MEDVISION-X: UNIFYING MEDICAL IMAGE ANALYSIS AND SYNTHESIS THROUGH GENERATIVE MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical image analysis and synthesis are served by separate specialist models: segmentation networks predict label maps, restoration networks recover intensities, and translation or editing networks synthesize new images, each with its own architecture, objective, and output head. We ask how far a single generative model with one image output pathway can go across these heterogeneous tasks. We present MedVision-X, which casts segmentation, super-resolution, cross-modal translation, virtual immunohistochemical staining, CT denoising, and lesion editing as instruction-conditioned image generation. Masks are rendered as color-coded images with an optional text legend and continuous targets are generated directly, so every task is trained with the same flow-matching objective and decoded by the same image decoder, without task-specific heads or losses. Starting from the SenseNova-Vision unified multimodal model, MedVision-X is adapted in two stages: broad adaptation on 1.04M medical records, followed by balanced joint training on the training folds of 16 benchmarks. Across 14 test benchmarks and 62 scored baseline entries, MedVision-X obtains the best primary endpoint on 8 and ranks second on 2 more. The gains are largest for physically meaningful targets: one checkpoint reduces body HU MAE from 99.0 to 39.1 on low-dose CT and from 83.1 to 55.2 on MRI-to-CT synthesis, and improves PSNR by 4.0 dB on LoDoPaB-CT and 3.2 dB on IXI. Dataset-specific segmentation checkpoints lead on BUSI and Kvasir-SEG, while ISIC, DRIVE, and pixel-fidelity super-resolution remain behind the strongest specialists. A mask round-trip diagnostic shows that the image representation preserves thin structures, locating the remaining gap in conditional spatial precision rather than in the unified output format. The code and model weights are publicly available at `https://anonymous.4open.science/r/MedVision-X/README.md`.

1 INTRODUCTION

Medical image computing comprises tasks whose outputs look very different. Segmentation delineates anatomy and lesions as discrete label maps. Restoration and super-resolution recover continuous intensities from degraded acquisitions. Cross-modal translation and virtual staining predict the same tissue in another modality or stain, and lesion editing realizes a requested local change while leaving the rest of the image intact. Each of these tasks is usually solved by its own specialist: self-configuring U-Nets for segmentation (Isensee et al., 2021), residual or transformer networks for restoration (Chen et al., 2017; Conde et al., 2022), and conditional adversarial or instruction-conditioned generators for translation, staining, and editing (Isola et al., 2017; Zhu et al., 2017; Li et al., 2023a; Brooks et al., 2023). Each task thus needs its own architecture, loss, output head, and training pipeline, and supervision for one task cannot inform another, although all of them depend on the same anatomy, tissue morphology, and pathology.

General computer vision has recently shown a way out of this fragmentation. Vision Banana instruction-tunes an image generator to output decodable RGB encodings of segmentation, depth, and surface normals (Gabeur et al., 2026). SenseNova-Vision expresses detection, segmentation, dense geometry, and multi-view prediction through the native text and image outputs of a unified multimodal model, and matches leading specialists without task-specific heads (Han et al., 2026). In both systems the output space of the generator becomes the task interface. This raises a question

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

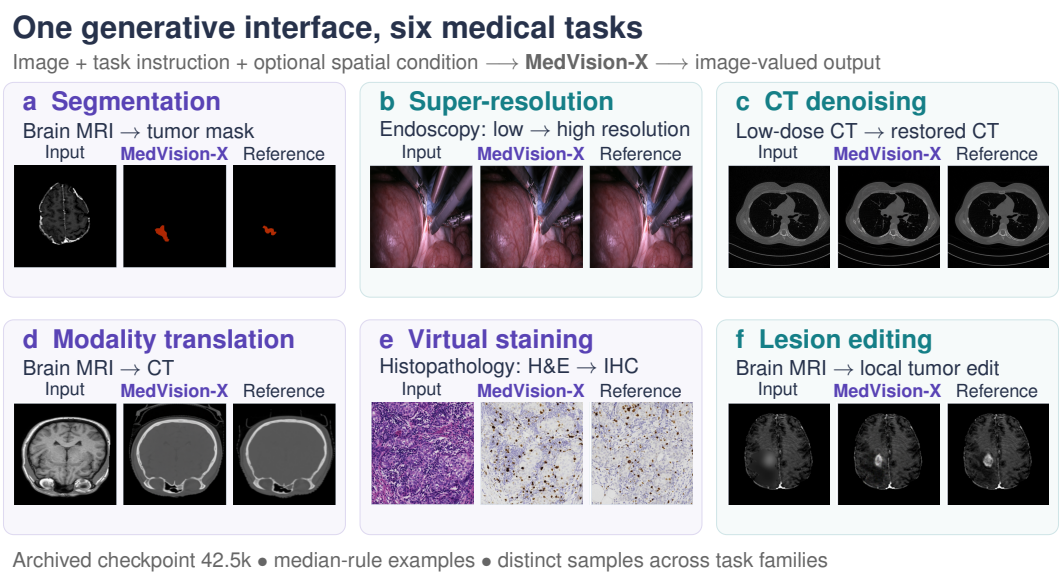

Figure 1: **Six medical tasks through one generative interface.** For each task family we show the input, the MedVision-X output, and the reference, spanning MRI, CT, endoscopy, and histopathology. Each example is the test sample closest to the task's median per-sample metric at an earlier development checkpoint (42.5k); image tiles are cropped without retouching. Segmentation colors: tumor (white), edema (orange), necrosis (purple).

for medicine: *can one generative model, using a single image output pathway, reach specialist-level accuracy across heterogeneous medical analysis and synthesis tasks?*

Medical imaging makes this question harder in three ways. First, outputs mix discrete and continuous semantics: a mask must decode to exact labels, whereas a synthetic CT must be accurate in Hounsfield units. Second, clinically relevant structures are often small, thin, or low in contrast, such as retinal vessels, polyp boundaries, and lesions, so small spatial errors change the result. Third, generation must be controlled: translation and staining must change appearance while preserving spatial correspondence, and editing must change only the requested region. Existing medical generalists address parts of this problem. MVG casts several tasks as in-context image-to-image prediction (Ren et al., 2024); MedVersa routes requests to dedicated visual modules through a language-model orchestrator (Zhou et al., 2024); UniMedVL trains one model for medical understanding and generation but reports a remaining gap to a segmentation specialist (Ning et al., 2025). What remains untested is whether a single native generative pathway can compete with task-trained specialists on dense outputs that span segmentation, restoration, translation, staining, and editing, under the benchmark protocols those specialists use.

We introduce MedVision-X, a generative generalist for medical imaging built on SenseNova-Vision. Given a source image, a natural-language instruction, and an optional spatial condition, MedVision-X produces the target image through the same latent flow-matching pathway and frozen image decoder for every task (Figure 1). Two design choices make this work. (i) *A decodable task representation.* Segmentation targets are rendered as color-coded masks whose classes are bound to names by a short text legend; restoration, translation, staining, and editing targets are generated directly as images and mapped back to physical units where the dataset defines them. No task-specific head and no Dice, adversarial, or perceptual loss is added. (ii) *A staged medical adaptation recipe.* Broad adaptation on 1,036,436 records from 18 weighted sampling streams exposes the model to many modalities and image transformations. Balanced joint training on the training folds of 16 benchmarks, with equal weight for each of the six task families, then specializes the shared weights without letting the largest task dominate.

We evaluate MedVision-X on 14 benchmarks spanning MRI, CT, ultrasound, dermoscopy, endoscopy, fundus photography, and histopathology, against 62 scored baseline entries that include target-trained specialists, released checkpoints, and reimplementations (Table 2). MedVision-X obtains the best

primary endpoint on 8 of the 14 benchmarks and ranks second on two more. The evaluation yields three findings. *First, a shared generator is most accurate where the target is a physical quantity.* With one checkpoint, MedVision-X lowers body HU MAE by 61% relative to the strongest low-dose CT baseline (39.11 vs. 99) and by 34% on SynthRAD2023 MRI-to-CT synthesis (55.2 vs. 83.11), and leads all endpoints on LoDoPaB-CT and SynthRAD2023. *Second, the remaining gap is spatial precision, not the output format.* MedVision-X leads Dice on BUSI and Kvasir-SEG but trails on the thin structures of DRIVE and ISIC borders, while a VAE round trip of ground-truth masks retains 0.963 Dice on DRIVE, so the image representation itself is not the bottleneck. *Third, single similarity scores hide task-specific tradeoffs.* Lower staining FID does not imply better marker agreement, the best perceptual super-resolution is not the best in PSNR, and edit realization must be read together with preservation; we report each pair explicitly.

In summary, our contributions are:

- **A unified generative formulation for medical imaging.** Six task families are expressed as instruction-conditioned image generation with decodable outputs, trained with one objective and decoded by one image decoder, without task-specific heads or losses. The backbone is inherited from SenseNova-Vision; the contribution is the medical task representation.
- **A staged medical adaptation recipe.** Broad weighted adaptation is followed by balanced joint training restricted to benchmark training folds, so that heterogeneous and highly imbalanced medical data train one shared model.
- **A provenance-aware benchmark study** against 62 scored baseline entries on 14 benchmarks, which identifies where a shared medical generator already matches or exceeds specialists and where it does not, with every comparator labeled by training provenance and every claim scoped to the checkpoints that produced it.

2 Related work

Vision tasks as native generation. Unified multimodal models such as BAGEL combine understanding and image generation in one pretrained transformer (Deng et al., 2025). Recent work shows that such generators can serve as general vision learners. Vision Banana instruction-tunes an image generator to emit decodable RGB encodings of segmentation, depth, and surface normals, and retains its generation ability (Gabeur et al., 2026). SenseNova-Vision, which we build on, expresses structured perception, dense geometry, segmentation, and multi-view prediction through native text, image, and mixed outputs without task-specific heads (Han et al., 2026). These studies target natural images, where most targets are categorical or geometric. Medical tasks additionally require continuous targets with physical meaning, such as Hounsfield units, and controlled changes such as local lesion edits. We test whether the same generative interface transfers to this setting and at what precision.

Medical generalist models. Several works unify medical tasks, but at different levels. MVG represents segmentation, cross-modal synthesis, denoising, and inpainting as image-to-image prediction conditioned on example image–label pairs (Ren et al., 2024). MedVersa unifies at the system level: a language-model orchestrator calls dedicated visual modules for detection and segmentation (Zhou et al., 2024). UniMedVL is the closest predecessor. Its progressive curriculum trains one set of parameters for medical understanding and generation, including segmentation, super-resolution, virtual staining, and cross-modal synthesis, and it reports a remaining gap to a segmentation specialist (Ning et al., 2025). Compared with these models, MedVision-X uses one native image output pathway for all dense tasks, adds CT restoration and lesion editing, and measures accuracy directly against task-trained specialists on their standard benchmarks.

Specialists for dense medical prediction. Task-specific systems define the accuracy that a generalist must reach. nnU-Net configures segmentation pipelines automatically (Isensee et al., 2021), and MedSAM and MedSAM3 bring promptable segmentation to medical images (Ma et al., 2024; Liu et al., 2025). RED-CNN and Swin2SR address low-dose CT denoising and super-resolution with residual and transformer architectures (Chen et al., 2017; Conde et al., 2022). Pix2Pix and CycleGAN established paired and unpaired translation (Isola et al., 2017; Zhu et al., 2017), Adaptive Supervised PatchNCE targets H&E-to-IHC staining (Li et al., 2023a), and InstructPix2Pix learns instruction-conditioned editing (Brooks et al., 2023). Each specialist encodes its task in a dedicated

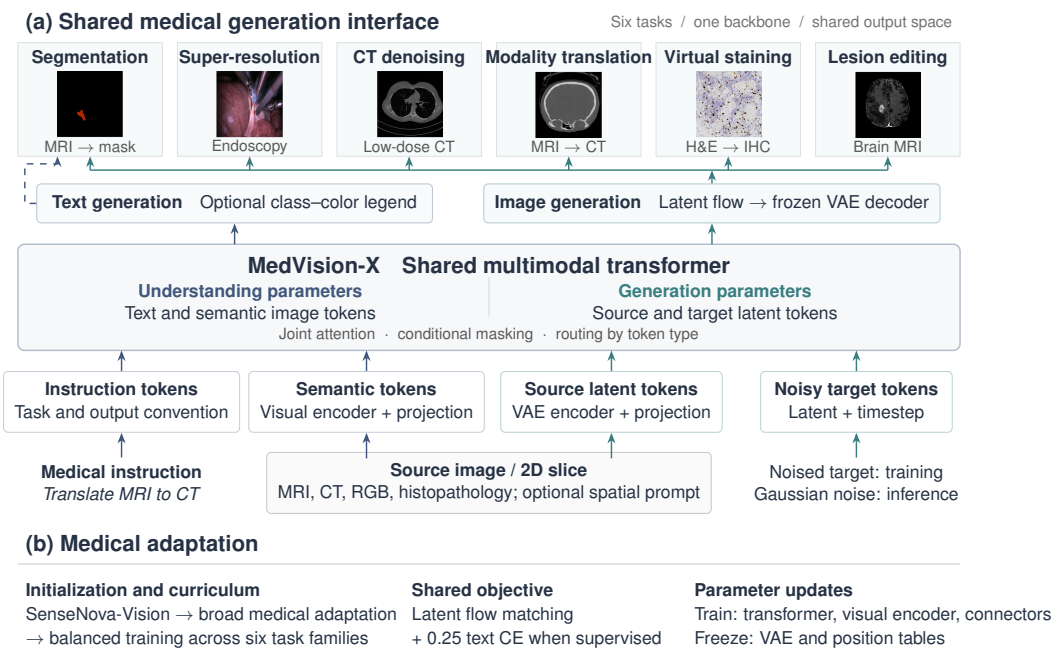

Figure 2: **MedVision-X architecture and medical adaptation. (a)** The instruction, semantic image features, source latents, and noisy target latents form one sequence in a shared transformer. Tokens are routed to understanding or generation parameters by token type, not by task, and every task decodes its output through the same latent flow and frozen VAE decoder; segmentation can additionally emit a class–color legend. **(b)** Adaptation updates the transformer, visual encoder, and connectors with a single objective. Thumbnails illustrate target types and come from an earlier development checkpoint. Layout adapted from Han et al. (2026).

architecture or loss; MedVision-X learns all of them from task-conditioned examples under one objective.

3 METHOD

MedVision-X poses every medical task as generating an image from a source image and an instruction, so that the task is defined by the data rather than by the architecture (Figure 2).

3.1 MEDICAL TASKS AS CONDITIONAL IMAGE GENERATION

A task instance consists of a source visual context x, an instruction p, and an image-valued target y. The context contains the source image and, when the task provides one, an additional spatial condition such as an edit mask. The instruction states the requested transformation, for example *segment the polyp*, *translate MRI to CT*, or *generate the Ki67 stain*. A single conditional generator serves all tasks:

$$\widehat{y} = D_\phi(G_\theta(E_\phi(x), V_\psi(x), p; \epsilon)), \qquad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where E_ϕ and D_ϕ are the encoder and decoder of a frozen image autoencoder, V_ψ extracts semantic visual features, and G_θ generates target latents. All parameters are shared across tasks. Datasets differ only in how source and target are paired and normalized, in the instruction, and in the rule used to read out the prediction.

3.2 DECODABLE TARGETS

Continuous targets. Restoration, super-resolution, translation, staining, and editing targets are images and are generated directly in normalized intensity space. Where a dataset defines physical

units, as for CT, the declared inverse intensity transform maps the prediction back to Hounsfield units for evaluation.

Segmentation targets. An integer annotation is rendered as an RGB mask with a black background, and foreground class k receives the kth color of a declared palette. A short text legend binds each class name to its color. During training the legend can be supervised as text before the mask, which yields a mixed text–image response; at inference it is either generated or supplied as conditioning. Multiclass predictions are decoded to the nearest palette color. Binary tasks use a white foreground, so the generated luminance acts as a score that is thresholded with a validation-selected rule. Segmentation is thus learned with the same image objective as every other task, while its decoded output remains a conventional label map. Because masks pass through a frozen autoencoder, we verify in Section 5.3 that this representation preserves thin structures.

3.3 Backbone

We initialize MedVision-X from SenseNova-Vision, which builds on the BAGEL Mixture-of-Transformers architecture (Han et al., 2026; Deng et al., 2025). The source image enters through two complementary representations: VAE latents carry spatial and appearance detail, and a SigLIP-style encoder provides semantic features (Tschannen et al., 2025). Instruction tokens, projected semantic tokens, source latent tokens, optional legend tokens, and noisy target latents form a single sequence. Text and semantic tokens use the understanding parameters of the transformer and latent tokens use its generation parameters, while joint attention lets both representations condition the output. Routing depends only on token type; no expert is selected per medical task. The attention mask isolates examples packed into one sequence; within an example, target tokens attend to the conditioning context and to each other, but conditioning tokens cannot read the noisy target. We update the transformer, visual encoder, and connectors (about 14.6B parameters, without LoRA) and keep the VAE and fixed position tables frozen. The architecture is inherited (Appendix A); our contribution lies in the task representation, adaptation recipe, and evaluation.

3.4 Training objective and inference

For a target latent $z_0 = E_\phi(y)$ and Gaussian noise ϵ, we train a velocity field on the interpolation path of the pretrained model (Lipman et al., 2023):

$$z_t = (1-t)z_0 + t\epsilon, \qquad \mathcal{L}_{\text{flow}} = \mathbb{E}\Big[\|v_\theta(z_t, x, p, t) - (\epsilon - z_0)\|_2^2\Big]. \tag{2}$$

The loss is averaged over supervised latent tokens. When the target contains text, mainly the segmentation legend, we add a token-normalized cross-entropy term with weight 0.25. This is the only objective for all six task families. Conditioning dropout of 0.05 for text and 0.10 for each image condition enables classifier-free guidance; the timestep distribution is given in Appendix B. At inference, the learned field is integrated with an Euler solver from Gaussian noise on a shifted time grid, and the result is decoded by the shared VAE decoder. Volumetric data are processed as independent 2D slices in both training and inference.

4 Medical data and staged adaptation

Training one generator for six medical task families raises two practical problems. The available data are highly imbalanced: a single segmentation aggregate contributes 861,344 of our 1,036,436 broad-adaptation records. In addition, benchmark comparisons require that test data never enter optimization. We therefore separate training into a broad adaptation stage, which teaches medical appearance and image transformations, and a joint stage restricted to benchmark training folds. Both stages use the objective in Equation 2 and the same architecture. Appendix D shows instruction–target examples for every task family.

Stage A: broad medical adaptation. The Stage-A corpus contains 1,036,436 training records in 18 sampling streams. Seven task aggregates (978,506 records) cover reconstruction, segmentation, super-resolution, modality translation, virtual staining, denoising, and counterfactual editing; reconstruction

Table 1: **Six task families in joint adaptation.** Each family receives one sixth of the nominal Stage-B sampling probability. All targets are images; segmentation targets additionally have a class–color interpretation.

Task family	Training datasets	Target representation
Segmentation	ISIC 2018, Kvasir-SEG, BUSI, DRIVE, BraTS-GLI	Binary or multiclass mask
Super-resolution	fastMRI SR, SurgiSR4K	Higher-resolution image
Cross-modal synthesis	SynthRAD2023, IXI	Paired target modality
Denoising / restoration	AAPM Mayo LDCT, LoDoPaB-CT, LDCT projection	Reference-quality CT image
Virtual staining	MIST, BCI	Paired IHC image
Counterfactual editing	BraTS tumor editing, ATLAS v2	Constructed edited target

serves only as an auxiliary mapping. Eleven named sources add 57,930 records spanning contrast-enhanced CT, breast, brain, and cardiac MRI, microscopy, low-dose CT, and PET (Contrast-X, 2026; Garrucho et al., 2025; Fridman et al., 2025; Newitt & Hylton, 2016; Baid et al., 2021; Kazerooni et al., 2025; Li et al., 2023b; Zhang et al., 2019; Zhou et al., 2020; McCollough et al., 2017; Xue et al., 2025). Streams are drawn with explicit weights rather than in proportion to their size: in the expanded mixture, the original aggregates receive 71.43% of the draw probability and the added sources 28.57%, although the latter hold only 5.59% of the records. This prevents the segmentation aggregate from dominating exposure. Counts are 2D records, not patients (Appendix C).

Stage B: balanced joint adaptation. Stage B starts from the Stage-A EMA weights and trains on 300,392 2D records derived only from the designated training folds of 16 benchmarks (Table 1). Validation folds are used for checkpoint and inference selection, and test folds are excluded. Each of the six task families receives one sixth of the sampling probability, and dataset weights set exposure within a family; very small datasets are down-weighted, for example DRIVE (16 training images) has weight 0.08 against 0.48 for each other segmentation dataset. Balancing at the family level keeps rare tasks such as lesion editing in every stage of training.

Preprocessing and optimization. Volumes are aligned as the source protocol requires, intensity-normalized (fixed HU windows for CT, foreground-based for MRI), and exported as 8-bit 2D slices; editing targets are constructed from annotations, not observed counterfactuals. Variable-resolution examples are packed into long sequences, and examples that overflow a pack are carried to the next one rather than dropped, which would distort the task mixture. Stage A runs 25,000 optimizer updates at learning rate 2.5×10^{-5} and Stage B runs 2,000 updates at 5×10^{-6}, both on 16 GPUs with BF16, fused AdamW, and EMA decay 0.995. Appendix B lists all settings, and Figure 3 visualizes both mixtures.

5 Experiments

We evaluate MedVision-X against task specialists to answer three questions. (Q1) Can one generative model reach specialist-level accuracy across six heterogeneous medical task families? (Q2) Where does it fall short? (Q3) Beyond global image similarity, does the generated output satisfy the structural, semantic, and preservation requirements of each task?

5.1 Experimental setup

Benchmarks and metrics. We report 14 test benchmarks: four for segmentation (ISIC 2018, Kvasir-SEG, BUSI, DRIVE) and two each for super-resolution (fastMRI, SurgiSR4K), CT denoising (LoDoPaB-CT, LDCT projection), cross-modal synthesis (SynthRAD2023 MRI-to-CT, IXI T1-to-PD), virtual staining (MIST, BCI), and lesion editing (BraTS, ATLAS v2). Metrics follow each task: Dice, IoU, and HD95 for segmentation; body-masked HU MAE, PSNR, and SSIM for CT; PSNR, SSIM, NMSE, and LPIPS for super-resolution; FID, KID, LPIPS, SSIM, and auxiliary-classifier

Table 2: **Results on 14 benchmarks across six task families.** Each row compares MedVision-X with the strongest scored baseline for the displayed primary endpoint. Bold and underline mark the best and second-best values among all methods in the full per-dataset table (Appendix H). Baseline provenance: [T] trained on the benchmark training split, [C] released checkpoint. Run: B, shared checkpoint (Stage-B microstep 3999); S1500/S1000, dataset-specific segmentation checkpoints; E4/E3, four- and three-checkpoint ensembles; M, lesion-mask compositing.

Dataset	Endpoint	Comparator	Baseline	MedVision-X	Run
Segmentation					
ISIC 2018	Dice ↑	MedSAM3 [C]	**0.9282**	0.8918	S1500
Kvasir-SEG	Dice ↑	MedCLIPSeg [C]	<u>0.9034</u>	**0.9046**	E4
BUSI	Dice ↑	nnU-Net v2 [T]	<u>0.8097</u>	**0.8271**	E3
DRIVE	Dice ↑	nnWNet [T]	**0.8252**	0.7911	S1000
Super-resolution					
fastMRI SR	PSNR ↑	GaussM2ASR [T]	**36.76**	35.86	B
SurgiSR4K	RGB PSNR ↑	HAT [C]	**31.83**	30.92	B
Denoising / restoration					
LoDoPaB-CT	PSNR ↑	Restormer [C]	<u>33.97</u>	**37.99**	B
LDCT projection	HU MAE ↓	WIA-LD2ND [T]	<u>99</u>	**39.11**	B
Cross-modal synthesis					
SynthRAD2023	HU MAE ↓	RegConDiS [T]	<u>83.11</u>	**55.2**	B
IXI	PSNR ↑	DRDD [T]	<u>22.43</u>	**25.58**	B
Virtual staining					
MIST	FID ↓	VSMT [C]	<u>44.88</u>	**31.66**	B
BCI	FID ↓	VSMT [C]	<u>47.27</u>	**39.12**	B
Lesion editing					
BraTS editing	Edit Dice ↑	InstructPix2Pix [C]	**0.4849**	<u>0.4466</u>	B
ATLAS v2	Edit Dice ↑	MedEdit [C]	**0.7428**	<u>0.5936</u>	B,M

ISIC uses eight-view D4 test-time augmentation. Kvasir ensembles steps 1500, 2000, 3000, and 3500; BUSI ensembles 2500, 3000, and 3500. Current sample coverage is unspecified in the updated export; earlier partial counts are documented in Appendix H. MedSAM3 has possible training overlap and was evaluated outside the sealed campaign. HAT, Restormer, and InstructPix2Pix use out-of-domain checkpoints. WIA-LD2ND is a paper-guided repair. Lower staining FID does not imply higher marker agreement. Earlier BraTS-GLI and AAPM results appear separately in Appendix H.1.

marker agreement for staining; and, for editing, Edit Dice of the requested lesion together with MAE and SSIM outside the edit region. The two remaining training benchmarks, BraTS-GLI and AAPM Mayo LDCT, were evaluated in an earlier snapshot and are reported in Appendix H.1. Splits, label conventions, and metric definitions are given in Appendix E.

Baselines. The comparison contains 68 baseline entries across the 14 benchmarks, 62 of which produced scores. They include nnU-Net v2 (Isensee et al., 2021), nnWNet (Zhou et al., 2025), MedCLIPSeg (Koleilat et al., 2026), and MedSAM3 (Liu et al., 2025) for segmentation; Restormer (Zamir et al., 2022), HAT (Chen et al., 2023), WIA-LD2ND (Zhao et al., 2024), and GaussM2ASR for restoration and super-resolution; RegConDiS and SynDiff (Özbey et al., 2023) for translation; VSMT and TDKStain for staining; and MedEdit (Ben Alaya et al., 2024) and InstructPix2Pix (Brooks et al., 2023) for editing. Every entry is labeled by provenance: trained on the benchmark training split [T], a released checkpoint [C], or a reimplementation [R]. Released checkpoints applied outside their training domain, and those with possible training overlap, are marked in Appendix H.

Checkpoints and inference. The ten non-segmentation benchmarks use one checkpoint, the end of Stage B (microstep 3999). Restoration, super-resolution, and translation use 20 Euler steps without guidance and average eight generation seeds; LoDoPaB-CT and SurgiSR4K add a VAE residual correction. Staining uses a marker-specific prompt, and ATLAS editing composites the output inside the lesion mask. Segmentation uses dataset-specific checkpoints: a single checkpoint with eight-view test-time augmentation on ISIC, checkpoint ensembles on Kvasir-SEG (four) and BUSI (three), and a single checkpoint on DRIVE, each averaged over three seeds. The segmentation rows therefore evaluate the shared architecture and recipe rather than one set of weights. Seeds measure sampling variability and are not independent training runs.

5.2 MAIN RESULTS

Table 2 summarizes the comparison (per-dataset tables and plots in Appendices H and I). MedVision-X obtains the best primary endpoint on 8 of the 14 benchmarks and is second on two more.

Restoration and cross-modal synthesis. These tasks require accurate continuous intensities, and here MedVision-X is strongest. On LDCT projection it reaches a body HU MAE of 39.11, against 99 for the target-trained WIA-LD2ND and 110.2 for Restormer, and its body PSNR of 40.22 exceeds the best baseline by 4.71 dB. On LoDoPaB-CT it improves PSNR from 33.97 to 37.99 and SSIM from 0.8124 to 0.8774. On SynthRAD2023 it leads all three endpoints, with HU MAE 55.2 against 83.11 for RegConDiS, PSNR 30.99, and SSIM 0.8455. On IXI it leads PSNR (25.58 vs. 22.43), SSIM, and MAE; only LPIPS is lower for CRAVIM (0.1353 vs. 0.1667). All of these results come from the same checkpoint.

Segmentation. MedVision-X leads all four BUSI metrics, with Dice 0.8271 against 0.8097 for nnU-Net v2 and no false-positive pixels on normal images. On Kvasir-SEG it leads Dice (0.9046) and IoU (0.8543), although the Dice margin over MedCLIPSeg is only 0.0012 and MedCLIPSeg has lower HD95. It trails on ISIC, where the released MedSAM3 checkpoint, which may have seen dermoscopy data in training, reaches 0.9282 Dice against 0.8918, and on DRIVE, where nnWNet and nnU-Net v2 reach 0.8252 and 0.8236 against 0.7911. On DRIVE, MedVision-X still outperforms the remaining four methods, including PMRNet (0.7339) and MedSAM3 (0.1057).

Super-resolution. On fastMRI, MedVision-X attains the best perceptual score (LPIPS 0.1498 against 0.1718 for GaussM2ASR), while GaussM2ASR leads PSNR by 0.9 dB (36.76 vs. 35.86), as well as NMSE and SSIM. On SurgiSR4K, HAT leads PSNR and SSIM and TADSR leads LPIPS, and MedVision-X ranks second in SSIM. This is the expected perception–distortion tradeoff of a generative model.

Virtual staining. MedVision-X produces the most realistic stain distributions on both datasets, lowering FID from 44.88 to 31.66 on MIST and from 47.27 to 39.12 on BCI relative to VSMT, with the lowest KID in both cases. On MIST it also has the highest marker agreement (0.6476), marginally above TDKStain (0.6452). On BCI its HER2 agreement (0.4084) is below VSMT (0.4442).

Lesion editing. On BraTS, MedVision-X achieves Edit Dice 0.4466, above the target-trained MedEdit (0.3499) and below InstructPix2Pix (0.4849), while leading every preservation and realism metric. On ATLAS, MedEdit reaches a higher Edit Dice (0.7428 vs. 0.5936).

5.3 ANALYSIS

Distribution realism is not marker fidelity. On BCI, MedVision-X has the lowest FID but trails VSMT in HER2 agreement and LPIPS, and on MIST its agreement lead is only 0.0024 (Figure 8 in the appendix). Lower FID means that generated patches look like real IHC, not that they express the correct marker level; both metrics remain proxies for biological equivalence.

Requested change versus preservation. Returning the input unchanged would preserve context perfectly, so preservation must be read with lesion realization. On BraTS, MedVision-X changes almost nothing outside the edit region (outside MAE 1.91×10^{-3}, SSIM 0.9961; MedEdit 5.52×10^{-3}, 0.9638) while realizing the lesion more accurately than MedEdit. On ATLAS, the zero outside error follows from mask compositing.

Is the frozen image representation a bottleneck for masks? Masks are generated in the latent space of a frozen VAE, which could erase thin structures. We encode and decode ground-truth validation masks and recover them with the same decoding rule (Appendix F). On DRIVE, Dice drops only from 0.976 (transform only) to 0.963 after the VAE round trip; on an ISIC subset the two are identical to three decimals (0.999). With MedVision-X at 0.791 on DRIVE, the gap to specialists lies mainly in conditional prediction; with four DRIVE and eight ISIC images, this indicates feasibility rather than a precise ceiling.

6 Discussion and limitations

Implications. The results suggest that the output space of a unified generator can serve as a common interface for medical analysis and synthesis. Its strength is highest where targets carry physical meaning, such as Hounsfield units, and its weakness is concentrated on fine and thin structures: DRIVE vessels, ISIC borders, small BraTS-GLI subregions, and pixel-exact super-resolution. Because the mask diagnostic rules out the representation as the main cause, improving conditional spatial precision is the most direct route to closing these gaps.

Scope and limitations. The single-checkpoint evidence covers the ten non-segmentation benchmarks; segmentation rows use dataset-specific checkpoints, ensembles, and test-time augmentation from the same training run. We evaluate the full recipe and do not yet isolate the contribution of each stage or test transfer against single-task training, so we make no claim about which component drives the gains. Stage B uses only benchmark training folds, but the public Stage-A aggregates were not audited image by image against the test splits. Baselines differ in provenance and cost: some are released out-of-domain checkpoints or reimplementations, MedVision-X has 14.6B parameters, and restoration outputs average eight samples. We report point estimates without confidence intervals or per-dataset sample counts, do not compare with MVG or UniMedVL, process 2D slices only, and note that generated stains and lesions require biological and clinical validation. Single-checkpoint evaluation of all 16 benchmarks, stage ablations, and an overlap audit are the immediate next steps.

7 Conclusion

We presented MedVision-X, a generative generalist that casts medical segmentation, super-resolution, cross-modal translation, virtual staining, CT denoising, and lesion editing as instruction-conditioned image generation with one output pathway and one objective. Under the reported protocols it obtains the best primary endpoint among the listed baselines on 8 of 14 benchmarks, with the largest gains on HU-accurate CT restoration and MRI-to-CT synthesis, and its gaps on thin structures define the next step for generative unification in medical imaging.

Reproducibility statement

The task representation, training stages, objective, and inference procedure are specified in Sections 3 and 4. The full architecture, optimization settings, and sampler are given in Appendices A and B; the Stage-A and Stage-B data mixtures in Appendix C; and dataset splits, label conventions, and metric definitions in Appendix E. For every dataset, Appendix H records the evaluated checkpoint, number of sampling steps and seeds, guidance setting, and postprocessing, together with the training provenance of each baseline. Code, configuration files, evaluation scripts, and result summaries are provided in an anonymized repository at `https://anonymous.4open.science/r/MedVision-X-ICLR27`, and the evaluated checkpoints will be released upon publication.

Ethics statement

This work uses retrospective, publicly available medical imaging benchmarks under their original data-use conditions; no new data were collected from patients. Generated images and edits are research outputs, and benchmark performance does not establish clinical validity. Virtual stains require biological validation, and constructed lesion edits do not represent observed patient counterfactuals. Clinical use would require independent validation of pathology preservation, population coverage, and three-dimensional consistency.

AI-use statement

Generative AI tools assisted with source-code inspection, literature organization, drafting and editing of the manuscript text, and writing LaTeX and TikZ code for tables, schematics, and plots. All numerical results come from the authors' recorded model evaluations, and every plot is drawn from those recorded values. No figure contains AI-generated medical images: all image tiles are

unretouched source images, references, or MedVision-X outputs. The authors verified the text, references, figures, and results and take full responsibility for the content.

REFERENCES

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020.

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

Malek Ben Alaya, Daniel M Lang, Benedikt Wiestler, Julia A Schnabel, and Cosmin I Bercea. MedEdit: Counterfactual diffusion-based image editing on brain MRI. *arXiv preprint arXiv:2407.15270*, 2024.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402, 2023.

Hu Chen, Yi Zhang, Mannudeep K Kalra, Feng Lin, Yang Chen, Peixi Liao, Jiliu Zhou, and Ge Wang. Low-dose CT with a residual encoder-decoder convolutional neural network. *IEEE Transactions on Medical Imaging*, 36(12):2524–2535, 2017.

Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen 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 (ISIC). *arXiv preprint arXiv:1902.03368*, 2019.

Marcos V Conde, Ui-Jin Choi, Maxime Burchi, and Radu Timofte. Swin2SR: SwinV2 transformer for compressed image super-resolution and restoration. In *European Conference on Computer Vision Workshops*, pp. 669–687. Springer, 2022.

Contrast-X. Contrast-X. Hugging Face dataset repository, `https://huggingface.co/datasets/Contrast-X/Contrast-X`, 2026. Accessed 2026-08-30.

Maria Correia de Verdier, Rachit Saluja, Louis Gagnon, Dominic LaBella, Ujjwall Baid, Nourel Hoda Tahon, Martha Foltyn-Dumitru, Jikai Zhang, Maram Alafif, Saif Baig, et al. The 2024 brain tumor segmentation (BraTS) challenge: Glioma segmentation on post-treatment MRI. *arXiv preprint arXiv:2405.18368*, 2024.

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.

Xiaogang Du, Dongxin Gu, Tao Lei, Yipeng Jiao, and Yibin Zou. PMR-Net: Parallel multi-resolution encoder-decoder network framework for medical image segmentation. *arXiv preprint arXiv:2409.12678*, 2024.

Naomi Fridman, Bubby Solway, Tomer Fridman, Itamar Barnea, and Anat Goldstein. BreastDCEDL: A comprehensive breast cancer DCE-MRI dataset and transformer implementation for treatment response prediction. *arXiv preprint arXiv:2506.12190*, 2025.

Valentin Gabeur, Shangbang Long, Songyou Peng, Paul Voigtlaender, Shuyang Sun, Yanan Bao, Karen Truong, Zhicheng Wang, Wenlei Zhou, Jonathan T Barron, et al. Image generators are generalist vision learners. *arXiv preprint arXiv:2604.20329*, 2026.

Lidia Garrucho, Kaisar Kushibar, Claire-Anne Reidel, Smriti Joshi, Richard Osuala, Apostolia Tsirikoglou, Maciej Bobowicz, Javier Del Riego, Alessandro Catanese, Katarzyna Gwoździewicz, et al. A large-scale multicenter breast cancer DCE-MRI benchmark dataset with expert segmentations. *Scientific Data*, 12(1):453, 2025.

Xiaoyang Han, Jianhua Li, Kewang Deng, Zukai Chen, Xuanke Shi, Sihan Wang, Boxuan Li, Linyan Wang, Siyi Xie, Xin You, et al. Vision as unified multimodal generation. *arXiv preprint arXiv:2607.06560*, 2026.

Yufan He, Pengfei Guo, Yucheng Tang, Andriy Myronenko, Vishwesh Nath, Ziyue Xu, Dong Yang, Can Zhao, Benjamin Simon, Mason Belue, et al. VISTA3D: A unified segmentation foundation model for 3D medical imaging. *arXiv preprint arXiv:2406.05285*, 2024.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, 2017.

IXI Project. IXI dataset. `https://brain-development.org/ixi-dataset/`, n.d. Accessed 2026-09-09.

Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-SEG: A segmented polyp dataset. In *International Conference on Multimedia Modeling (MMM)*, volume 11962 of *Lecture Notes in Computer Science*, pp. 451–462. Springer, 2020.

Fengyi Jiang, Xiaorui Zhang, Lingbo Jin, Ruixing Liang, Yuxin Chen, Adi Chola Venkatesh, Jason Culman, Tiantian Wu, Lirong Shao, Wenqing Sun, Cong Gao, Hallie McNamara, Jingpei Lu, and Omid Mohareri. SurgiSR4K: A high-resolution endoscopic video dataset for robotic-assisted minimally invasive procedures. *Machine Learning for Biomedical Imaging*, pp. 875–885, 2025. doi: 10.59275/j.melba.2025-f593.

Anahita Fathi Kazerooni, Nastaran Khalili, Xinyang Liu, Debanjan Haldar, Zhifan Jiang, Anna Zapaishchykova, Julija Pavaine, Lubdha M Shah, Blaise V Jones, Nakul Sheth, et al. BraTS-PEDs: Results of the multi-consortium international pediatric brain tumor segmentation challenge 2023. *Machine Learning for Biomedical Imaging*, pp. 72–87, 2025. doi: 10.59275/j.melba.2025-f6fg.

Taha Koleilat, Hojat Asgariandehkordi, Omid Nejati Manzari, Berardino Barile, Yiming Xiao, and Hassan Rivaz. MedCLIPSeg: Probabilistic vision-language adaptation for data-efficient and generalizable medical image segmentation. *arXiv preprint arXiv:2602.20423*, 2026.

Nicholas Konz, Yuwen Chen, Haoyu Dong, and Maciej A Mazurowski. Anatomically-controllable medical image generation with segmentation-guided diffusion models. *arXiv preprint arXiv:2402.05210*, 2024.

Johannes Leuschner, Maximilian Schmidt, Daniel Otero Baguer, and Peter Maass. LoDoPaB-CT, a benchmark dataset for low-dose computed tomography reconstruction. *Scientific Data*, 8(1):109, 2021.

Fangda Li, Zhiqiang Hu, Wen Chen, and Avinash Kak. Adaptive supervised PatchNCE loss for learning H&E-to-IHC stain translation with inconsistent groundtruth image pairs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 632–641. Springer, 2023a.

Guangyuan Li, Chen Rao, Juncheng Mo, Zhanjie Zhang, Wei Xing, and Lei Zhao. Rethinking diffusion model for multi-contrast MRI super-resolution. *arXiv preprint arXiv:2404.04785*, 2024.

Lei Li, Fuping Wu, Sihan Wang, Xinzhe Luo, Carlos Martín-Isla, Shuwei Zhai, Jianpeng Zhang, Yanfei Liu, Zhen Zhang, Markus J Ankenbrand, et al. MyoPS: A benchmark of myocardial pathology segmentation combining three-sequence cardiac magnetic resonance images. *Medical Image Analysis*, 87:102808, 2023b.

Sook-Lei Liew, Bethany P Lo, Miranda R Donnelly, Artemis Zavaliangos-Petropulu, Jessica N Jeong, Giuseppe Barisano, Alexandre Hutton, Julia P Simon, Julia M Juliano, Anisha Suri, et al. A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. *Scientific Data*, 9(1):320, 2022.

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

Anglin Liu, Xu R Cao, Yifan Shen, Yi Lu, Xiang Li, Qianqian Chen, and Jintai Chen. MedSAM3: Delving into segment anything with medical concepts. *arXiv preprint arXiv:2511.19046*, 2025.

Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Yizhou Yu, Yong Liang, Guangming Shi, Shaoting Zhang, Hairong Zheng, and Shanshan Wang. Swin-UMamba: Mamba-based UNet with ImageNet-based pretraining. *arXiv preprint arXiv:2402.03302*, 2024.

Shengjie Liu, Chuang Zhu, Feng Xu, Xinyu Jia, Zhongyue Shi, and Mulan Jin. BCI: Breast cancer immunohistochemical image generation through pyramid Pix2pix. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1814–1823, 2022.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

Cynthia H McCollough, Adam C Bartley, Rickey E Carter, Baiyu Chen, Tammy A Drees, Phillip Edwards, David R Holmes III, Alice E Huang, Farhana Khan, Shuai Leng, et al. Low-dose CT for the detection and classification of metastatic liver lesions: Results of the 2016 low dose CT grand challenge. *Medical Physics*, 44(10):e339–e352, 2017.

Hyungi Min, Taeseung You, Hangyeul Lee, Yeongjae Cho, and Sungzoon Cho. An interpretable local editing model for counterfactual medical image generation. *arXiv preprint arXiv:2603.00423*, 2026.

Taylor R Moen, Baiyu Chen, David R Holmes III, Xinhui Duan, Zhicong Yu, Lifeng Yu, Shuai Leng, Joel G Fletcher, and Cynthia H McCollough. Low-dose CT image and projection dataset. *Medical Physics*, 48(2):902–911, 2021. doi: 10.1002/mp.14594.

David Newitt and Nola Hylton. Single site breast DCE-MRI data and segmentations from patients undergoing neoadjuvant chemotherapy. The Cancer Imaging Archive, 2016.

Junzhi Ning, Wei Li, Cheng Tang, Jiashi Lin, Chenglong Ma, Chaoyang Zhang, Jiyao Liu, Ying Chen, Shujian Gao, Yuandong Pu, et al. UniMedVL: Unifying medical multimodal understanding and generation through observation-knowledge-analysis. *arXiv preprint arXiv:2510.15710*, 2025.

Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Özturk, Alper Güngör, and Tolga Çukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 2023.

Md Mostafijur Rahman, Mustafa Munir, and Radu Marculescu. EMCAD: Efficient multi-scale convolutional attention decoding for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Sucheng Ren, Xiaoke Huang, Xianhang Li, Junfei Xiao, Jieru Mei, Zeyu Wang, Alan Yuille, and Yuyin Zhou. Medical vision generalist: Unifying medical imaging tasks in context. *arXiv preprint arXiv:2406.05565*, 2024.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, 23(4):501–509, 2004.

Adrian Thummerer, Erik van der Bijl, Arthur Galapon Jr, Joost JC Verhoeff, Johannes A Langendijk, Stefan Both, Cornelis AT van den Berg, and Matteo Maspero. SynthRAD2023 grand challenge dataset: Generating synthetic CT for radiotherapy. *Medical Physics*, 50(7):4664–4674, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. SegMamba: Long-range sequential modeling mamba for 3D medical image segmentation. *arXiv preprint arXiv:2401.13560*, 2024.

Song Xue, Hanzhong Wang, Yizhou Chen, Fanxuan Liu, Hong Zhu, Marco Viscione, Rui Guo, Axel Rominger, Biao Li, and Kuangyu Shi. UDPET: Ultra-low dose PET imaging challenge dataset. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 616–623. Springer, 2025.

Chao Yang, Boqian Zhang, Jinghao Xu, and Guang Jiang. HDW-SR: High-frequency guided diffusion model based on wavelet decomposition for image super-resolution. *arXiv preprint arXiv:2511.13175*, 2025.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, et al. fastMRI: An open dataset and benchmarks for accelerated MRI. *arXiv preprint arXiv:1811.08839*, 2018.

Tianyi Zhang, Zheng-Peng Duan, Peng-Tao Jiang, Bo Li, Ming-Ming Cheng, Chun-Le Guo, and Chongyi Li. Time-aware one step diffusion network for real-world image super-resolution. *arXiv preprint arXiv:2508.16557*, 2025.

Yide Zhang, Yinhao Zhu, Evan Nichols, Qingfei Wang, Siyuan Zhang, Cody Smith, and Scott Howard. A Poisson-Gaussian denoising dataset with real fluorescence microscopy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11702–11710, 2019.

Haoyu Zhao, Yuliang Gu, Zhou Zhao, Bo Du, Yongchao Xu, and Rui Yu. WIA-LD2ND: Wavelet-based image alignment for self-supervised low-dose CT denoising. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 15007 of *Lecture Notes in Computer Science*, pp. 764–774. Springer, 2024. doi: 10.1007/978-3-031-72104-5_73.

Hong-Yu Zhou, Julián Nicolás Acosta, Subathra Adithan, Suvrankar Datta, Eric J Topol, and Pranav Rajpurkar. A generalist learner for multifaceted medical image interpretation. *arXiv preprint arXiv:2405.07988*, 2024.

Ruofan Zhou, Majed El Helou, Daniel Sage, Thierry Laroche, Arne Seitz, and Sabine Süsstrunk. W2S: Microscopy data with joint denoising and super-resolution for widefield to SIM mapping. In *European Conference on Computer Vision Workshops*, pp. 474–491. Springer, 2020.

Yanfeng Zhou, Lingrui Li, Le Lu, and Minfeng Xu. nnWNet: Rethinking the use of transformers in biomedical image segmentation and calling for a unified evaluation benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20852–20862, 2025.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, 2017.

APPENDIX

A ARCHITECTURE AND TARGET REPRESENTATION

MedVision-X continues the SenseNova-Vision/BAGEL architecture. Medical task families share the routed transformer, visual encoder, connectors, latent output head, and VAE decoder. The vocabulary head supports the optional segmentation legend; it is not a separate medical prediction head.

Table 3: Architecture used for medical adaptation. Counts refer to physically stored parameters, including both transformer routes.

Component	Configuration
Qwen2 routed transformer	28 layers; hidden width 3,584; MLP width 18,944; 28 query heads and 4 key/value heads; vocabulary 152,064; RoPE base 10^6
Visual encoder	SigLIP-style encoder; 26 active blocks; hidden width 1,152; MLP width 4,304; 16 heads; patch size 14
Frozen VAE	RGB input; 16 latent channels; spatial downsampling factor 8; latent scale 0.3611 and shift 0.1159; 83,819,683 parameters
Latent patchification	2×2 latent patches; 64 values per token; projection to width 3,584; effective image stride 16
Semantic connector	Two-layer GELU MLP, $1152 \to 3584 \to 3584$
Image output	Shared $3584 \to 64$ velocity projection and frozen VAE decoder
Position representation	Sequence RoPE and fixed 2D sinusoidal tables for visual and latent grids
Routed transformer parameters	14,141,252,608
Visual encoder parameters	402,552,736
Visual connector parameters	16,980,992
Timestep embedder parameters	13,769,728
Fixed 2D position parameters	32,241,664
Latent projection parameters	462,400
Main-model total	**14,607,260,128**; approximately 14.575B trainable

The inherited checkpoint family name contains `7B`, but it does not describe the total stored size of the two-route model. The VAE is stored separately and remains frozen. Medical adaptation updates the transformer, visual encoder, and connectors without LoRA; the external fixed position tables are not optimized.

Token routing and attention. Text and semantic-image tokens use the understanding route; clean source latents and noisy or generated target latents use the generation route. Each route has its own RMSNorm, attention projections, and MLP parameters. Tokens communicate through attention in the same sequence, so this is deterministic modality routing rather than learned selection among medical-task experts. Packed examples are mutually isolated. Text follows causal attention; clean visual blocks use bidirectional attention within their blocks. A noisy target block attends to the preceding conditioning context and to itself, while conditioning tokens cannot attend back into that target block.

Segmentation representation. The data loader reads a 2D integer annotation and renders label $k > 0$ using the kth declared RGB color; label zero is black. Class names and RGB values are provided by the same record. The general palette is inherited from the SenseNova segmentation interface, with colors drawn from 200 separated anchors. The supervised text enumerates class–color bindings and a segmentation marker precedes the image output. Binary benchmark tasks use a white foreground, so generated luminance yields a score in $[0, 1]$. These scores are not assumed to be calibrated probabilities. Multiclass predictions use nearest-palette decoding. The output is resized to the target dimensions with nearest-neighbor interpolation before categorical evaluation.

Spatial scope. Volumetric source data are materialized and loaded as individual 2D slices. There are no adjacent-slice inputs, volume tokens, through-plane position embeddings, or cross-slice attention in this implementation. Some tasks provide additional spatial conditioning: ATLAS editing records

contain a factual source image and an edit mask. These enter the shared sequence as conditioning images, without an additional task head.

B Optimization and flow integration

Let $a \sim \mathcal{N}(0,1)$ and $\tau = \sigma(a)$. The training timestep is shifted toward noise according to

$$t = \frac{4\tau}{1+3\tau}, \qquad z_t = (1-t)z_0 + t\epsilon, \qquad v^\star = \epsilon - z_0. \tag{3}$$

For the set Ω of supervised target-latent tokens, the image objective is

$$\mathcal{L}_{\text{flow}} = \frac{1}{|\Omega|} \sum_{i \in \Omega} \frac{1}{64} \left\| v_\theta(z_t, x, p, t)_i - v_i^\star \right\|_2^2 . \tag{4}$$

The distributed implementation normalizes image loss by the number of supervised latent tokens and text loss by the number of supervised text tokens. If target text is present, the total objective is $\mathcal{L}_{\text{flow}} + 0.25\mathcal{L}_{\text{CE}}$. Independent classifier-free conditioning dropout probabilities are 0.05 for text and 0.10 each for semantic features and clean source latents.

Table 4: Training schedules of the two adaptation stages. A microstep is one packed batch per rank; two microsteps constitute one optimizer update.

Setting	Stage A	Stage B
Packed microsteps	50,000	4,000
Optimizer updates	25,000	2,000
Learning rate	2.5×10^{-5}	5×10^{-6}
Warmup optimizer updates	500	100
Gradient accumulation	2	2
EMA decay	.995	.995
Checkpoint interval (microsteps)	250	500
GPUs	16	16

All schedules use fused AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-15}$, zero weight decay, gradient norm clipping at 1, and a constant learning rate after linear warmup. Model training uses BF16, hybrid sharded FSDP, non-reentrant activation checkpointing, and eight nodes with two GPUs per node. New adaptation stages initialize from the preceding EMA weights. Warmup is counted in optimizer updates, whereas checkpoint names count packed microsteps; the final index of a 4,000-microstep run is `0003999`.

Resolution and packing. The VAE image transform uses stride 16 and configured minimum/maximum image sizes of 512/1,024; the semantic image transform uses stride 14 and limits of 224/980. The soft packed-token target is 32,768 per rank and the ordinary per-example limit is 24,064 tokens. Earlier Stage-A training uses a 36,864-token hard pack limit; the continuation configuration and Stage-B launcher use 32,768. The corrected Stage-B loader buffers examples that do not fit a partial pack. When buffering cannot accommodate a valid example, it emits the partial pack and places that example into the next pack. This prevents overflow-induced rejection of otherwise admissible examples. The explicit length filter remains; sampling probabilities need not equal realized token proportions.

Image sampling. The default sampler uses EMA weights and a shifted Euler grid. For $N = 50$, the implementation constructs N points from one to zero, applies the same rational timestep shift, and integrates over the $N - 1 = 49$ adjacent intervals. Writing these points as $t_0 > \cdots > t_{49}$, the update is

$$z_{k+1} = z_k - v_\theta(z_k, x, p, t_k)(t_k - t_{k+1}), \qquad k = 0, \ldots, 48, \tag{5}$$

with classifier-free guidance applied to the velocity when enabled. Thus `num_timesteps=50` denotes 50 grid points, not 50 Euler updates. The current restoration, super-resolution, and translation evaluations report 20 sampling steps with eight-seed averaging and CFG off. Dataset-specific checkpoint, aggregation, and postprocessing settings appear in Appendix H; the 50-point default above is not the configuration reported for those runs.

C MEDICAL TRAINING DATA

Table 5: Stage-A data mixture. Counts are training records. Expanded-mixture draw probabilities equal the listed weight divided by 11.9.

Stream	Mapping	Records	Weight
Original generation	Reconstruction	41,133	2.00
Original segmentation	Segmentation	861,344	1.50
Original super-resolution	Resolution enhancement	695	1.00
Original cross-modal	Modality translation	41,174	1.00
Original IHC	Virtual staining	17,295	1.00
Original denoising	Restoration	13,922	1.00
Original counterfactual	Image editing	2,943	1.00
Contrast-X	Noncontrast→contrast CT	5,450	.45
MAMA-MIA	Pre→post breast MRI	5,029	.35
BreastDCEDL	DCE phase 0→1	1,067	.30
Breast-MRI-NACT	Pre→post breast MRI	7,254	.20
BraTS 2021	T1→T1ce	9,920	.15
BraTS-PEDs	T1n→T1c	5,078	.30
MyoPS 2020	C0/T2→delayed MRI	204	.25
FMD	Noisy→averaged microscopy	2,400	.35
W2S	Widefield→SIM	72	.25
AAPM LDCT	Low→full-dose CT	16,628	.40
UDPET	Low/noisy→standard PET	4,828	.40
Total	**18 streams**	**1,036,436**	**11.90**

The seven original streams are task aggregates; the eleven additions preserve named-source membership. The original segmentation aggregate includes 857,364 SA-Med2D-derived rows and 3,980 BraTS-GLI slices. Its generated records do not retain fine-grained upstream labels, so these counts cannot establish the number of independent source datasets or patients. Stage A uses the seven original streams through microstep 27,250 and all 18 from microstep 27,251. A comparison across this boundary changes both training duration and source mixture.

Table 6: Stage-B training records and mixture weights. Each task family has total weight 2. Records are not counts of independent subjects.

Family	Dataset	Records	Weight
Segmentation	ISIC 2018	806	.48
	Kvasir-SEG	800	.48
	BUSI	624	.48
	DRIVE	16	.08
	BraTS-GLI	29,947	.48
Cross-modal	SynthRAD2023	44,562	1.00
	IXI	69,004	1.00
Denoising/restoration	AAPM Mayo LDCT	4,800	$2/3$
	LoDoPaB-CT	35,820	$2/3$
	LDCT-and-Projection-data	20,522	$2/3$
Super-resolution	fastMRI SR	70,748	1.60
	SurgiSR4K	603	.40
Counterfactual	BraTS tumor editing	2,900	1.80
	ATLAS v2	168	.20
Virtual staining	MIST	15,566	1.00
	BCI	3,506	1.00
Total	**16 datasets**	**300,392**	**12.00**

Stage-B records are materialized from fixed benchmark training folds, with validation and test roles rejected by the preparer. This restriction does not establish absence of earlier exposure in the pretrained backbone or Stage-A aggregates. Fine-grained provenance is incomplete for some original aggregates; the original MIMIC-CXR reconstruction split is image-level, and the original MIST

split is patch-level and may share whole slides. Patient-disjoint or slide-disjoint generalization must therefore be tied to the specific dataset and stage rather than applied to the entire training lineage.

Preprocessing follows the source contracts, including alignment for paired tasks, foreground filtering, and intensity normalization before 2D export. The LoDoPaB input here is an FBP reconstruction, not a sinogram; the auxiliary Stage-A reconstruction mapping uses the source image itself as the target. Image fidelity can be measured in the normalized target space or, where an inverse intensity transform is defined, in physical units. Physical-unit metrics must state that transform and the precision of the exported images. Constructed editing supervision and evaluation against an unpaired population of real images answer different questions and must be reported separately.

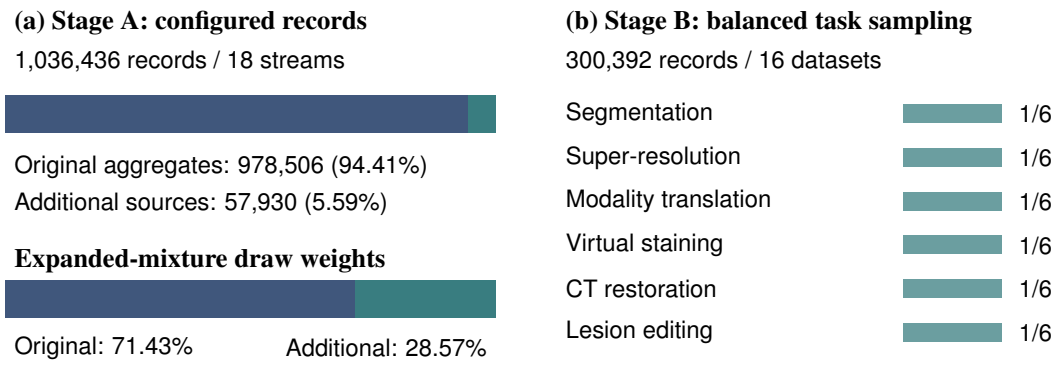

Figure 3: **Medical data composition and configured sampling.** Record counts and sampler probabilities measure different quantities: Stage A increases exposure to smaller additional sources, while Stage B allocates equal total weight to six task families. Bars show configured probabilities, not measured realized draws. Counts are records rather than independent patients.

D INSTRUCTION–TARGET EXAMPLES

Figure 4 illustrates the source, instruction, and target format of each task family used in Stages A and B.

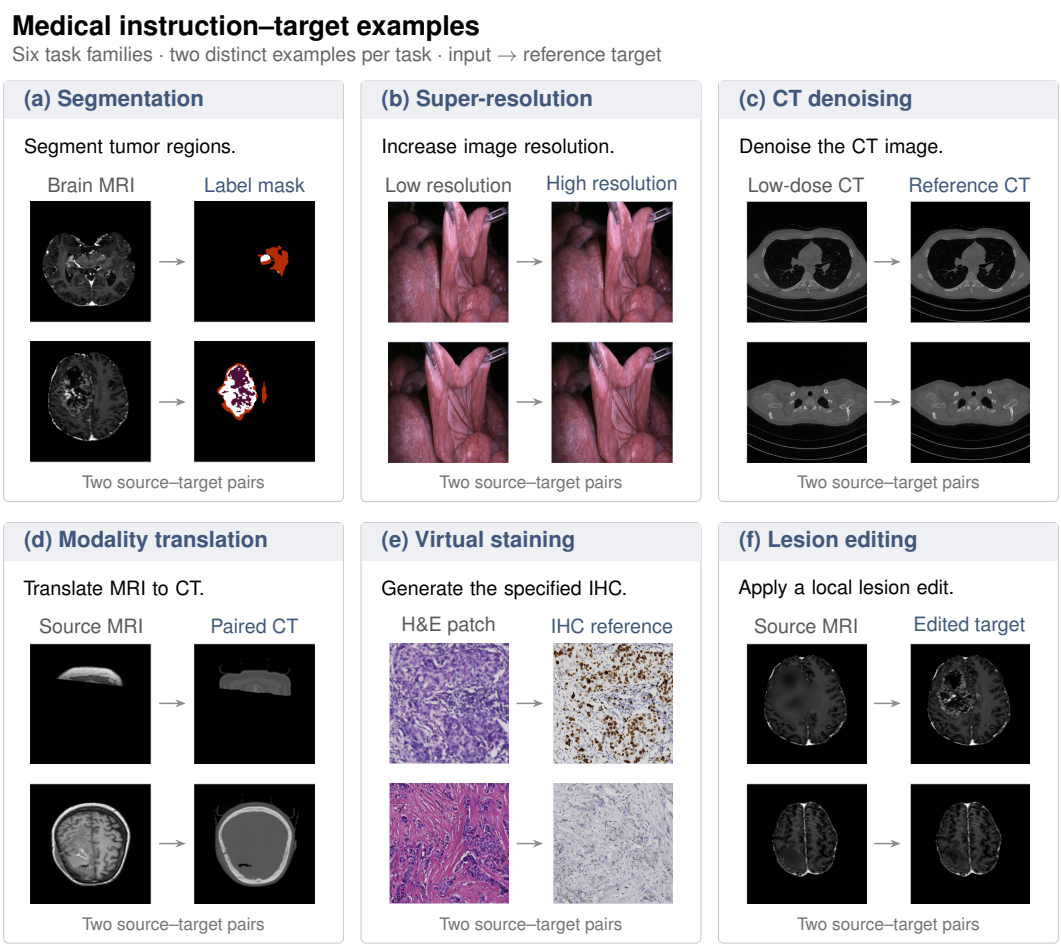

Figure 4: **Instruction–target examples across the six task families.** Each panel shows two source images and their reference targets, distinct from the examples in Figure 1; no model prediction is shown. Instructions are paraphrased, and marker and spatial conditions are omitted for compactness. Segmentation colors: tumor (white), edema (orange), necrosis (purple). Editing targets are constructed supervision, not observed patient counterfactuals.

E Dataset and evaluation contracts

E.1 Dataset inventory and partitions

Table 7 lists the frozen benchmark partitions. These downstream training records are distinct from the broad adaptation corpus. Counts refer to the stated units: a subject, volume, slice, patch, and frame are not interchangeable. Image-domain super-resolution is also distinct from accelerated MRI reconstruction from undersampled k-space. The LoDoPaB condition is a reconstructed FBP image, and the LDCT projection dataset is used through its reconstructed DICOM pairs.

Segmentation sources are ISIC 2018, Kvasir-SEG, BUSI, DRIVE, and postoperative BraTS-GLI (Codella et al., 2019; Jha et al., 2020; Al-Dhabyani et al., 2020; Staal et al., 2004; de Verdier et al., 2024). Translation uses SynthRAD2023 and IXI (Thummerer et al., 2023; IXI Project, n.d.). Restoration uses the AAPM challenge data, LoDoPaB-CT, the low-dose CT image and projection resource, fastMRI, and SurgiSR4K (McCollough et al., 2017; Leuschner et al., 2021; Moen et al., 2021; Zbontar et al., 2018; Jiang et al., 2025). Staining uses MIST and BCI (Li et al., 2023a; Liu et al., 2022); the second editing source is ATLAS v2 (Liew et al., 2022).

ISIC, Kvasir, and BUSI follow the pinned author-provided split workbooks recorded in the benchmark registry, with the documented ISIC split correction. DRIVE uses 16/4 training/validation cases from

Table 7: **Frozen downstream splits.** Counts are train/validation/test in the indicated units. Partitions reproduce the registered benchmark protocol; they are not uniformly official or patient-disjoint across all sources.

Dataset	Train	Validation	Test	Unit
ISIC 2018	806	90	379	Images
Kvasir-SEG	800	100	100	Images
BUSI	624	78	78	Images
DRIVE	16	4	20	Eyes/images
BraTS-GLI	217	27	27	Subjects
SynthRAD2023	288	36	36	Subjects
IXI	462	58	57	Subjects
AAPM Mayo LDCT	4,800	526	610	Slices
LoDoPaB-CT	35,820	3,522	3,553	Slices
LDCT projection	20,522	2,489	2,461	Slices
fastMRI SR	4,469	1,378	558	Volumes
SurgiSR4K	603	72	125	Frames
BraTS editing	2,900	442	363	Slices
ATLAS v2 editing	168	10	21	Constructed samples
MIST	15,566	1,729	4,000	Paired patches
BCI	3,506	390	977	Paired patches

official training eyes 21–40 and all 20 official test eyes 01–20. BraTS is partitioned by subject. SynthRAD partitions brain and pelvis independently, and IXI uses a site-stratified subject split. AAPM uses an 8/1/1 patient partition; the fastMRI volume counts correspond to 70,748/21,842/8,852 slices. SurgiSR4K uses 19/2/3 videos. BraTS editing contains 211/27/27 usable cases. ATLAS construction preserves study-level separation; its lesion-positive reference comes from another subject and is unpaired with the factual input.

MIST preserves the author test split and uses 1,000 test patches per marker; whole-slide overlap exists across its source partitions. BCI preserves the author test split, while patient and whole-slide identities are unavailable. Results on these partitions characterize performance under the stated source protocol, not independently demonstrated patient-level generalization. Training overlap of released baseline weights is a separate property and must accompany the affected method's result. Similarly, the aggregate sources in broad medical adaptation require source-level overlap checks before a dataset can be described as unseen.

E.2 Segmentation labels and metrics

ISIC and Kvasir report Dice, IoU, and HD95. BUSI unions all lesion masks for an image and additionally measures the false-positive pixel fraction on normal images, not a patient-level false-diagnosis rate. DRIVE reports Dice, IoU, AUC, sensitivity, and specificity inside the official field of view. Thresholds are chosen on validation data. Empty-mask handling and the support of finite boundary distances must be reported; discarding infinite HD95 values without disclosing those cases changes the meaning of a cohort summary.

The BraTS-GLI contract is the *2024 post-treatment* label convention, with four foreground labels and six evaluated regions:

Region	Meaning	Label set
ET	Enhancing tumor	$\{3\}$
NETC	Non-enhancing tumor core	$\{1\}$
RC	Resection cavity	$\{4\}$
SNFH	Surrounding non-enhancing FLAIR hyperintensity	$\{2\}$
TC	Tumor core	$\{1, 3\}$
WT	Whole tumor	$\{1, 2, 3\}$

Dice and HD95 are reported for each of these six regions. Resection cavity is excluded from TC and WT. A WT-only baseline supplies a partial comparison and cannot establish performance across

the full six-region contract. Slice-wise model outputs must be reconstructed and evaluated with the registered spatial convention; the model itself remains two-dimensional.

E.3 Restoration and modality translation

SynthRAD reports body-masked HU MAE, PSNR, and SSIM. IXI evaluates registered T1-to-PD translation using PSNR, SSIM, MAE, and LPIPS. AAPM and the reconstructed LDCT projection pairs report body-masked HU MAE, RMSE, PSNR, and SSIM. LoDoPaB reports RMSE, PSNR, and SSIM in its normalized image domain. fastMRI SR reports NMSE, PSNR, SSIM, and LPIPS for image-domain $4\times$ super-resolution; SurgiSR4K reports RGB PSNR, RGB SSIM, and LPIPS for its native $4\times$ pairs.

All methods must use the same output grid, valid region, data range, and aggregation rule for a given endpoint. A normalized-image MAE is not a HU MAE. The source image resizing used during training does not itself establish performance at a benchmark's native resolution; the final inference procedure must specify any tiling, reconstruction, or resampling. In particular, SurgiSR4K uses native 960×540 to 3840×2160 pairs. The checkpoint, number of sampling steps and seeds, guidance setting, and postprocessing used for each reported run are listed with the corresponding table in Appendix H.

E.4 Virtual staining

MIST evaluates H&E-to-IHC synthesis for ER, PR, HER2, and Ki67. BCI evaluates H&E-to-HER2 synthesis. Both report FID, KID, LPIPS, SSIM, and an auxiliary classifier agreement measure. MIST FID/KID are computed separately for each marker and then averaged equally; BCI uses its HER2 distribution. Distribution metrics require the full feature population and are not averages of per-image FID values.

Let $a(\hat{y}_i)$ be the prediction of the frozen IHC classifier and c_i the intended label. For groups $g \in \mathcal{G}$, agreement is

$$A = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \frac{1}{|g|} \sum_{i \in g} \mathbf{1}[a(\hat{y}_i) = c_i]. \tag{6}$$

For MIST, c_i is one of the four stain markers. For BCI, it is the inherited source score $0, 1+, 2+, 3+$. The auxiliary classifier uses frozen Inception features and a classifier trained only on real training IHC, selected using real validation macro-F1. Its agreement on real test targets is reported alongside generated-image agreement to characterize the measurement. These endpoints are classifier-based proxies, not pathologist agreement or proof of biomarker equivalence.

E.5 Lesion editing

The two editing benchmarks use lesion-addition contracts. BraTS provides a constructed factual input and a paired original T1c target. Its construction removes tumor labels $\{1, 2, 3\}$ while preserving cavity label 4. ATLAS pairs a real lesion-free factual slice with a desired lesion mask and an unpaired lesion-positive reference from another subject. That reference supports realism metrics and does not define a pixelwise target for the factual input.

Let m denote the desired lesion mask, r the locked edit region, x the factual input, and $s(\hat{y})$ a fixed domain-specific segmentation of the generated image. The principal structural and preservation endpoints are

$$\mathrm{EditDice} = \mathrm{Dice}\big(s(\hat{y}), m\big), \tag{7}$$

$$\mathrm{OutsideMAE} = \frac{\sum_{u \notin r} |\hat{y}_u - x_u|}{|\{u : u \notin r\}|}. \tag{8}$$

The full contract also includes outside-region SSIM, FID, and KID. The auxiliary lesion segmenter and region construction are fixed across methods. An identity prediction $\hat{y} = x$ would satisfy the preservation metrics without realizing the requested lesion, so Edit Dice and preservation are always reported together.

E.6 Checkpoints, seeds, and comparison scope

The ten non-segmentation datasets are evaluated with one checkpoint, the end of Stage B (microstep 3999). The four segmentation datasets use dataset-specific checkpoints or checkpoint ensembles from the same Stage-B run (Appendix H), so their rows evaluate the shared architecture and training recipe rather than a single set of weights. Thresholds, inference modes, and seed aggregation are chosen on validation data before the test run. Inference seeds measure generation variability and do not represent independent training runs. The September 22 export reports aggregate scores without per-dataset sample counts or confidence intervals; we therefore report point estimates and avoid statements of statistical significance.

F Mask-representation diagnostic

A completed validation-only diagnostic tests how much mask structure survives the frozen image representation. Ground-truth masks undergo the corresponding spatial transform, followed either by direct mask recovery or by posterior-mean VAE encoding and decoding before recovery. Table 8 uses thresholds selected within this diagnostic validation subset. DRIVE uses all four validation images and official FOVs; ISIC uses eight of its 90 validation images. These small-sample results support representational feasibility for the tested masks. They do not measure the final model's segmentation accuracy or establish a universal representation ceiling.

Table 8: **Validation mask round trips.** Native-size macro Dice after transform-only recovery and after posterior-mean VAE reconstruction. No medical model prediction is involved. Source: the completed September 8 diagnostic.

Dataset	Images	Transform only	VAE round trip
DRIVE	4	0.97636	0.96260
ISIC 2018 (subset)	8	0.99853	0.99849

G Reproducibility materials

The manuscript source includes a dated experiment manifest covering dataset counts, metric contracts, configuration locations, and final-result fields. Model implementation and data conversion remain in the parent repository. The shared model's architecture, stage configurations, and complete dataset contracts are described above. The anonymized release will contain the source code, dataset identifiers and usage conditions, the evaluated checkpoints, the fixed inference configurations, the result summaries behind every table, and the plotting script for Appendix I.

H Detailed results from the September 22 evaluation snapshot

The current results cover 14 datasets across six task families. The source label SenseNova (ours) denotes MedVision-X. T denotes a target-trained run, C a released checkpoint, and R a reimplementation. Dataset notes specify domain shifts, split overlap, adapters, and repairs. A dash denotes an unavailable score, including a blocked run; it never denotes zero. Bold marks the best reported value and underlining marks the second-best distinct value in each column, including ties. Rankings are numerical comparisons under the listed protocols, not tests of statistical significance or equal training and inference budgets. Baselines with a public paper record are cited as follows: U-Net (Ronneberger et al., 2015), nnU-Net v2 (Isensee et al., 2021), nnWNet (Zhou et al., 2025), MedCLIPSeg (Koleilat et al., 2026), MedSAM3 (Liu et al., 2025), PMRNet (Du et al., 2024), EMCAD (Rahman et al., 2024), SegMamba (Xing et al., 2024), Swin-UMamba (Liu et al., 2024), VISTA3D (He et al., 2024), HAT (Chen et al., 2023), DiffMSR (Li et al., 2024), TADSR (Zhang et al., 2025), HDW-SR (Yang et al., 2025), Swin2SR (Conde et al., 2022), Restormer (Zamir et al., 2022), WIA-LD2ND (Zhao et al., 2024), RED-CNN (Chen et al., 2017), SynDiff (Özbey et al., 2023), CycleGAN (Zhu et al., 2017), Pix2Pix (Isola et al., 2017), InstructPix2Pix (Brooks et al., 2023), MedEdit (Ben Alaya et al., 2024), InstructX2X (Min et al., 2026), and SegGuidedDiff (Konz et al., 2024).

Segmentation uses v3b step 1500 with D4 test-time augmentation on ISIC, a 1500+2000+3000+3500 checkpoint ensemble on Kvasir, a 2500+3000+3500 ensemble on BUSI, and step 1000 on DRIVE. The ten other datasets use checkpoint 3999 with the inference settings below. This snapshot therefore evaluates a shared architecture with dataset-specific checkpoints and aggregation. It does not constitute a single-checkpoint comparison.

The updated export supplies aggregate scores without evaluated-sample counts. The earlier snapshot reported partial coverage of 17/36 subjects on SynthRAD2023, 20/57 on IXI, 3,296/3,553 slices on LoDoPaB-CT, and 3,992/8,852 slices on fastMRI SR. Those counts describe the earlier runs; full or matched-cohort coverage of the updated scores is not established by the new export.

Normal FP is the false-positive pixel fraction on BUSI normal images. HD95 units and metric aggregation follow the dataset contracts in Appendix E. Body metrics use the body evaluation region; HU MAE is measured in Hounsfield units. Staining agreement is an auxiliary-classifier measure. ATLAS lesion-mask compositing preserves the source outside the mask, so its zero outside MAE and unit outside SSIM characterize the composited pipeline.

Table 9: ISIC 2018. Reported test results.

Model	Dice ↑	IoU ↑	HD95 ↓
MedCLIPSeg [C]	<u>0.9259</u>	<u>0.868</u>	<u>50.96</u>
MedSAM3 [C]	**0.9282**	**0.8731**	**44.71**
nnU-Net v2 [T]	0.9127	0.8517	224.5
nnWNet [T]	0.8872	0.8182	249.3
PMRNet [R]	0.9063	0.8435	70.78
U-Net [T]	0.9022	0.8366	86.64
MedVision-X	0.8918	0.8263	71.7

Ours: v3b step 1500, v2 with D4 test-time augmentation (8 views; prompt directions remapped). MedSAM3: official v1 LoRA checkpoint, text prompt selected on validation, possible dermoscopy training overlap; evaluated outside the sealed campaign. PMRNet: paper-guided reimplementation using the pinned official architecture, trained on the frozen target split.

Table 10: Kvasir-SEG. Reported test results.

Model	Dice ↑	IoU ↑	HD95 ↓
HGM [C]	–	–	–
MedCLIPSeg [C]	<u>0.9034</u>	0.8378	**38.08**
MedSAM3 [C]	0.8999	<u>0.8514</u>	43.87
nnU-Net v2 [T]	–	–	–
nnWNet [T]	0.9029	0.8498	43.8
U-Net [T]	0.8357	0.7602	73.66
MedVision-X	**0.9046**	**0.8543**	<u>42.56</u>

Ours: v3b checkpoint ensemble 1500+2000+3000+3500, v2 dense perception, official prompts, mean of 3 seeds. HGM: official checkpoint with canonical split mismatch and possible training overlap; scores unavailable. MedSAM3: official v1 LoRA checkpoint, text prompt selected on validation, possible endoscopy training overlap; evaluated outside the sealed campaign. nnU-Net v2 scores are unavailable in this snapshot.

Table 11: BUSI. Reported test results.

Model	Dice ↑	IoU ↑	HD95 ↓	Normal FP ↓
EMCAD [T]	0.6715	0.6039	63.2	7.27×10^{-3}
MedSAM3 [C]	0.6568	0.5918	77.61	0.0862
nnU-Net v2 [T]	<u>0.8097</u>	<u>0.7486</u>	<u>39.08</u>	3.58×10^{-3}
nnWNet [T]	0.7786	0.7147	42.91	3.14×10^{-3}
PMRNet [R]	0.7815	0.7148	49.33	5.20×10^{-3}
U-Net [T]	0.6449	0.5642	91.6	<u>3.10×10^{-3}</u>
MedVision-X	**0.8271**	**0.7652**	**38.29**	**0**

Ours: v3b checkpoint ensemble 2500+3000+3500, v2 dense perception, official prompts, mean of 3 seeds. MedSAM3: official v1 LoRA checkpoint, text prompt selected on validation, possible ultrasound training overlap; evaluated outside the sealed campaign. PMRNet: paper-guided reimplementation using the pinned official architecture, trained on the frozen target split.

Table 12: DRIVE. Reported test results.

Model	Dice ↑	IoU ↑	AUC ↑	Sens. ↑	Spec. ↑
MedCLIPSeg [T]	0.3809	0.2355	0.7513	0.3603	0.9234
MedSAM3 [C]	0.1057	0.05588	0.5395	0.06556	0.9752
nnU-Net v2 [T]	0.8236	0.7005	0.9787	0.8111	**0.9775**
nnWNet [T]	**0.8252**	**0.7027**	**0.9809**	**0.8172**	0.9768
PMRNet [R]	0.7339	0.5806	0.9427	0.686	0.9741
U-Net [T]	0.6463	0.4784	0.9188	0.7677	0.9117
MedVision-X	0.7911	0.6546	0.9606	0.7827	0.972

Ours: v3b step 1000, v2 dense perception, official prompts, mean of 3 seeds, threshold 0.5. MedCLIPSeg: target-trained. MedSAM3: official v1 LoRA checkpoint with validation-selected text prompt, evaluated outside the sealed campaign; its AUC is near chance, with an instance-detection versus thin-structure mismatch noted in the evaluation record. PMRNet: paper-guided reimplementation using the pinned official architecture, trained on the frozen target split.

Table 13: SynthRAD2023. Reported test results.

Model	Body HU MAE ↓	Body PSNR ↑	Body SSIM ↑
CRAVIM [T]	150.1	23.12	0.6
CycleGAN [T]	160.3	22.86	0.4188
DRDD [T]	149	25.41	0.4417
RegConDiS [T]	83.11	26.78	0.7604
SynDiff [T]	127.6	23.97	0.6435
MedVision-X	**55.2**	**30.99**	**0.8455**

Ours: checkpoint 3999, v2, CFG off, mean of 8 seeds with 20 steps each. CRAVIM, DRDD, and SynDiff: target-trained official architectures. RegConDiS: target-trained.

Table 14: IXI. Reported test results.

Model	PSNR ↑	SSIM ↑	MAE ↓	LPIPS ↓
CRAVIM [T]	22.3	0.7313	0.04368	**0.1353**
CycleGAN [T]	21.49	0.6905	0.05123	0.1522
DRDD [T]	22.43	0.301	0.04221	0.4637
MISA-LDM [R]	18.12	0.6761	0.06635	0.2905
SynDiff [C]	17.89	0.6841	0.09465	0.2535
MedVision-X	**25.58**	**0.8439**	**0.0291**	0.1667

Ours: checkpoint 3999, v2, CFG off, mean of 8 seeds with 20 steps each. CRAVIM and DRDD: target-trained official architectures. MISA-LDM: target-trained paper reimplementation.

Table 15: LoDoPaB-CT. Reported test results.

Model	RMSE ↓	PSNR ↑	SSIM ↑
Restormer [C]	0.02134	33.97	0.8124
VQ-SCD [R]	0.04479	27.04	0.6325
WIA-LD2ND [T]	0.03845	28.6	0.8114
MedVision-X	**0.01467**	**37.99**	**0.8774**

Ours: checkpoint 3999, v2, CFG off, mean of 8 seeds with 20 steps each, plus VAE residual. Restormer: official Gaussian-gray checkpoint, zero-shot out of domain. VQ-SCD: target-trained paper-guided reimplementation. WIA-LD2ND: target-trained paper-guided repair.

Table 16: LDCT projection. Reported test results.

Model	Body HU MAE ↓	Body RMSE ↓	Body PSNR ↑	Body SSIM ↑
Restormer [C]	110.2	142.1	35.51	0.6676
VQ-SCD [R]	239.5	270.8	23.96	0.03678
WIA-LD2ND [T]	99	122.9	33.56	0.5891
MedVision-X	**39.11**	**51.04**	**40.22**	**0.8542**

Ours: checkpoint 3999, v2, CFG off, mean of 8 seeds with 20 steps each. Restormer: official Gaussian-gray checkpoint, zero-shot out of domain. VQ-SCD: target-trained paper-guided reimplementation. WIA-LD2ND: target-trained paper-guided repair.

Table 17: fastMRI SR. Reported test results.

Model	NMSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
DiffMSR [T]	0.0128	<u>36.66</u>	0.9255	0.1736
GaussM2ASR [T]	**0.01217**	**36.76**	**0.9315**	<u>0.1718</u>
HAT [C]	<u>0.01263</u>	36.56	<u>0.9282</u>	0.1871
Swin2SR [C]	0.07125	30.16	0.6313	0.2132
MedVision-X	0.01485	35.86	0.9208	**0.1498**

Ours: checkpoint 3999, v2, CFG off, mean of 8 seeds with 20 steps each. DiffMSR and GaussM2ASR: target-trained single-contrast self-guide adapters. HAT and Swin2SR: released natural-image x4 checkpoints, zero-shot out of domain.

Table 18: SurgiSR4K. Reported test results.

Model	RGB PSNR ↑	RGB SSIM ↑	LPIPS ↓
HAT [C]	**31.83**	**0.7004**	<u>0.4946</u>
HDW-SR [R]	20.85	0.4188	0.699
Swin2SR [C]	<u>31.01</u>	0.6597	0.6044
TADSR [C]	28.79	0.6048	**0.4712**
MedVision-X	30.92	<u>0.6728</u>	0.5508

Ours: checkpoint 3999, v2, CFG off, mean of 8 seeds with 20 steps each, plus VAE residual. HAT, Swin2SR, and TADSR: released natural-image x4 checkpoints, zero-shot out of domain. HDW-SR: target-trained paper-guided reimplementation.

Table 19: BraTS editing. Reported test results.

Model	Edit Dice ↑	Out. MAE ↓	Out. SSIM ↑	FID ↓	KID ↓
ControlCG [R]	0.01874	0.1285	0.4314	32.87	0.02496
InstructPix2Pix [C]	**0.4849**	0.05865	0.7963	7.808	9.15×10^{-3}
InstructX2X [C]	0.09558	0.1273	0.6795	<u>5.729</u>	<u>6.33×10^{-3}</u>
MedEdit [T]	0.3499	<u>5.52×10^{-3}</u>	<u>0.9638</u>	8.06	0.01066
SegGuidedDiff [C]	0.06637	0.08861	0.2302	20.85	0.02032
MedVision-X	<u>0.4466</u>	$\mathbf{1.91 \times 10^{-3}}$	**0.9961**	**0.1417**	$\mathbf{7.49 \times 10^{-5}}$

Ours: checkpoint 3999, v1 protocol with v2 scorer. ControlCG: target-trained paper-guided reimplementation. InstructPix2Pix: author natural-RGB checkpoint, zero-shot out of domain on brain MRI. InstructX2X: author chest-X-ray v2 checkpoint, zero-shot out of domain on brain MRI, RSE without anatomical pseudomask. MedEdit: target-trained, validation-selected. SegGuidedDiff: author mask-ablated breast-MRI checkpoint, cross-anatomy out of domain, ROI-guided.

Table 20: ATLAS v2. Results with lesion-mask compositing.

Model	Edit Dice ↑	Out. MAE ↓	Out. SSIM ↑	FID ↓	KID ↓
ControlCG [R]	0.02125	0.1163	0.4237	28.11	0.01947
InstructPix2Pix [C]	0	0.0416	0.8435	<u>21.99</u>	9.48×10^{-3}
InstructX2X [C]	0.08336	0.1786	0.2756	90.31	0.1071
MedEdit [C]	**0.7428**	<u>3.28×10^{-3}</u>	<u>0.993</u>	**21.23**	$\mathbf{7.29 \times 10^{-3}}$
SegGuidedDiff [C]	0.05213	0.2292	0.1899	92.37	0.1161
MedVision-X	<u>0.5936</u>	**0**	**1**	23.91	<u>7.54×10^{-3}</u>

Ours: checkpoint 3999, v2, CFG off, mean of 3 seeds, lesion-mask composite. ControlCG: target-trained paper-guided reimplementation. InstructPix2Pix: author natural-RGB checkpoint, zero-shot out of domain on brain MRI. InstructX2X: author chest-X-ray v2 checkpoint, zero-shot out of domain on brain MRI, RSE without anatomical pseudomask. MedEdit: official ATLAS checkpoint on the canonical frozen test. SegGuidedDiff: author mask-ablated breast-MRI checkpoint, cross-anatomy out of domain, ROI-guided.

Table 21: MIST. Reported test results.

Model	FID ↓	KID ↓	LPIPS ↓	SSIM ↑	Marker agr. ↑
ConfusionGAN [R] (blocked)	–	–	–	–	–
PD-UNIST [R] (blocked)	–	–	–	–	–
Pix2Pix [T]	288.4	0.2499	0.6913	**0.3855**	0.3392
TDKStain [T]	94.11	0.06773	0.5516	0.294	0.6452
VSMT [C]	44.88	0.01347	**0.5434**	0.2479	0.6447
MedVision-X	**31.66**	$\mathbf{3.32 \times 10^{-3}}$	0.5484	0.2735	**0.6476**

Ours: checkpoint 3999, v2, CFG off, Stage-A marker-specific prompt. Pix2Pix, TDKStain, and VSMT: official MIST patch split with known whole-slide overlap.

Table 22: BCI. Reported test results.

Model	FID ↓	KID ↓	LPIPS ↓	SSIM ↑	HER2 agr. ↑
ConfusionGAN [R] (blocked)	–	–	–	–	–
PD-UNIST [R] (blocked)	–	–	–	–	–
Pix2Pix [T]	258.5	0.1993	0.5839	**0.5712**	0.3367
TDKStain [T]	140.1	0.09358	0.5317	0.5586	0.3695
VSMT [C]	47.27	9.87×10^{-3}	**0.4867**	0.5464	**0.4442**
MedVision-X	**39.12**	$\mathbf{6.24 \times 10^{-3}}$	0.5109	0.5273	0.4084

Ours: checkpoint 3999, v2, CFG off. Pix2Pix, TDKStain, and VSMT: BCI release patch-only split; whole-slide identities unavailable.

H.1 Earlier results outside the updated snapshot

BraTS-GLI and AAPM Mayo LDCT remain in the 16-dataset training registry but have no entries in the September 22 export. Their earlier results are retained here with their original checkpoints and protocols, separately from the 14-dataset current comparison. For BraTS-GLI, D and H denote Dice and HD95 for the indicated regions.

Table 23: BraTS-GLI. Earlier snapshot; not updated in the September 22 results.

Model	D-WT ↑	D-TC ↑	D-ET ↑	D-SNFH ↑	D-NETC ↑	D-RC ↑	H-WT ↓
nnU-Net v2 [T]	**0.8802**	0.8033	0.8142	**0.8559**	0.7211	**0.6105**	**4.979**
SegMamba [T]	0.8736	**0.8090**	**0.8173**	0.8469	**0.7362**	0.5917	6.881
Swin-UMamba [T]	0.6496	0.6851	0.6967	0.6012	0.6220	0.3142	60.760
VISTA3D [C]	0.4555	–	–	–	–	–	26.930
MedVision-X	0.6977	0.5383	0.5501	0.6506	0.1785	0.3100	43.490

Ours: v3b step 1500 (interim), multi-parametric RGB input, one seed. SegMamba and Swin-UMamba: target-trained. VISTA3D: official NV-Segment CT/MR checkpoint, BraTS23 class 176, T1c, WT labels 1/2/3 with cavity label 4 excluded; partial metric contract.

Table 24: AAPM Mayo LDCT. Earlier snapshot; not updated in the September 22 results.

Model	HU MAE ↓	RMSE ↓	PSNR ↑	SSIM ↑
RED-CNN [C]	–	–	–	–
Restormer [C]	38.36	49.34	38.43	0.87890
Text-guided unified [R]	–	–	–	–
VQ-SCD [R]	205.90	212.00	25.72	0.05187
WIA-LD2ND [T]	**22.81**	**29.97**	**42.73**	**0.94720**
MedVision-X	24.79	32.28	42.09	0.93780

Ours: checkpoint 3999, v2, CFG off, mean of 8 seeds with 20 steps each. RED-CNN: released LDCTBench checkpoint, cross-release patient overlap unknown. Restormer: released Gaussian-gray checkpoint, zero-shot out of domain. Text-guided unified and VQ-SCD: target-trained paper-guided reimplementations. WIA-LD2ND: target-trained paper-guided repair.

I VISUAL ANALYSIS OF THE SEPTEMBER 22 RESULTS

All plots in this section are drawn directly from the per-dataset tables in Appendix H (and, for Figure 6, from the earlier BraTS-GLI snapshot in Appendix H.1). No value is interpolated or estimated. The plotting source that recomputes every rank from those tables is included in the supplementary material. Rankings and distances are numerical comparisons under the listed protocols; they are not significance tests.

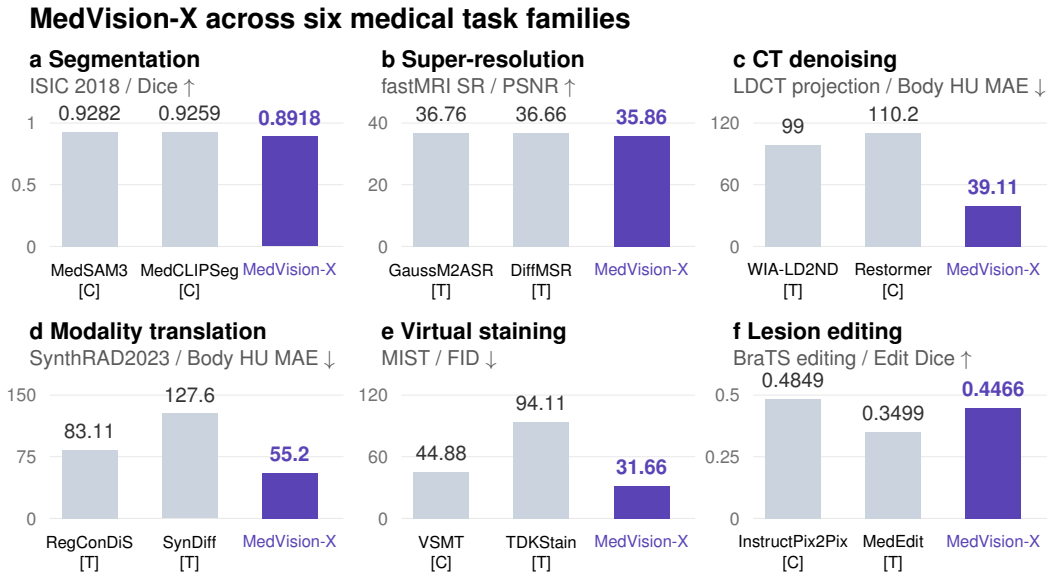

Figure 5: **Benchmark overview across six medical task families.** Each panel shows one primary endpoint for MedVision-X and the two strongest scored baselines on the same test split; axes start at zero and arrows give the preferred direction. [T] denotes a baseline trained on the benchmark training split and [C] a released checkpoint. ISIC uses a dataset-specific segmentation checkpoint with test-time augmentation; the other five panels use one shared checkpoint. Baseline provenance, including domain shift and possible training overlap, is listed in Appendix H. Layout follows Han et al. (2026).

	WT	TC	ET	SNFH	NETC	RC
nnU-Net v2	0.880	0.803	0.814	0.856	0.721	0.611
SegMamba	0.874	0.809	0.817	0.847	0.736	0.592
Swin-UMamba	0.650	0.685	0.697	0.601	0.622	0.314
VISTA3D	0.456	–	–	–	–	–
MedVision-X	0.698	0.538	0.550	0.651	0.178	0.310

Figure 6: **Earlier BraTS-GLI segmentation by region.** Dice for each of the six regions of the 2024 post-treatment label convention (Appendix E), from the earlier snapshot in Table 23; BraTS-GLI is absent from the September 22 export. The color scale is fixed from zero to one, and dashes mark regions that the WT-only VISTA3D contract does not score. MedVision-X uses an interim single-seed checkpoint here. Its largest gap is on the non-enhancing tumor core (NETC).

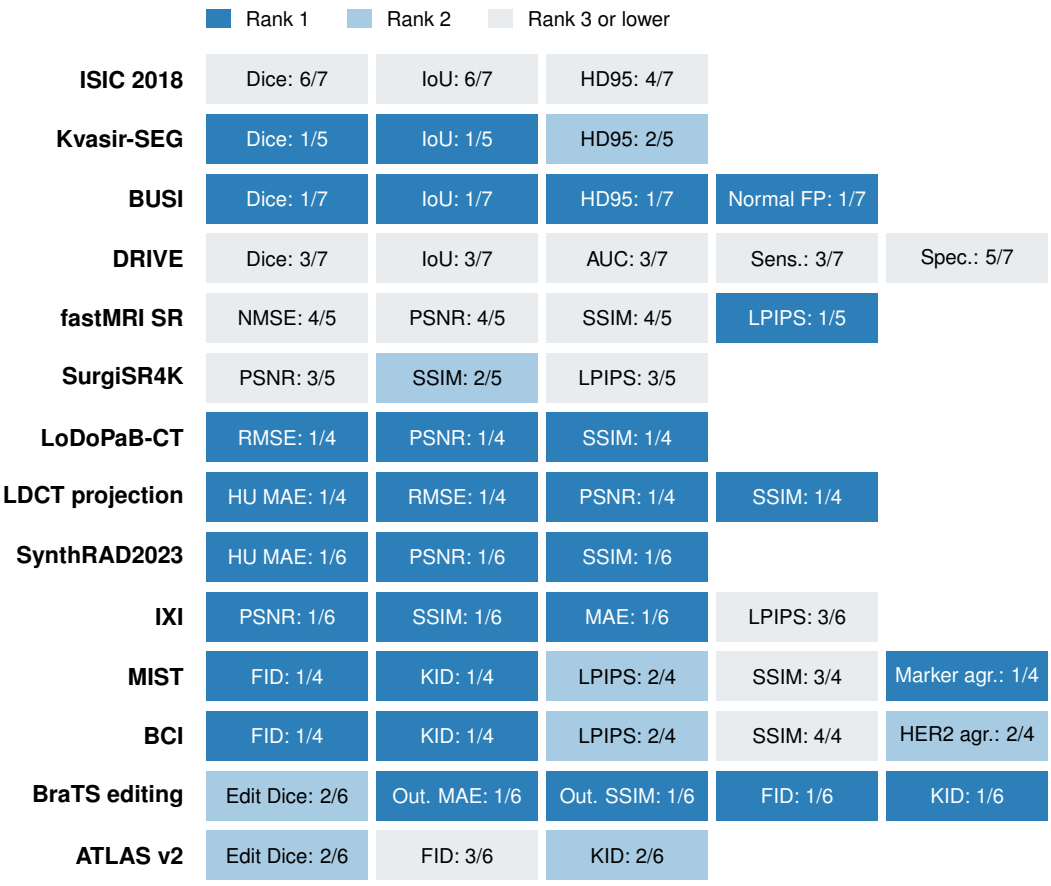

Figure 7: **Rank of MedVision-X on every reported metric.** Each cell gives the rank of MedVision-X among all methods with a score for that metric (rank/number of scored methods), with the preferred direction of each metric applied. Across 54 dataset–metric pairs, MedVision-X ranks first on 29 and in the top two on 37. It ranks first on the primary endpoint of 8 of the 14 datasets and second on two more (BraTS and ATLAS editing). ATLAS outside-region metrics are omitted because lesion-mask compositing makes them exact by construction. Segmentation rows use dataset-specific checkpoints or ensembles; all other rows use checkpoint 3999.

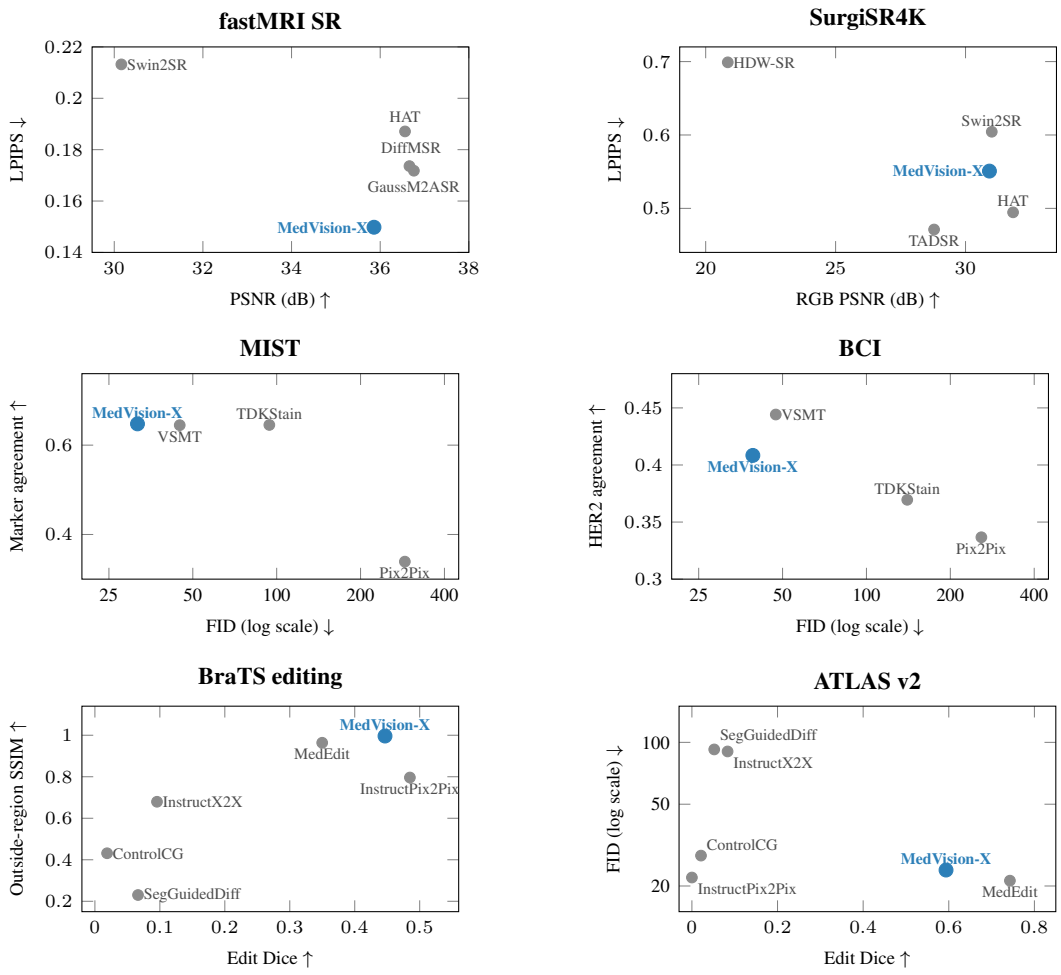

Figure 8: **Tradeoffs that a single endpoint hides.** Each panel plots two reported metrics for every scored method; MedVision-X is shown in blue. **Top:** super-resolution separates pixel fidelity from perceptual quality. MedVision-X has the best LPIPS on fastMRI but lower PSNR than three target-trained or released specialists; on SurgiSR4K it lies between HAT and TADSR on both axes. **Middle:** virtual staining separates distribution realism from classifier-based marker agreement. MedVision-X has the lowest FID on both datasets; its MIST agreement lead is 0.0024, and on BCI VSMT has higher HER2 agreement. **Bottom:** lesion editing separates realization of the requested lesion from preservation and realism. On BraTS, MedVision-X combines the second-highest Edit Dice with the highest outside-region SSIM; on ATLAS, MedEdit is better on both Edit Dice and FID. ATLAS outside-region metrics are omitted because MedVision-X uses lesion-mask compositing there.