

For each target-specific record, a canonical instruction specifies the segmentation target, imaging modality, overlay color, and requirement to preserve non-target content. Qwen3.6-35B-A3B receives this instruction together with its semantic fields as text input and produces alternative imperative phrasings (Qwen Team, 2026). Verb choice, clause order, and sentence structure may vary, while the target reference and rendering operation remain fixed. The returned candidates are checked for agreement with the record's target, modality, color, and preservation requirement, followed by exact-string deduplication. Accepted phrases remain bound to the corresponding target-specific supervision, allowing linguistic diversification without changing the source image or mask. Figure A2 contrasts valid rephrasings with seven types of semantic violation. Figure A3 gives an instantiated prompt specification and six illustrative phrasings for a dermoscopy record, while Figure A4 pairs stored examples with rephrased instructions. The corpus contains 3,725 distinct instruction strings.

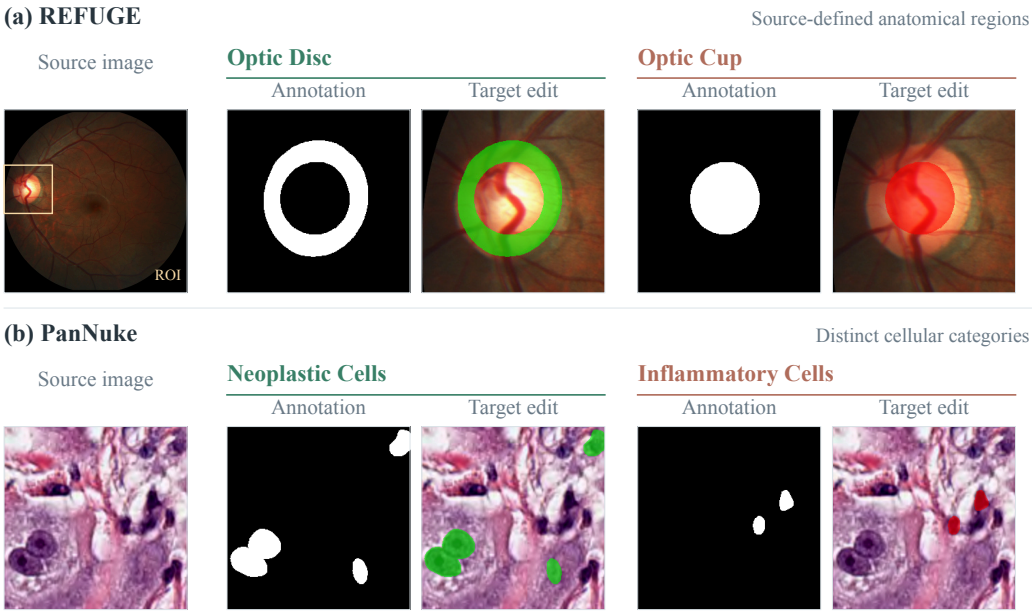

Figure A1: **One source image, separate target-specific records.** Each row pairs a source image with two annotations and their target edits. (a) REFUGE uses the supplied disc and cup supports. The disc-labeled support excludes the cup, and both detail pairs depict the marked source region. (b) PanNuke shows separate neoplastic-cell and inflammatory-cell supports at full image extent. These are stored construction examples, not model predictions.

Semantic consistency under rephrasing

Preserve the requested edit while varying its expression.

Target	Modality	Overlay color	Non-target content
Skin lesion	**Dermoscopy**	**GREEN**	**Unchanged**

Accept Different phrasing, identical editing semantics

- ✓ Apply a **GREEN** overlay to the **skin lesion** in the dermoscopy image, leaving all non-target content unchanged.
- ✓ Keep all non-target content in the dermoscopy image intact while highlighting the **cutaneous lesion** with a **GREEN** overlay.

Reject or rewrite One changed requirement is enough to invalidate a candidate.

Violation	Changed instruction fragment	Required meaning
✗ Target substitution	Overlay the **hair** in GREEN ...	Skin lesion or cutaneous lesion
✗ Modality mismatch	... in the **endoscopy** image ...	Dermoscopy
✗ Color mismatch	... mark the skin lesion in **RED** ...	GREEN overlay
✗ Output substitution	**Return a separate mask** of the skin lesion ...	Overlay on the source image
✗ Expanded edit scope	... cover the lesion **and surrounding skin** ...	Requested target only
✗ Background change	... and **increase the background contrast.**	Non-target content unchanged
✗ Unsupported diagnosis	... highlight the **malignant** skin lesion ...	No inferred diagnosis

Excerpts isolate the change. Omitted clauses follow the valid reference instruction.

Figure A2: **Semantic consistency under rephrasing.** Two valid phrasings are contrasted with seven violations of the editing specification, each paired with its required meaning. Highlighted fragments isolate the changes. These authored examples illustrate acceptance criteria, not outputs from a model run.

Instruction rephrasing for Segmentation as Editing

Qwen3.6-35B-A3B | An instantiated OptiMedEdit-37K prompt

Rewrite the canonical instruction as six natural English commands for Segmentation as Editing. Each command must describe the same target-specific color overlay on the supplied medical image. Change the phrasing while preserving the requested operation and its spatial scope. Use the fields below as the sole basis for rewriting, without inferring a diagnosis or inventing image findings.

01 Record to be rephrased

Target	**skin lesion**	Allowed alias	**cutaneous lesion**
Modality	**dermoscopy**	Overlay color	**GREEN**

Canonical edit: Render a **GREEN** overlay over the **skin lesion** in the dermoscopy image without altering any non-target content.

02 Semantic requirements

Target identity	Keep skin lesion or cutaneous lesion. Retain any supplied anatomical qualifier.
Imaging context	Specify dermoscopy without inventing acquisition details or visible findings.
Overlay operation	Request a GREEN overlay on the source image, rather than a separate mask image.
Spatial scope	Restrict the edit to the requested target. Do not add or remove target regions.
Preservation	Explicitly leave all non-target content unchanged, including other structures.
Medical meaning	Do not infer malignancy, severity, treatment, or any additional diagnosis.

03 Ways to vary the wording

Use different action verbs and sentence openings. Alternate between target-first, image-first, and preservation-first phrasing, and vary clause order without changing the edit. Prefer concise, natural commands over repetitive templates. Use only the allowed alias when replacing the target name.

04 Review before returning the candidates

Read each candidate as a standalone instruction. Check the target, modality, color, edit scope, and preservation requirement, then rewrite any mismatch. Remove duplicate strings and vary near-identical openings before returning exactly six complete instructions.

05 Output format and six example phrasings

Authored illustration of the requested JSON response

```
{"instructions": [
  "Apply a GREEN overlay to the skin lesion in the dermoscopy image, leaving all non-target content unchanged.",
  "In the dermoscopy image, mark the cutaneous lesion in GREEN and preserve every non-target region.",
  "Keep all non-target content in the dermoscopy image intact while highlighting the skin lesion with a GREEN overlay.",
  "Render a GREEN mask over the skin lesion on the dermoscopy image without altering content outside the target.",
  "Identify the skin lesion in the dermoscopy image and overlay it in GREEN, preserving the rest of the image.",
  "Restrict the GREEN overlay to the cutaneous lesion in the dermoscopy image and keep all other content unchanged."
]}
```

Six distinct strings. No commentary or additional fields.

Figure A3: **Constrained instruction rephrasing for OptiMedEdit-37K.** An instantiated dermoscopy prompt specifies fixed semantics, allowed variation, and review criteria. Six authored outputs illustrate different phrasings of the same edit, rather than a historical generation log.

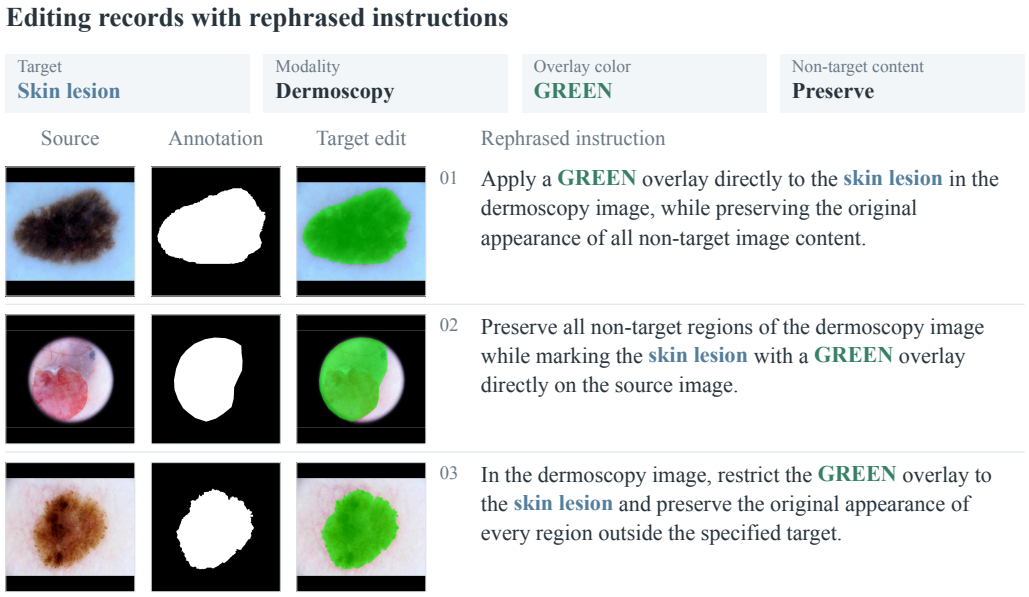

Figure A4: **Editing records with rephrased instructions.** Each row pairs a stored ISIC 2018 source image, its annotation, and its target edit with a semantically equivalent instruction rewritten for this illustration. The records share the skin-lesion target, dermoscopy modality, GREEN overlay, and background-preservation requirement. Target and color expressions are highlighted in blue and green. All images show the full source extent. The target edits are supervision examples, not model predictions.

A.4 Target Rendering and Record Validation

Rendering and storage. Annotation-derived edits follow Eq. (1), using the instruction-specified RGB color and opacity $\rho = 0.6$. Compositing is evaluated in floating point on the aligned source grid before model-side resizing. The result is clipped to the image range and converted to unsigned 8-bit RGB for PNG storage. Non-target pixels are copied directly from the decoded source. Red, green, and blue are rendering attributes assigned across records, not fixed anatomical class identifiers.

Pair verification. Checks require a shared spatial grid and agreement between the instruction, annotation, and rendering color. For saved annotation-derived edits, background pixels must match the decoded source exactly, and foreground values must agree with floating-point compositing to within one 8-bit intensity level per channel. The original mask is used for forward verification, including pixels where the source RGB value equals the overlay color and the edit direction vanishes. Thus, verification does not depend on detecting a nonzero image difference at every foreground pixel.

Record binding. The verified pair is linked to its instruction, target, color, opacity, and provenance fields. Annotation-derived entries retain source-image and mask paths, while collection-derived pairs retain their available source identifiers. Direct array checks for the examples in Figures A1 and A4 confirm exact background preservation and a maximum foreground discrepancy below one 8-bit intensity level. These checks concern the displayed construction examples only. Evaluation uses the source annotations as reference masks, as described in Appendix D.1.

Table A.1: Source datasets, training/test sample counts, imaging modalities, and segmentation targets in OptiMedEdit-37K. Vessel targets use a unified category name. License entries reproduce the recorded source/release metadata and do not imply a common redistribution license.

Dataset	Modality	Train	Test	Target	Recorded License
BBBC010	Microscopy	80	20	Organism	CC BY 4.0
BKAI-IGH	Endoscopy	800	200	Polyp	CC BY 4.0
BriFiSeg	Microscopy	1,200	240	Nuclei	CC BY 4.0
CellNuclei	Microscopy	536	134	Nuclei	CC BY 4.0
CHASE_DB1	Fundus	21	7	Vessel	CC BY 4.0
CystoidFluid	OCT	804	202	Cystic Fluid	CC BY 4.0
DeepBacs	Microscopy	19	15	Organism	CC BY 4.0
DRIVE	Fundus	20	20	Vessel	CC BY 4.0
DynamicNuclearNet	Microscopy	6,367	717	Nuclei	CC BY 4.0
FIVES	Fundus	600	200	Vessel	CC BY 4.0
G1020	Fundus	1,448	362	Optic Cup; Optic Disc	No formal license stated (research use invited)
GlaS	Pathology	85	80	Glandular Structure	Custom (Research Only; Non-Commercial)
IDRiD	Fundus	53	27	Optic Disc	CC BY 4.0
ISIC2016	Dermoscopy	900	379	Skin Lesion	CC0
ISIC2018	Dermoscopy	2,694	1,000	Skin Lesion	CC0
Kvasir-SEG	Endoscopy	800	200	Polyp	Custom (Research Only)
MoNuSAC	Microscopy	209	101	Nuclei	CC BY 4.0
NeoPolyp	Endoscopy	1,640	410	Polyp; Neoplastic Polyp; Non-neoplastic Polyp	Custom (Subject to Competition Rules)
NuSeT	Microscopy	2,725	683	Nuclei	CC BY 4.0
OCT-CME	OCT	1,177	283	Retinal Edema	CC BY-NC-SA 4.0
PanNuke	Pathology	4,249	858	Connective Tissue Cells; Epithelial Cells; Inflammatory Cells; Neoplastic Cells	CC BY-NC-SA 4.0
PolypGen	Endoscopy	1,112	299	Polyp	CC BY 4.0
REFUGE	Fundus	800	400	Optic Cup; Optic Disc	CC BY 4.0
RoboTool	Surgical imaging	400	100	Surgical Tool	CC BY 4.0
STARE	Fundus	10	10	Vessel	Unspecified
TNBCNuclei	Microscopy	40	10	Nuclei	CC BY 4.0
UWaterloo SkinCancer	Dermoscopy	162	44	Melanoma; Skin Lesion	CC BY 4.0
WBC	Microscopy	320	80	Nuclei	CC BY 4.0
YeaZ	Microscopy	456	251	Organism	CC BY 4.0
Total (29 datasets)		**29,727**	**7,332**	**37,059 samples**	

B METHOD DETAILS AND ANALYSIS OF GRADE

This appendix details parameter ownership, recovery, and gradient transport in Section 4 and Figure 3. Flow Matching, low-rank parameterization, and cbDice are established components (Lipman et al., 2023; Hu et al., 2022; Shi et al., 2024). Our method connects geometric supervision on analytically recovered masks to target-activated residuals through the model's native edit output.

B.1 STAGED OPTIMIZATION AND PARAMETER ALLOCATION

Each training stage starts from the preceding checkpoint. The multimodal encoder H and VAE (E, D) remain fixed throughout. The stages differ in their trainable capacity and supervision:

Stage	Updated parameters	Training records	Supervision and computation
I	Condition connector	All training triplets	Flow Matching, Eq. (4)
II	Complete MMDiT, including the connector	All training triplets	Flow Matching, Eq. (4)
III	Residual parameters ϕ only	Vessel-target triplets	Joint objective, Eq. (7), with gradients through frozen D

Stage III retains a single backbone with target-gated residuals. Although D is frozen, its input Jacobian must be retained so that edit-space errors reach ϕ. The numerical optimization settings are collected in Appendix C.

B.2 CLEAN-TARGET ESTIMATION AND ERROR TRANSPORT

The reconstruction in Eq. (6) follows algebraically from the linear Flow Matching path. Define the velocity error $\Delta\mathbf{v} = \widehat{\mathbf{v}} - \mathbf{v}^\star$, with all latent arrays vectorized in $\mathbb{R}^L$. Combining Eqs. (4) and (6) gives

$$\widehat{\mathbf{z}}_y - \mathbf{z}_y = -\sigma\Delta\mathbf{v}, \qquad \|\widehat{\mathbf{z}}_y - \mathbf{z}_y\|_2 = \sigma\|\Delta\mathbf{v}\|_2. \tag{A1}$$

Indeed, the sampled latent satisfies $\mathbf{z}_\sigma = \mathbf{z}_y + \sigma\mathbf{v}^\star$, so subtracting $\sigma\widehat{\mathbf{v}}$ yields the identity directly. For fixed sampled inputs, the reconstruction Jacobian with respect to $\widehat{\mathbf{v}}$ is $-\sigma\mathbf{I}_L$. This gives the sign and noise-level factor in Eq. (8).

A perfect velocity prediction recovers the target latent at any sampled noise level. With an imperfect prediction, the estimated endpoint exposes its error to image-space supervision through D (Figure A5a). At $\sigma = 0$, the estimate equals the encoded target and its auxiliary gradient with respect to the velocity vanishes. At other noise levels, it provides a differentiable training estimate that need not match the output of the iterative inference sampler.

B.3 RESIDUAL CAPACITY AND PRESERVATION OF THE SHARED PATH

Capacity and initialization. The factorization in Eq. (5) gives $\mathrm{rank}(\Delta\mathbf{W}_\ell) \le r$. An adapted projection adds $r(d_{\mathrm{in}}+d_{\mathrm{out}})$ parameters, corresponding to the fraction $\eta_\ell = r(d_{\mathrm{in}}+d_{\mathrm{out}})/(d_{\mathrm{in}}d_{\mathrm{out}})$ of that dense projection. For a square projection of width d, this reduces to $2r/d$ (Figure A6c). This is a per-projection parameter count, not a claim about the total model's memory or runtime. As in LoRA (Hu et al., 2022), $\mathbf{B}_\ell$ is initialized to zero and $\mathbf{A}_\ell$ to nonzero random values. The initial residual is therefore zero while allowing a gradient into $\mathbf{B}_\ell$.

Functional preservation. Let $\boldsymbol{\theta}^{(2)}$ denote the frozen Stage-II parameters. For any instruction whose normalized target has $g_{\mathrm{ves}}(p) = 0$, every residual vanishes, giving the explicit identity

$$f_{\boldsymbol{\theta}^{(2)},\boldsymbol{\phi}}(\mathbf{z}_\sigma, \mathbf{z}_x, \mathbf{h}, \sigma; p) = f_{\boldsymbol{\theta}^{(2)}}(\mathbf{z}_\sigma, \mathbf{z}_x, \mathbf{h}, \sigma). \tag{A2}$$

The equality holds for arbitrary trained ϕ when the inputs and remaining computation are identical. It follows layer by layer from the zero gate and frozen shared parameters. Using the same initial noise and sampler, this also preserves the non-vessel sampling trajectory in exact arithmetic. The gate follows the requested target, not the mere presence of a vessel in an image. This preservation property concerns the disabled residual path. It does not guarantee improvement when the gate is active.

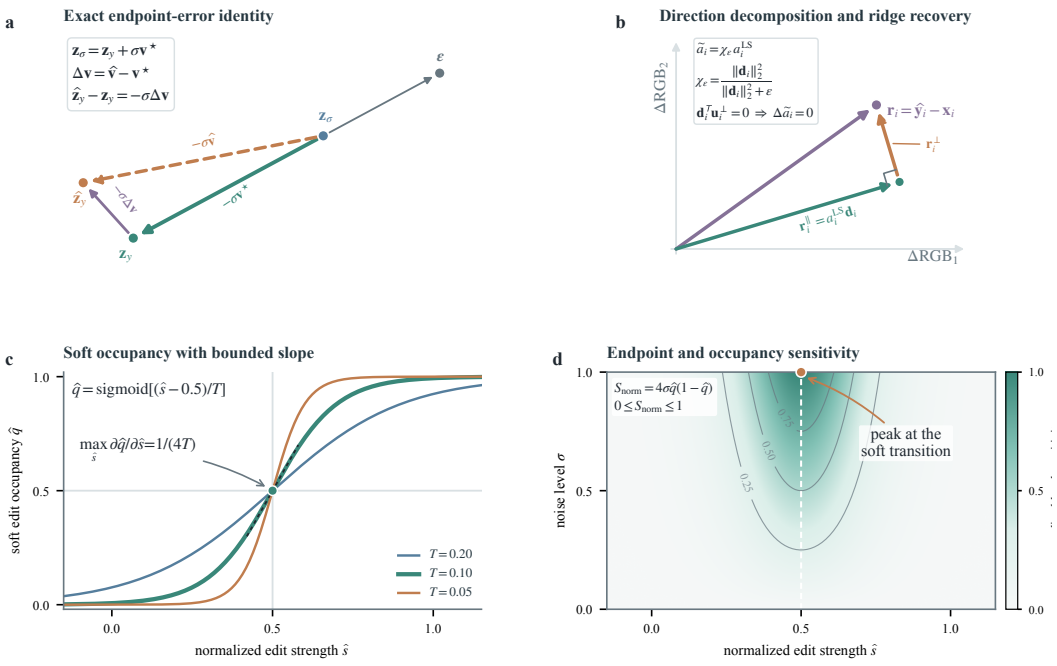

Figure A5: **Endpoint reconstruction and edit recovery.** (a) The endpoint error is exactly the velocity error scaled by $-\sigma$. (b) The RGB residual separates into components parallel and orthogonal to the edit direction. Ridge recovery shrinks the parallel coefficient and ignores the orthogonal component. (c) Temperature controls the transition from normalized edit strength to soft occupancy. (d) $4\sigma\widehat{q}_i(1-\widehat{q}_i)$ isolates the endpoint and sigmoid factors, not the complete loss gradient. Vector diagrams use illustrative two-dimensional coordinates. Curves and fields are analytical evaluations, not experimental measurements.

B.4 Analytic Recovery and Continuous Supervision

Scalar recovery from an RGB edit. We reuse the direction $\mathbf{d}_i$ defined in Eq. (1), and write $\mathbf{r}_i = \widehat{\mathbf{y}}_i - \mathbf{x}_i$ for the RGB edit residual. The unclipped coefficient in Eq. (2) uniquely minimizes

$$\underset{a\in\mathbb{R}}{\text{minimize}} \quad \|\mathbf{r}_i - a\mathbf{d}_i\|_2^2 + \varepsilon a^2. \tag{A3}$$

Its second derivative is $2(\|\mathbf{d}_i\|_2^2 + \varepsilon) > 0$. Setting the first derivative to zero gives Eq. (2), so no separate inverse model is fitted. This is the regularized scalar form of alpha-deblending used for medical image edits (Lai et al., 2026).

For nonzero $\mathbf{d}_i$, decompose $\mathbf{r}_i$ into its Euclidean parallel and orthogonal components. If a_i^{LS} denotes the unregularized parallel coefficient, ridge recovery gives $\widetilde{a}_i = \chi_\varepsilon(s_d)a_i^{\mathrm{LS}}$, with $s_d = \|\mathbf{d}_i\|_2$ and

$$\chi_\varepsilon(s_d) = \frac{s_d^2}{s_d^2+\varepsilon}, \qquad \gamma_\varepsilon(s_d) = \frac{s_d}{s_d^2+\varepsilon} \le \frac{1}{2\sqrt{\varepsilon}}. \tag{A4}$$

The first factor quantifies shrinkage, while the second is the operator norm of the RGB-to-strength derivative. Consequently, an RGB perturbation $\mathbf{u}_i$ changes the recovered strength by at most $\gamma_\varepsilon(s_d)\|\mathbf{u}_i\|_2$, and has no effect when $\mathbf{d}_i^{\mathsf{T}}\mathbf{u}_i = 0$. Figure A5b distinguishes the Euclidean decomposition from the subsequent ridge shrinkage.

For an ideal target edit, the recovered coefficient is $\rho m_i \chi_\varepsilon(s_d)$. It therefore approaches the rendered strength ρm_i when $s_d^2 \gg \varepsilon$, with a controlled bias when contrast is small. If $\mathbf{d}_i = \mathbf{0}$, the image contains no information about the overlay strength at that pixel and the regularized solution is zero. Regularization defines the inverse in this case but does not restore missing information. The derivative gain peaks at $s_d = \sqrt{\varepsilon}$, where $\chi_\varepsilon = 1/2$ (Figure A6b).

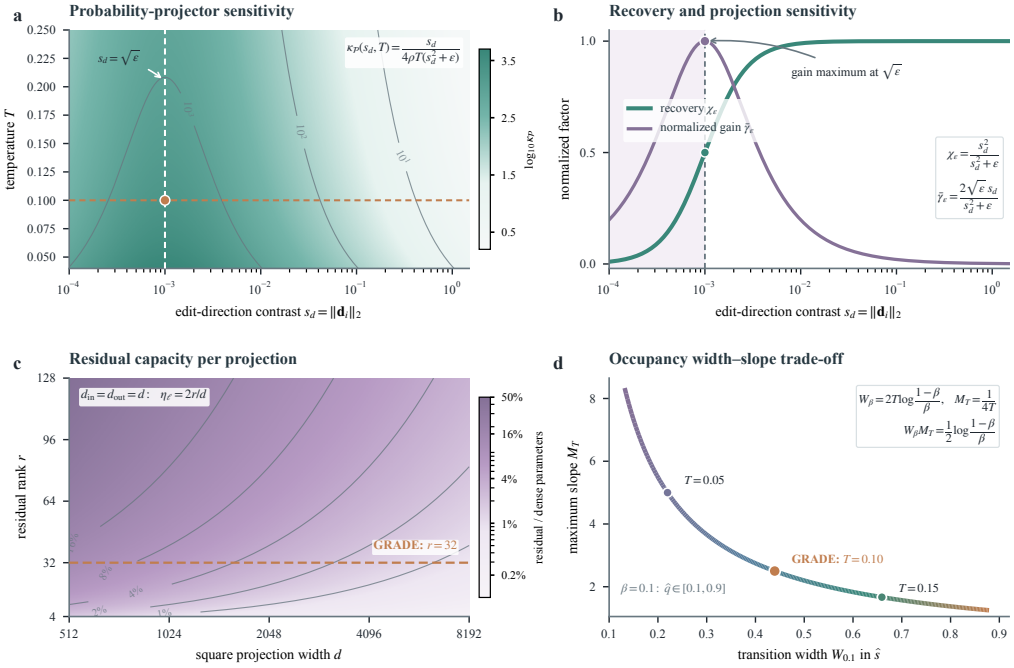

Figure A6: **Analytical trade-offs in the recovery and residual components.** (a) Upper bound on the probability projector's local sensitivity as edit contrast and temperature vary. (b) Ridge shrinkage χ_ε and normalized derivative gain $2\sqrt{\varepsilon}\,\gamma_\varepsilon$. (c) Residual-to-dense parameter ratio for a square projection, rather than whole-model overhead. (d) Occupancy transition width and maximum sigmoid slope as temperature changes. Highlighted settings are specified in Appendix C. All panels evaluate analytical expressions and do not constitute hyperparameter ablations.

Continuous training coordinate. The differentiable map described in Section 4.3 is

$$\widehat{s}_i = \widetilde{a}_i/\rho, \qquad \ell_i = (\widehat{s}_i - b)/T, \qquad \widehat{q}_i = \mathrm{sigmoid}(\ell_i), \qquad \frac{\partial \widehat{q}_i}{\partial \widehat{s}_i} = \frac{\widehat{q}_i(1-\widehat{q}_i)}{T} \le \frac{1}{4T}. \tag{A5}$$

Neither clipping nor thresholding is applied to $\widetilde{a}_i$ along this training path. Normalization converts the rendered foreground strength into an occupancy coordinate, with ideal background and foreground near zero and one, respectively. For $0 < \beta < 1/2$, the interval in $\widehat{s}_i$ over which $\widehat{q}_i$ changes from β to $1-\beta$ has width $W_\beta = 2T\log((1-\beta)/\beta)$. Together with the maximum slope $M_T = 1/(4T)$, this yields $W_\beta M_T = \frac{1}{2}\log((1-\beta)/\beta)$. This is a transition in edit strength, not the width of a spatial vessel boundary (Figure A6d).

Training probabilities and inference masks. The soft training map and the hard evaluation mask use different decision coordinates. The midpoint $\widehat{q}_i = 1/2$ corresponds to $\widetilde{a}_i = \rho b$. In contrast, inference applies the fixed threshold τ to the clipped strength in Section 3.3, as specified in Appendix D.1. For $0 < \tau < 1$, this is equivalent to testing $\widetilde{a}_i > \tau$, and equals thresholding $\widehat{q}_i$ at $\mathrm{sigmoid}((\tau/\rho - b)/T)$, not generally at $1/2$. The $1/2$ cutoff used to construct cbDice's training geometry is therefore not the evaluation threshold.

B.5 Geometry Supervision and Its Gradient Route

Region supervision. Let Ω be the valid image region and $N_\Omega > 0$ its pixel count.

The region losses in Eq. (7) are evaluated per image and then averaged over the batch:

$$\ell_{\mathrm{BCE}} = \frac{1}{N_\Omega}\sum_{i\in\Omega}\big[\mathrm{softplus}(\ell_i) - m_i\ell_i\big],$$
$$\ell_{\mathrm{Dice}} = 1 - \frac{2\sum_{i\in\Omega}\widehat{q}_i m_i + \delta}{\sum_{i\in\Omega}\widehat{q}_i + \sum_{i\in\Omega} m_i + \delta}. \tag{A6}$$

Here $\mathrm{softplus}(u) = \log(1 + \exp u)$ is evaluated in a numerically stable form, and $\delta > 0$ smooths the Dice ratio. BCE provides pixelwise foreground/background supervision, while Dice couples predictions through regional overlap.

Centerline and boundary supervision. We use cbDice (Shi et al., 2024) as an existing geometric loss, with its skeleton and distance construction held fixed during each backward pass. The implementation follows the authors' released formulation with linear inverse-radius weighting.[1] Valid-region masking is applied before computing geometry. For prediction and reference, indexed by $k \in \{p, g\}$, construct the binary mask, skeleton S_i^k, distance to background d_i^k, and skeleton radius $r_i^k = S_i^k d_i^k$. The prediction mask uses $\widehat{q}_i > 1/2$, and the reference uses m_i. All these geometry quantities are detached from automatic differentiation.

For completeness, define $R_k = \max(1, \max_i r_i^k)$ and $r_k^{\min} = \max(1, \min_i r_i^k)$ over the image grid, with zeros outside the corresponding mask. The three fixed geometry maps are

$$u_i^k = \frac{\min(d_i^k, R_k)}{R_k}, \qquad v_i^k = \frac{r_i^k}{R_k}, \qquad w_i^k = S_i^k \frac{R_k - r_i^k + r_k^{\min}}{R_k}. \tag{A7}$$

Use the continuous weights $\pi_i^p = \widehat{q}_i$ and $\pi_i^g = m_i$ to form distance, radius, and inverse-radius maps $Q_{v,i}^k = u_i^k \pi_i^k$, $Q_{sv,i}^k = v_i^k \pi_i^k$, and $Q_{s,i}^k = w_i^k \pi_i^k$. Unlike the geometry maps, their prediction-side probability factors retain gradients. Define the pointwise combination $C(a, b, s) = s\,a$ if $a \neq 0$ and $b = 0$, and $C(a, b, s) = s\,b$ otherwise. The smoothed precision, sensitivity, and loss are

$$\begin{aligned} P_{\mathrm{cb}} &= \frac{\sum_i Q_{s,i}^p Q_{v,i}^g + \delta_{\mathrm{cb}}}{\sum_i C(Q_{sv,i}^p, Q_{sv,i}^g, Q_{s,i}^p) + \delta_{\mathrm{cb}}}, \\ S_{\mathrm{cb}} &= \frac{\sum_i Q_{s,i}^g Q_{v,i}^p + \delta_{\mathrm{cb}}}{\sum_i C(Q_{sv,i}^g, Q_{sv,i}^p, Q_{s,i}^g) + \delta_{\mathrm{cb}}}, \qquad \ell_{\mathrm{cbDice}} = 1 - \frac{2P_{\mathrm{cb}} S_{\mathrm{cb}}}{P_{\mathrm{cb}} + S_{\mathrm{cb}}}. \end{aligned} \tag{A8}$$

Thus gradients pass through the soft prediction factors in both numerators and denominators, but not through thresholding, skeletonization, or distance transforms. The resulting coefficient is clipped to $[0, 1]$ for numerical protection. Geometry is recomputed at the next forward pass, making the implemented objective piecewise differentiable. For an empty reference mask, the cbDice contribution is zero and the region losses remain active. Likewise, BCE provides a learning signal even when an empty predicted skeleton gives little geometric supervision. This construction encourages centerline and local width agreement without imposing a hard connectivity constraint.

From edit-space errors to residual updates. To unpack Eq. (8), let $\mathbf{g}_\ell = \nabla_{\boldsymbol{\ell}} \ell_{\mathrm{aux}}$ be the logit-space loss gradient on a fixed differentiable branch. At each pixel, and then at the velocity output, the intermediate gradients are

$$\nabla_{\widehat{\mathbf{y}}_i} \ell_{\mathrm{aux}} = \frac{g_{\ell,i}}{\rho T} \frac{\mathbf{d}_i}{\|\mathbf{d}_i\|_2^2 + \varepsilon}, \qquad \nabla_{\widehat{\mathbf{v}}} \ell_{\mathrm{aux}} = -\sigma \mathbf{J}_D^{\mathsf{T}} \nabla_{\widehat{\mathbf{y}}} \ell_{\mathrm{aux}}. \tag{A9}$$

Multiplication by $\mathbf{J}_{f,\boldsymbol{\phi}}^{\mathsf{T}}$ yields the main-text gradient formula. The first expression directs the auxiliary RGB gradient along the prescribed edit direction. Appearance changes orthogonal to that direction are invisible to the projector, which is one reason to retain the generative objective alongside geometry supervision. The decoder Jacobian includes the conversion to the RGB convention used by recovery. If this conversion clips out-of-range values, its derivative is zero on those clipped coordinates.

What the sensitivity bound establishes. For the probability-valued map $\mathcal{P} : \widehat{\mathbf{y}} \mapsto \widehat{\mathbf{q}}$, disjoint RGB pixel blocks and Eq. (A5) imply

$$\|\mathbf{J}_{\mathcal{P}}\|_2 \leq \max_i \frac{\|\mathbf{d}_i\|_2}{4\rho T(\|\mathbf{d}_i\|_2^2 + \varepsilon)} \leq \frac{1}{8\rho T\sqrt{\varepsilon}}. \tag{A10}$$

The logit-valued map has the analogous bound without the factor $1/4$. Consequently, Eq. (8) gives $\|\nabla_{\boldsymbol{\phi}} \ell_{\mathrm{aux}}\|_2 \leq \sigma \|\mathbf{J}_{f,\boldsymbol{\phi}}\|_2 \|\mathbf{J}_D\|_2 \|\mathbf{g}_\ell\|_2 / (2\rho T\sqrt{\varepsilon})$. This is a conditional bound on gradient transport. It does not bound the network Jacobians or establish convergence of the complete optimization. Figure A6a isolates the recovery component of this bound.

[1] cbDice reference implementation: `loss/cbdice_loss.py`, revision `a970c1d`.

The sigmoid derivative peaks at its occupancy transition, but the full auxiliary gradient also depends on the chosen loss. In particular, $\partial \ell_{\mathrm{BCE}} / \partial \ell_i = (\widehat{q}_i - m_i)/N_\Omega$, so a confident but incorrect prediction retains a strong BCE logit gradient. The sigmoid sensitivity field in Figure A5d should therefore not be interpreted as the full training-gradient magnitude.

B.6 Training Computation and Native Inference

Stage III uses four operations:

1. Encode the source, target edit, and instruction with frozen E and H, and sample the noise state as in Eq. (4).
2. Evaluate the residual-augmented MMDiT once. Use its velocity prediction for both Flow Matching and the endpoint estimate in Eq. (6).
3. Decode that estimate with frozen D while retaining its input gradient. Recover continuous logits and probabilities through Appendix B.4.
4. Evaluate region losses and construct detached cbDice geometry. Combine the losses using Eq. (7), then take one joint backward pass and update only ϕ.

The auxiliary branch adds decoder activation and geometry-processing costs, but requires no second MMDiT evaluation or unrolled inference trajectory.

Inference retains the target-gated MMDiT, native sampler, VAE decoding, and deterministic recovery in Section 3.3. It uses neither the training endpoint estimate nor reference masks or binary geometry.

C EXPERIMENTAL DETAILS

This appendix specifies the implementation and comparison settings for Section 5. We distinguish the primary Qwen-Image-Edit configuration, released pretrained baselines, and controlled adaptations of the shared backbone. Method derivations are given in Appendix B, and evaluation metrics and published-reference protocols in Appendix D.

C.1 TRAINING CONFIGURATION

Table C.1 collects the numerical settings for the Qwen-based experiments. Stages I and II use the 29,727-sample training partition summarized in Table A.1. Stages I and II share the optimization configuration and differ in trainable capacity and duration. Stage III starts from the Stage-II backbone and optimizes the residual parameters under the settings listed separately below. The parameter allocation and checkpoint progression are specified in Appendix B.1.

Table C.1: **Training and implementation configuration.** Settings for the primary Qwen-Image-Edit backbone. The final group specifies GRADE adaptation.

Setting	Configuration
Backbone and training schedule	
Generative backbone	Qwen-Image-Edit, approximately 20B MMDiT parameters
Training duration	Stage I: 2 epochs Stage II: 15 epochs Stage III: 5 epochs
Optimization and distributed execution	
Hardware	48 $\times$ NVIDIA A100 80GB
Distributed engine	Accelerate with DeepSpeed ZeRO-2
Precision and memory	BF16 mixed precision, CPU optimizer offload, activation checkpointing
Optimizer	AdamW: $\beta = (0.9, 0.999)$, epsilon 10^{-8}, weight decay 0.01
Learning rate	10^{-5}, with the first 5 updates at one-third rate and a constant rate thereafter
Effective batch size	1 sample per GPU, accumulation factor 1, global batch size 48
Input geometry and conditioning	
Stages I/II resolution	Aspect ratio preserved, at most $1{,}048{,}576$ pixels. Both spatial dimensions are multiples of 16.
Stage III resolution	1024×1024
Diffusion input	Noised target-edit latent, conditioned on source image and instruction
Source conditioning	Unnoised source latent, with source timestep fixed at zero
Training guidance	Classifier-free guidance scale 1
GRADE residual adaptation	
Residual configuration	Rank $r = 32$, scale $\alpha = 32$, dropout 0, effective scale $\alpha/r = 1$
Residual placement	Attention, feed-forward, and modulation projections
Parameter precision	Frozen shared modules in BF16, residual and optimizer states in FP32
Stage III optimizer	AdamW, learning rate 10^{-5}, weight decay 0.01, maximum gradient norm 1.0
Stage III execution	ZeRO-2 with CPU optimizer offload, random seed 20260723
Full objective weights	$\lambda_{\mathrm{GRADE}} = 0.1$, $(\omega_{\mathrm{B}}, \omega_{\mathrm{D}}, \omega_{\mathrm{C}}) = (1/2, 1/6, 1/3)$
Recovery parameters	$\rho = 0.6$, $b = 1/2$, $T = 0.1$, $\varepsilon = 10^{-6}$
Loss stabilizers	Dice: 10^{-5} cbDice: 10^{-3}
Skeleton construction	10 iterations on detached binary masks

C.2 Backbones and Input Preparation

Table C.2 identifies the model releases used in Figure 4a and the transfer study. Qwen-Image-Edit supplies both the pretrained baseline and the initialization for UniMedOpti. The five pretrained editing baselines are evaluated without fine-tuning on OptiMedEdit-37K. The corresponding comparison assesses medical segmentation before and after task adaptation, while the within-backbone controls below examine specific design choices.

Table C.2: **Backbone and baseline inventory.** Model releases and their roles in the experiments.

Model	Release / variant	Experimental role
Qwen-Image-Edit (Wu et al., 2025a)	2511 checkpoint	Primary backbone and pretrained baseline
BAGEL (Deng et al., 2025)	7B-MoT	Pretrained editing baseline
OmniGen2 (Wu et al., 2025b)	OmniGen2	Pretrained editing baseline
Step1X-Edit (Liu et al., 2025)	v1p2	Pretrained editing baseline
FLUX.1 Kontext (Black Forest Labs et al., 2025)	Kontext-dev	Pretrained editing baseline
DeepGen (Wang et al., 2026)	1.0	Second backbone for transfer

The Qwen training resolutions are stage-specific: Stages I and II use the aspect-ratio-preserving pixel budget in Table C.1, whereas Stage III uses a fixed square resolution. The target edit is the diffusion variable, and the source image and instruction provide conditioning. The source latent remains unnoised. Evaluation uses the reference spatial grid and valid region specified in Appendix D.1.

C.3 GRADE Implementation Settings

The residual configuration and complete-objective coefficients are centralized in Table C.1. The opacity ρ matches target rendering in Eq. (1), while b, T, and ε specify the continuous recovery map in Appendix B.4. Training uses the unclipped edit strength, and BCE is evaluated from logits. The Dice and cbDice stabilizers are distinct from the RGB projection stabilizer. The listed skeleton iterations belong to the detached geometry construction used by the cbDice loss, not the evaluation-time centerline metrics in Appendix D.3.

During residual adaptation, shared model parameters are fixed and only ϕ is updated. The decoder retains its input gradient despite having frozen weights, as detailed in Appendix B.6. Flow Matching and auxiliary supervision reuse the same noise level, noise realization, and velocity prediction, with the same backbone timestep weight $w(\sigma)$. These settings connect the numerical configuration here to the objective in Eq. (7) without introducing a separate sampling branch.

C.4 Inference and Evaluation Interface

The adapted backbones retain their native iterative sampler and VAE decoder at inference. The training endpoint estimate and auxiliary losses are not evaluated during sampling. For the controlled variants of a given backbone, inference settings, evaluation split, and valid region are held fixed. The training guidance scale in Table C.1 refers to optimization and does not specify the inference guidance setting.

Generated edits are converted to binary masks by Section 3.3. The fixed threshold, reference-grid evaluation, and absence of morphological refinement are specified in Appendix D.1. Dice, IoU, clDice, and CRecall follow the definitions and aggregation rules in Appendix D. This common evaluation interface is also used for the paired DeepGen comparison.

C.5 Controlled Comparisons

Output representation and training recipe. All three variants in Figure 4b follow the same training data schedule: 2 epochs followed by 15 epochs on the full training partition, then 5 epochs on the vessel training subset. Each epoch traverses the training subset specified for that phase. The direct-mask baseline in Figure 4b uses the same Qwen backbone, training data, split, initialization, and optimization configuration, but replaces each target edit with its corresponding binary mask. Its diffusion target is therefore a mask latent rather than a source-preserving edit latent. The non-staged editing baseline uses editing targets and the same data schedule, but does not use the staged optimization strategy. Its comparison with UniMedOpti evaluates the overall training recipe, while the residual controls below isolate the additional adaptation and supervision components.

Residual adaptation and objective components. Table C.3 defines the variants in Figures 5a,b and 6. The Stage-II checkpoint is the reference before further adaptation. The residual variants share this frozen checkpoint, residual initialization and capacity, vessel training set, and five-epoch schedule. Their inference and evaluation settings are also shared. Flow-only training controls for additional optimization and residual capacity. The region-supervised variant adds BCE and soft Dice, and full GRADE additionally includes cbDice.

Table C.3: **Objectives in the residual controls.** Checkmarks denote active training terms. Stage II is evaluated without further adaptation.

Variant	Additional epochs	Flow Matching	BCE	Dice	cbDice
Stage II	0	—	—	—	—
Flow-only residuals	5	✓	—	—	—
Residuals + region losses	5	✓	✓	✓	—
Full GRADE	5	✓	✓	✓	✓

Transfer to DeepGen. DeepGen 1.0 is adapted on OptiMedEdit-37K using the same staged training strategy, followed by GRADE in Stage III. Figure 5c compares the adapted DeepGen backbone with and without the residual refinement on FIVES and DRIVE under matched inference and evaluation settings. The comparison tests whether the edit-space adaptation applies to a second generative backbone. Published specialist references in Table 1 are taken from their respective studies, with selection and protocol details in Appendix D.5.

D Evaluation Metrics and Comparison Protocols

This appendix specifies how predictions are scored and how the published references in Table 1 are interpreted. It complements the experimental configuration in Appendix C with foreground and centerline metrics, aggregation rules, and dataset-specific reference provenance. The controlled comparisons in Figures 4–6 use our evaluation pipeline, while external reference scores retain the protocols of their cited studies.

D.1 Evaluation Inputs and Mask Recovery

Each evaluation record pairs a source image and target instruction with a binary reference mask $\mathbf{m}^{(n)} \in \{0,1\}^{H_n \times W_n}$. Multiple targets in one source image define separate segmentation tasks, as described in Appendix A.2. The instruction identifies the requested structure and editing operation without providing ground-truth locations at inference.

For an editing model, the generated image is converted to a binary prediction $\widehat{\mathbf{m}}^{(n)}$ using Eqs. (2)–(3). We use $\varepsilon = 10^{-6}$ and the fixed threshold $\tau = 0.5$ for all editing models and controlled editing variants. This threshold acts on the recovered edit strength $\widehat{a}_i$, which estimates ρm_i, rather than on the soft training probability $\widehat{q}_i$. Their relationship is derived in Appendix B.4. The recovered prediction is evaluated on the reference mask's spatial grid without additional trainable processing or morphological refinement. The direct-mask baseline supplies a binary segmentation output to the same region metrics.

Let Ω_n be the valid evaluation region, equal to the full reference image unless the adopted protocol specifies a field of view. All region counts and centerline overlaps below are restricted to this region. Changing the model's input resolution does not by itself define the scoring resolution: the reference grid and target support determine the evaluated task.

D.2 Foreground Region Metrics

Define the predicted and reference foreground sets as $P_n = \{i \in \Omega_n : \widehat{m}_i^{(n)} = 1\}$ and $G_n = \{i \in \Omega_n : m_i^{(n)} = 1\}$. For a nonempty foreground union, the per-record scores are

$$\mathrm{Dice}_n = \frac{2|P_n \cap G_n|}{|P_n| + |G_n|}, \qquad \mathrm{IoU}_n = \frac{|P_n \cap G_n|}{|P_n \cup G_n|}. \tag{A11}$$

These metrics score the requested foreground, without averaging it with a separate background-class score. For cell and gland targets, the support contains the pixels assigned to that target. No instance matching or object-detection criterion enters these region scores. Foreground pixel F1 is equivalent to Dice under the same counting and aggregation rules, whereas instance-detection F1 is a different quantity.

For a single nondegenerate mask pair, $\mathrm{Dice}_n = 2\mathrm{IoU}_n/(1+\mathrm{IoU}_n)$. The transformation is nonlinear, so it cannot generally convert a dataset's mean Dice into its mean IoU. Accordingly, an unreported reference IoU remains a dash in Table 1.

D.3 Centerline-Sensitive Metrics

The vessel analysis in Figure 6 supplements region overlap with clDice (Shit et al., 2021) and centerline recall (CRecall). Let $\mathcal{S}$ denote binary, topology-preserving 2-D thinning within the valid region, and write $S_{P,n} = \mathcal{S}(P_n)$ and $S_{G,n} = \mathcal{S}(G_n)$ for the predicted and reference centerline sets. For nonempty centerlines, topology precision and sensitivity are

$$T_{\mathrm{prec},n} = \frac{|S_{P,n} \cap G_n|}{|S_{P,n}|}, \qquad T_{\mathrm{sens},n} = \frac{|S_{G,n} \cap P_n|}{|S_{G,n}|}. \tag{A12}$$

The first measures how much of the predicted centerline lies in the reference foreground. The second measures how much of the reference centerline is covered by the prediction. We report their harmonic mean and the sensitivity term itself:

$$\mathrm{clDice}_n = \frac{2T_{\mathrm{prec},n}T_{\mathrm{sens},n}}{T_{\mathrm{prec},n} + T_{\mathrm{sens},n}}, \qquad \mathrm{CRecall}_n = T_{\mathrm{sens},n}. \tag{A13}$$

When both topology terms are zero for nonempty centerlines, clDice is zero. The intersections compare each centerline with the opposite foreground support, not with the opposite centerline.

CRecall exposes omitted centerline segments but does not penalize foreground expansion on its own. clDice balances centerline coverage and precision, while Dice and IoU constrain region extent. These complementary scores support a geometric interpretation of the ablation, without establishing that every branch is connected or anatomically correct. Evaluation uses hard binary masks. The cbDice training loss (Shi et al., 2024), including its detached geometric weights and differentiable prediction path, is specified in Appendix B.5 and is not an evaluation metric reported in Figure 6.

D.4 Aggregation and Reporting

For our model and controlled variants, each metric is computed separately for each evaluated image–target record. For an evaluation collection $\mathcal{E}$ of $N_{\mathcal{E}}$ records,

$$\overline{M}_{\mathcal{E}} = \frac{1}{N_{\mathcal{E}}} \sum_{n \in \mathcal{E}} M_n, \qquad M \in \{\mathrm{Dice}, \mathrm{IoU}, \mathrm{clDice}, \mathrm{CRecall}\}. \tag{A14}$$

Each record has equal weight irrespective of image resolution or foreground size. Pixels are not pooled across images. Dataset-specific entries use the corresponding test subset in Table A.1. Full-test-set entries aggregate all 7,332 test records directly, so datasets contribute according to their numbers of evaluated records, rather than through an equal-weight average of dataset means. clDice is computed before this averaging, rather than from separately averaged topology precision and sensitivity.

Values are rounded after aggregation, to four decimal places on the unit scale or equivalently two decimal places on the 0–100 display scale. Reported gains are absolute differences on the latter scale. Controlled variants share the evaluation collection, valid region, scoring grid, binarization, skeletonization, and averaging rule, with editing outputs using the same recovery settings. The qualitative panel in Figure 7 illustrates selected best-case examples across modalities and does not replace these aggregate evaluations.

D.5 Reference Selection and Comparison Protocols

Table 1 compares a shared model with published references for individual segmentation tasks. Table D.1 records the reporting source and configuration for every reference entry. These scores are transcribed from the cited studies and have not been recomputed through our evaluation pipeline.

Reference configurations. For the six MedSegBench entries, we select a U-Net configuration with the highest reported foreground F1 in that benchmark and retain its paired IoU (Kuş & Aydin, 2024). Scores from different encoders are not combined into a single result. The encoder names in Table D.1 identify these U-Net configurations, rather than separate segmentation methods. Kvasir-SEG instead uses the later Citrus-V 8B result reported on MedSegBench (Wang et al., 2025). This selection provides traceable reference points within the cited experiments, not an exhaustive ranking of all subsequent methods.

Partitions and scoring conventions. A common dataset name or split ratio is insufficient to establish equivalent evaluation. The held-out records, target support, valid region, scoring resolution, aggregation, and inference inputs must also correspond. Using a benchmark partition does not imply that the source lacks an official split: the benchmark may preserve the source test set while standardizing inputs and annotations. MedSegBench reports baselines at 512×512 and averages three runs. Its released binary test script uses batch size one and computes foreground F1 and IoU per image before averaging.[2] This identifies the reference's convention without assuming that every later study citing MedSegBench uses the same pipeline.

For ISIC 2016, we use the paired Dice and IoU from the main comparison in Table 3 of Ren et al. (2026). CFG-MambaNet describes the source training/test partition and resizes ISIC images and labels to 224×224. We retain these values as author-reported references, without combining metric maxima from different ablation configurations. For Kvasir-SEG, Citrus-V reports a MedSegBench result and releases a per-image Dice evaluator.

[2] MedSegBench binary evaluation script: `examples/test_pytorch.py`.

Table D.1: **Provenance of the published references in Table 1.** The last column identifies the evaluation collection reported by the reference study and the source of the adopted scores. Protocol qualifications are discussed in the text.

Dataset / target	Reference configuration	Reported collection and source
Dataset-specific studies		
FIVES	SuperVessel	Official held-out test, Table VII of Qiu et al. (2024)
REFUGE disc / cup	GCN	Official test, separate disc/cup scores, Table 1 of Tian et al. (2020)
ISIC 2016	CFG-MambaNet	Source train/test partition, Table 3 of Ren et al. (2026)
ISIC 2018	TESL-Net	Task-1 test collection, Table IV of Iqbal et al. (2024)
GlaS	ODS-SAM	Source challenge test collection, Table 3 of Huang et al. (2025)
Studies using MedSegBench		
DeepBacs	U-Net / ResNet-50	Released benchmark test, Table 5 of Kuş & Aydin (2024)
BriFiSeg	U-Net / DenseNet-121	Same benchmark source
NuSeT	U-Net / DenseNet-121	Same benchmark source
YeaZ	U-Net / DenseNet-121	Same benchmark source
CystoidFluid	U-Net / ResNet-50	Same benchmark source
RoboTool	U-Net / MiT-B0	Same benchmark source, encoder denoted MVT
Kvasir-SEG	Citrus-V 8B	MedSegBench result, Table 10 of Wang et al. (2025). IoU unreported.

Task and inference information. The semi-automatic deep level set method uses four user-provided boundary extreme points (Zheng et al., 2021), introducing spatial information unavailable to our image-and-instruction interface. For BriFiSeg, cell-line-specific or multiple-focal-plane results (Gendarme et al., 2022) require alignment with the single-channel aggregate benchmark task before substitution.

Interpretation of comparisons. Higher published scores under other settings do not by themselves overturn a comparison within a specified task. For example, the five-fold cross-validation results of Deconver (Ashtari et al., 2025) are not interchangeable with the fixed FIVES holdout. Conversely, incomplete protocol information does not establish that an unlisted method is incompatible or weaker. Table 1 provides published performance context across modalities. The component-level conclusions about editing targets, residual adaptation, and geometric supervision rely on the controlled experiments in Sections 5.2–5.4, whose training configurations are given in Appendix C.5.

E Additional Qualitative Comparisons

Figures A7–A11 extend Figure 7 with selected qualitative examples across datasets and targets. Each row pairs a source image with an annotation-derived reference edit and outputs from five pretrained editing models and our model. The displayed instructions are paraphrased for presentation, with modality and target terminology made explicit from the corresponding records in accordance with Appendix A.3. The panels retain complete image frames and the outputs generated with the original recorded prompts. These selected cases illustrate edit localization and background preservation. Aggregate performance is reported in Section 5 under the evaluation rules in Appendix D.

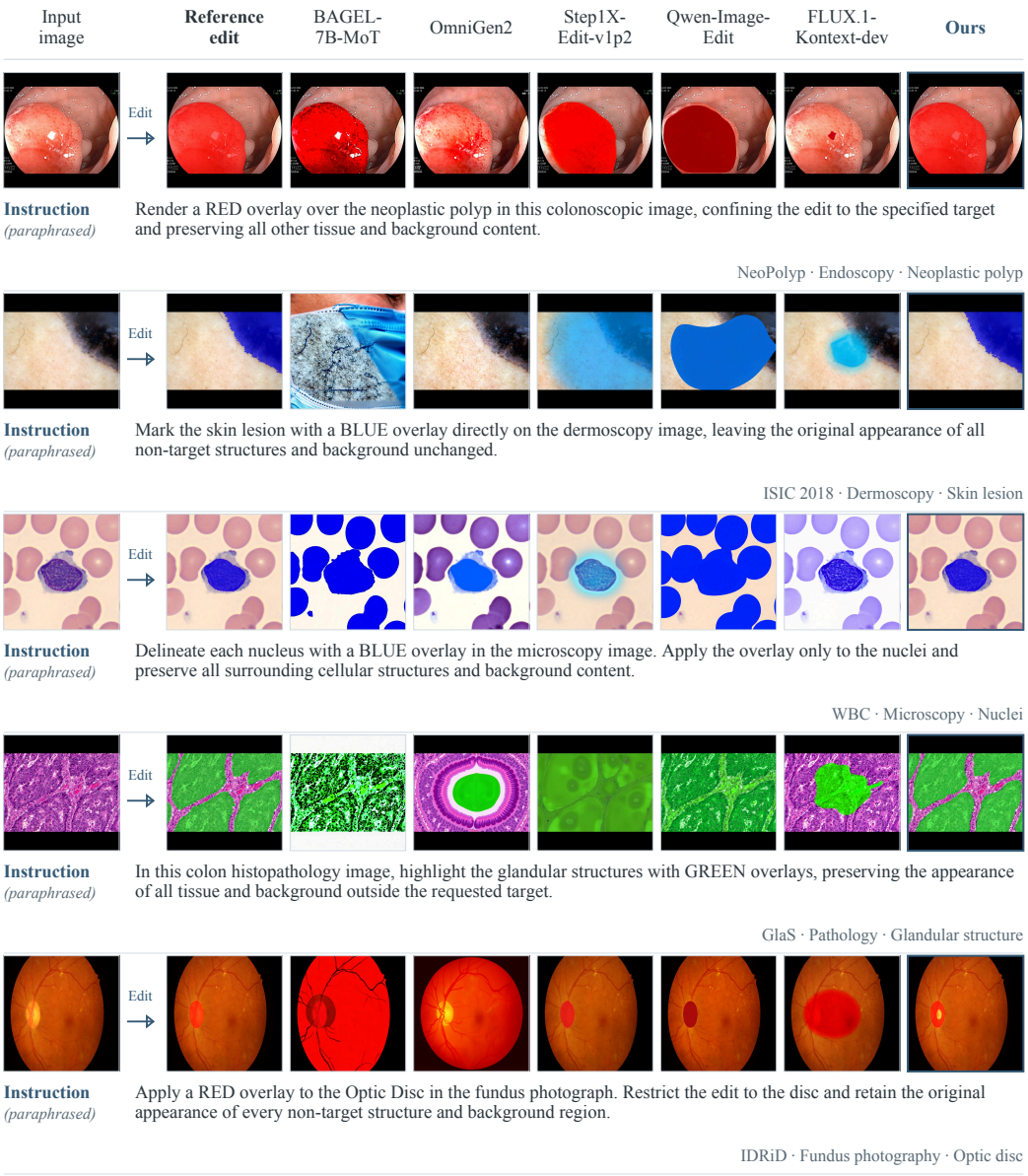

Figure A7: **Additional qualitative comparisons across optical imaging targets.** Selected examples from NeoPolyp, ISIC 2018, WBC, GlaS, and IDRiD.

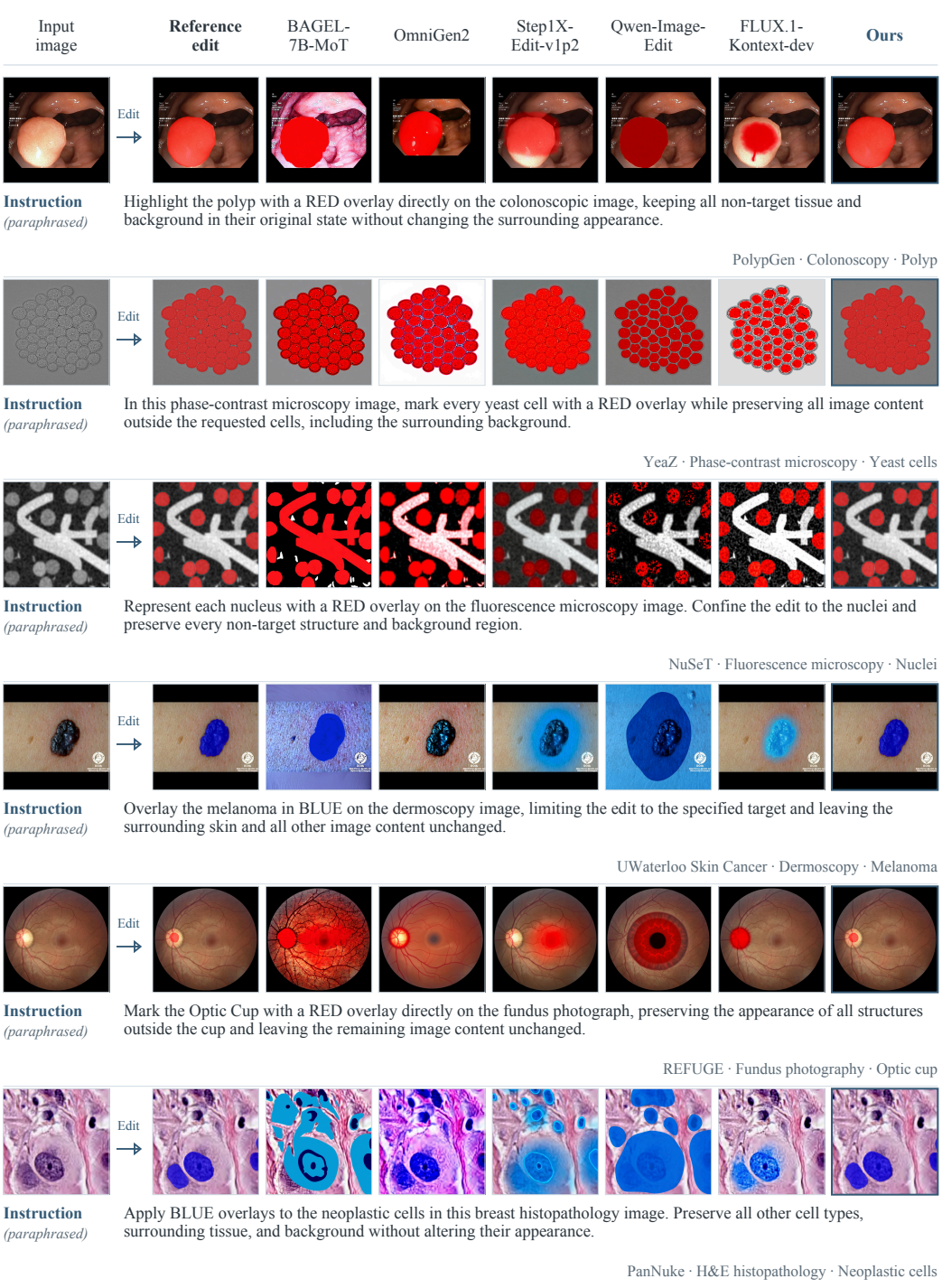

Figure A8: **Qualitative comparisons for cell, lesion, and anatomical targets.** Examples from PolypGen, YeaZ, NuSeT, UWaterloo Skin Cancer, REFUGE, and PanNuke, using the same column order as Figure A7.

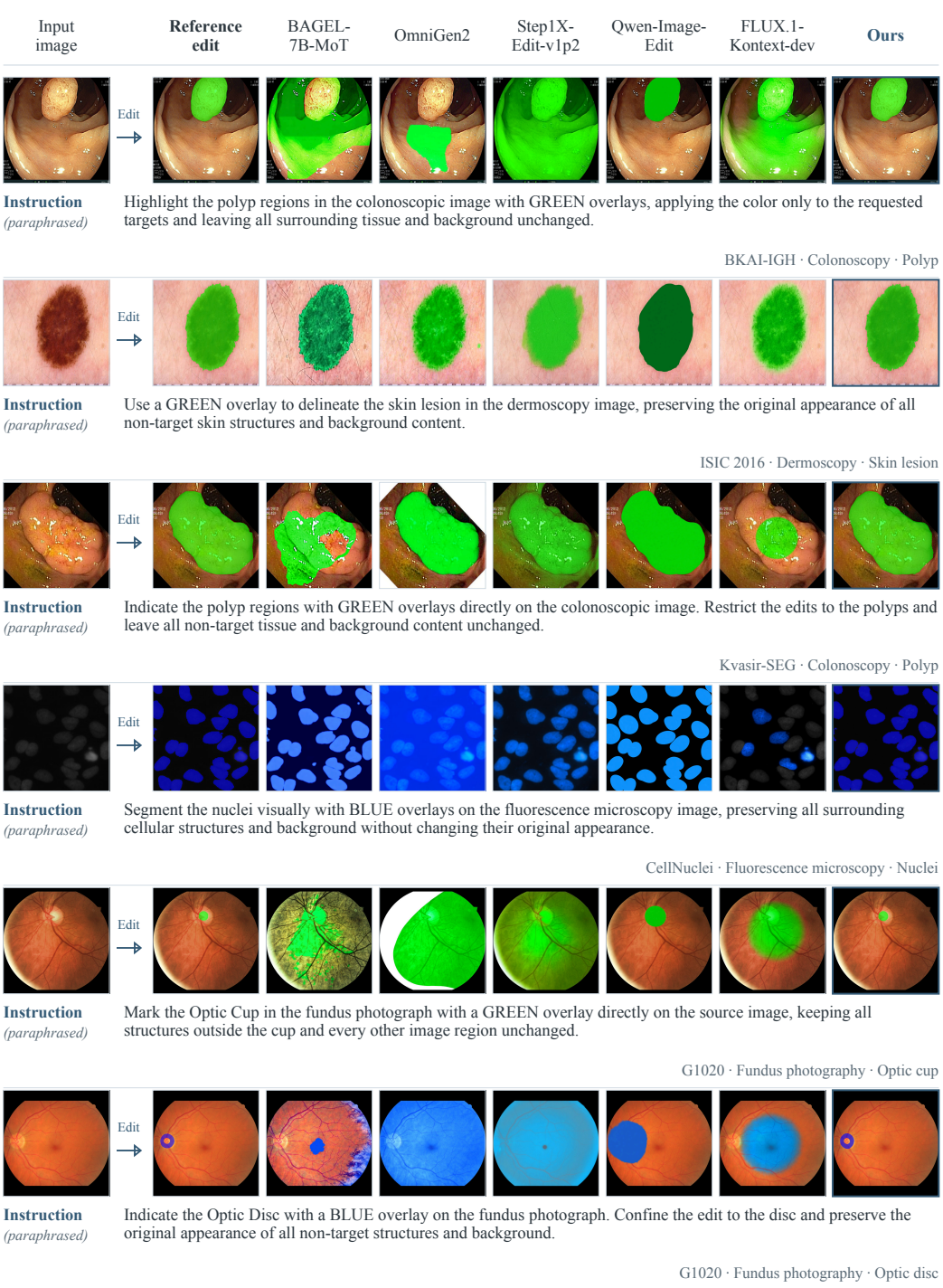

Figure A9: **Further comparisons across datasets and target structures.** Examples from BKAI-IGH, ISIC 2016, Kvasir-SEG, CellNuclei, and G1020. The two G1020 rows depict separate image–target records.

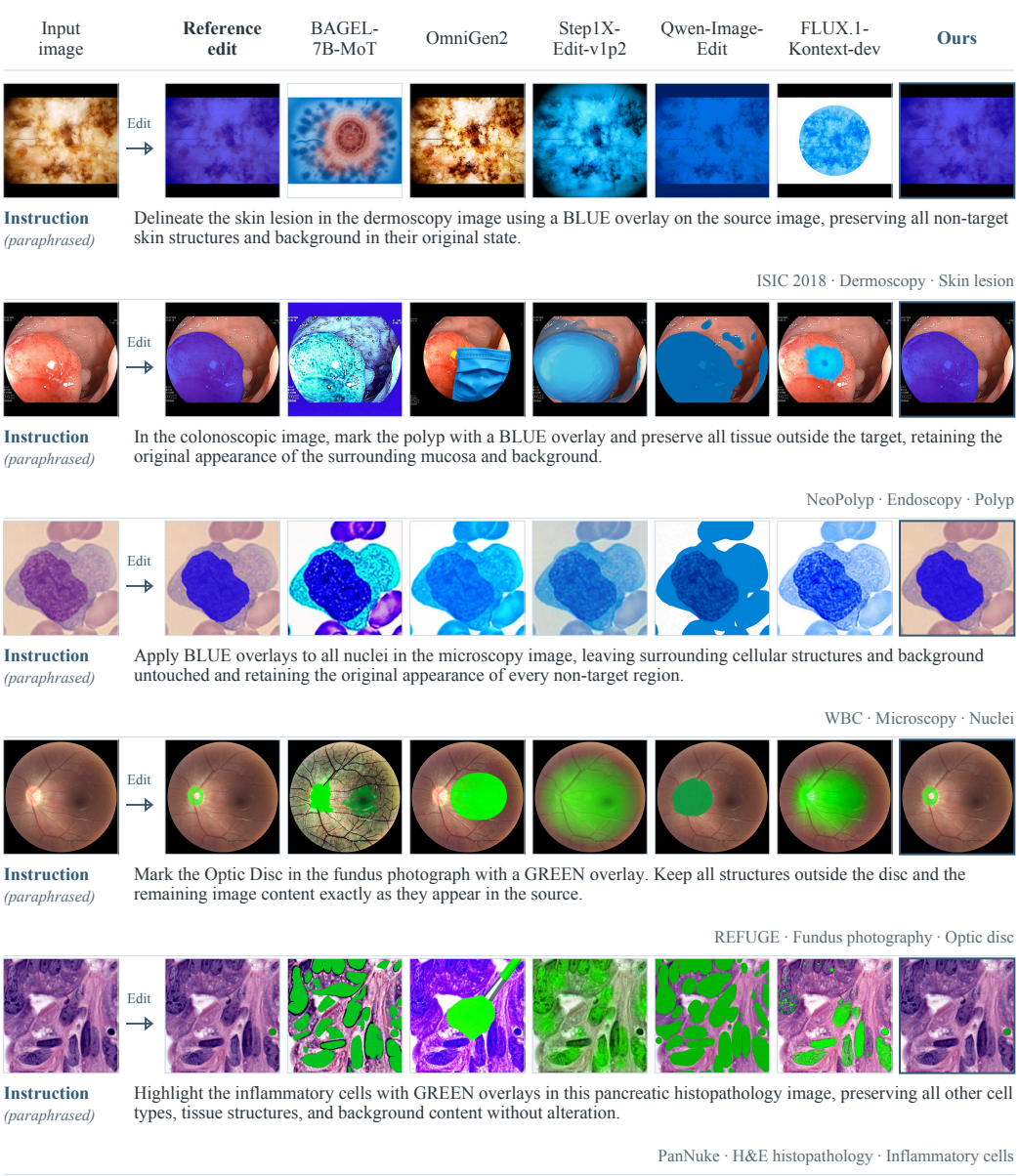

Figure A10: **Additional cases from previously represented datasets.** Different source images from ISIC 2018, NeoPolyp, WBC, REFUGE, and PanNuke complement the preceding comparisons.

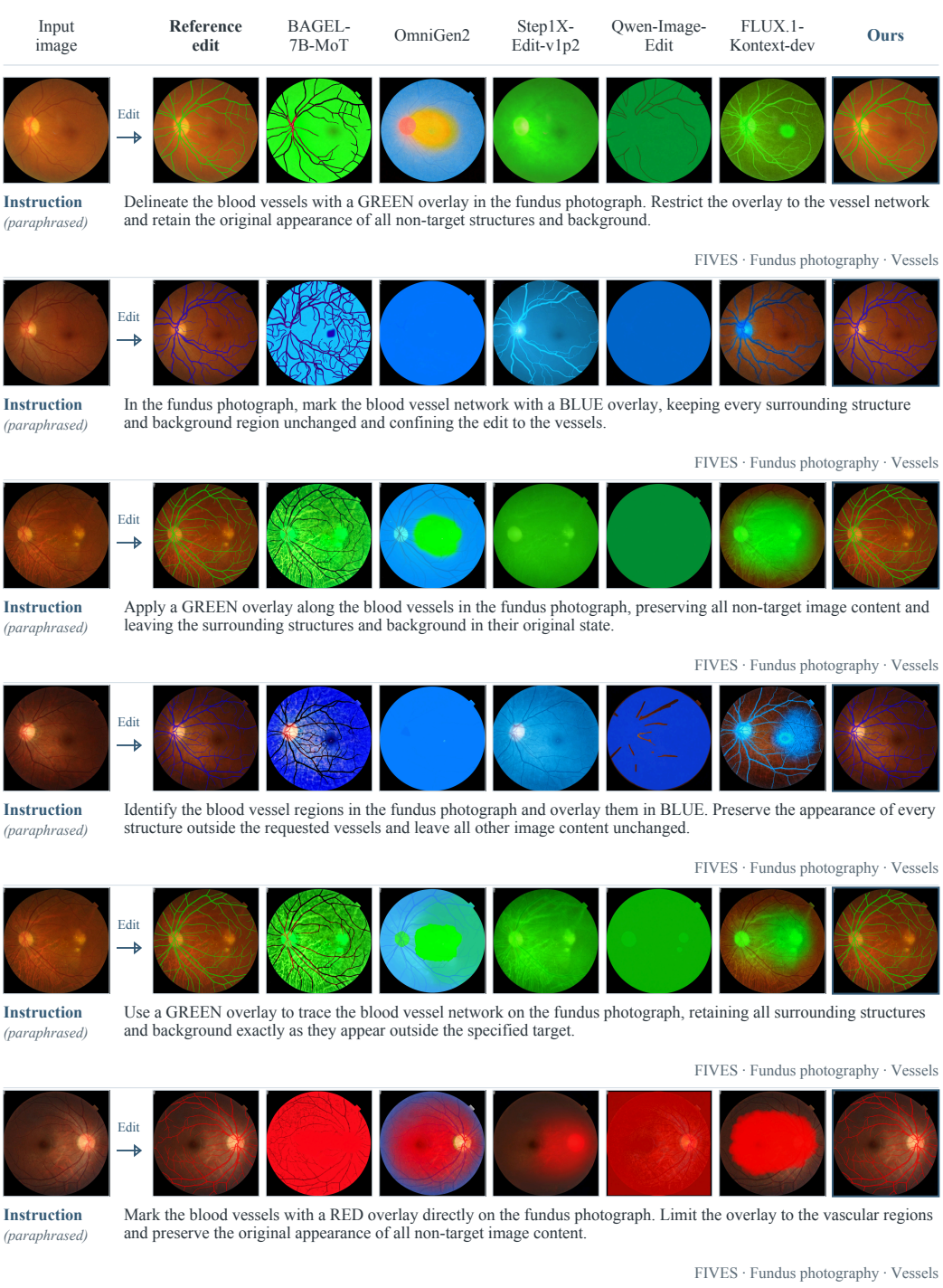

Figure A11: **Vessel-structure editing on FIVES.** Six selected cases show instruction-guided vessel overlays. These are generated image edits, with reference edits in the second image column.