DUALBRIDGE: WHAT TO TRANSMIT BETWEEN HETEROGENEOUS FOUNDATION MODELS FOR DOMAIN-GENERALIZED MEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-source domain generalization requires medical image segmentation models to generalize well using supervision from a single source domain. Recently, vision foundation models (VFMs) have attracted growing interest for their transferable knowledge acquired through large-scale pretraining. Self-supervised VFMs such as DINOv3 provide broadly transferable representations, whereas promptable VFMs such as SAM 3 offer segmentation-specific capabilities. However, bridging these complementary strengths is nontrivial: effective integration requires deciding what information to transmit and where to inject it. Prompt-only interfaces can discard dense semantics, while unrestricted feature fusion can increase the adaptation burden under limited medical supervision. To balance these competing demands, we propose DualBridge, a scale-selective framework that connects DINOv3 and SAM 3 through two complementary information-transfer pathways. The spatial pathway converts DINOv3-derived predictions into explicit spatial and shape prompts for SAM 3. The feature pathway complements this compact interface by selectively transferring dense semantic features. Specifically, lightweight adapters project intermediate DINOv3 features into SAM 3's high-level representations through gated residual connections that regulate their contribution. Together, these pathways combine explicit spatial guidance with selective feature transfer, without access to target-domain data during training. Experiments on cross-domain retinal fundus and polyp segmentation benchmarks demonstrate that DualBridge outperforms state-of-the-art methods under the single-source protocol. These results support selective information transfer as an effective strategy for harnessing complementary foundation models under unseen domain shifts.

1 INTRODUCTION

Medical image segmentation is a key component of disease diagnosis, treatment planning, and quantitative analysis (Pham et al., 2000). Although deep learning methods have achieved strong performance across many segmentation tasks, their reliability can deteriorate under cross-institutional domain shifts (Wei et al., 2024a). Differences in imaging devices, acquisition protocols, and preprocessing affect not only image appearance but also tissue contrast, noise characteristics, and boundary visibility (Yoon et al., 2024). Models trained on one institution's data may therefore fail to generalize to unseen clinical environments. Test-time adaptation addresses this challenge by updating models using unlabeled target-domain samples (Wang et al., 2021; 2022), but introduces additional computational overhead and requires target-data access for adaptation during deployment. These requirements motivate single-source domain generalization (SSDG) (Su et al., 2023; Ouyang et al., 2022), particularly in settings that demand cross-institutional generalization while restricting target-data use for privacy reasons. In SSDG, models are trained exclusively on labeled data from a single source domain and must segment images from unseen domains without target-specific adaptation.

Recently, vision foundation models (VFMs) have attracted growing interest for their transferable knowledge acquired through large-scale pretraining (Awais et al., 2025). Self-supervised models such as DINOv3 provide broadly transferable dense representations with strong regional coherence (Siméoni et al., 2026). However, these representations require task-specific decoding to identify and delineate the structures of interest. Promptable models such as SAM 3 offer complemen-

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.

tary segmentation capabilities, using spatial prompts to guide fine-grained mask prediction (Carion et al., 2026). These strengths have motivated efforts to combine self-supervised representations with promptable segmentation. For example, Matcher (Liu et al., 2024) and ProtoSAM (Ayzenberg et al., 2025) use DINO representations to derive spatial prompts for SAM in one-shot or few-shot segmentation settings. Other approaches investigate feature fusion or cross-model distillation (Xiong et al., 2025; Zhuo et al., 2026). Together, these studies demonstrate the potential of integrating heterogeneous VFMs through different forms of knowledge transfer.

For SSDG, the integration mechanism itself need to generalize well. It is learned using only source-domain supervision but deployed under unseen domain shifts without target-specific adaptation. Transferable foundation-model representations alone do not guarantee that this mechanism will exploit them reliably under unseen domain shifts. The challenge is therefore not merely to connect complementary models, but to transfer their knowledge without over-relying on source-specific cues. This motivates selective cross-model information transfer, guided by two questions: *what information should be transmitted, and where should it be injected?*

The information bottleneck principle provides a way to reason about how much task-relevant information an interface can convey (Tishby et al., 2001). A binary mask prompt carries compact spatial guidance, while a limited feature channel can carry complementary regional semantics. This view motivates two pathways with distinct capacities and destinations within SAM 3.

Accordingly, we propose **DualBridge**, a scale-selective framework for single-source domain-generalized medical image segmentation. The *spatial pathway* converts DINOv3-derived predictions into localization and shape prompts for SAM 3, establishing a compact task-specific interface. The *feature pathway* transfers DINOv3 features to selected high-level representations of SAM 3 through gated adapters. The two pathways differ in the information they carry and where it enters SAM 3, while its native high-resolution pathway remains intact. This design uses a small trainable interface under source-only medical supervision.

Our contributions are summarized as follows:

- We use an information-bottleneck perspective to motivate a compact spatial prompt and a restricted feature channel for integrating heterogeneous vision foundation models under source-only supervision.
- We propose DualBridge, an asymmetric, scale-selective framework that bridges self-supervised and promptable VFMs through compact spatial prompts and selective semantic transfer, enriching high-level semantics while preserving the promptable model's native fine-grained spatial processing.
- Experiments on five retinal fundus datasets and four polyp segmentation datasets demonstrate improvements over state-of-the-art methods on unseen domains, highlighting selective information transfer as a promising strategy for exploiting complementary foundation-model knowledge under limited source-domain supervision.

2 Related Work

2.1 Test-Time Adaptation and Single-Source Domain Generalization

Cross-institutional changes in acquisition and image appearance can degrade medical-image segmentation (Zhang et al., 2025). Test-time adaptation (TTA) responds using unlabeled target images: TENT minimizes prediction entropy (Wang et al., 2021), while CoTTA uses teacher predictions and weight restoration to adapt across changing domains (Wang et al., 2022). These methods update a model during deployment and can be sensitive to the quality and order of target samples (Niu et al., 2026). Single-source domain generalization (SSDG) instead learns from one labeled source and is evaluated on unseen domains without target-specific updates. Earlier medical SSDG methods diversify source appearances or seek more stable representations through augmentation and causal modeling (Su et al., 2023; Ouyang et al., 2022). Recent methods draw on pretrained models: Rein refines frozen VFM features with lightweight trainable modules (Wei et al., 2024b), Causal-Tune uses causal guidance for parameter-efficient adaptation (Zhang et al., 2026), and MFuser combines complementary pretrained representations (Zhang & Tan, 2025). SAM-based approaches, including DAPSAM, use domain-adaptive prompts to improve segmentation on unseen domains (Wei et al.,

2024a). Our setting follows SSDG; the distinction lies in how information from a self-supervised encoder reaches a promptable segmenter under source-only supervision.

2.2 Self-Supervised and Promptable Vision Foundation Models

Self-supervised encoders such as DINOv2 and DINOv3 provide dense features with useful object and region structure (Oquab et al., 2023; Siméoni et al., 2026). However, their features still need a task-specific mechanism to identify the medical structure to segment. In parallel, the SAM family predicts masks conditioned on prompts such as points, boxes, and masks (Kirillov et al., 2023; Ravi et al., 2025; Carion et al., 2026). Pretraining gives these models different capabilities, but neither guarantees robustness to every clinical domain shift. DualBridge uses DINOv3 to provide task-specific spatial guidance and complementary features, and uses SAM 3 for prompted mask prediction. The interface between them must therefore be learned from one source while remaining useful on unseen images.

2.3 Combining Heterogeneous Foundation Models

Prior work connects DINO-like encoders and SAM mainly through prompts or feature transfer. Matcher and ProtoSAM derive prompts from self-supervised features for one-shot or few-shot segmentation (Liu et al., 2024; Ayzenberg et al., 2025); related work also studies automatic grounding and prompt generation (Rasaee et al., 2025; Jiang et al., 2026). Such interfaces pass localization to SAM, but a discrete spatial prompt does not expose the full dense representation or its continuous prediction scores. Other approaches combine representations through feature fusion or distillation, as in SAM2-UNeXT and UINO-FSS (Xiong et al., 2025; Zhuo et al., 2026). Feature transfer can preserve richer information, but adds an integration mechanism whose behavior under unseen domains depends on how it is trained. DualBridge combines the two routes for SSDG: a DINOv3-derived mask enters SAM 3's prompt interface, while gated residual adapters pass selected DINOv3 features to high-level SAM 3 features. The adapters leave the finest feature level on SAM 3's native path, separating spatial guidance from semantic transfer without requiring target-domain adaptation.

3 Method

3.1 Problem Setting

We consider single-source domain generalization (SSDG) for medical image segmentation. Each run trains on labeled images from one source domain, $\mathcal{D}_s = \{(x_i, y_i)\}_{i=1}^{n_s}$, and is evaluated on unseen target domains $\{\mathcal{D}_{t_k}\}_{k=1}^{K}$. Target-domain images, labels, and statistics are unavailable during training, and at inference each target image is processed independently, without target-specific adaptation. Each foreground class is treated as a binary segmentation task.

We build on two frozen vision foundation models with complementary properties: a self-supervised encoder E_ψ (DINOv3), whose patch tokens are broadly transferable but carry no notion of the target classes, and a promptable segmenter S_θ (SAM 3), which produces masks but must be told what to segment. Neither alone solves the task, and under SSDG the mechanism that couples them is itself learned from a single domain: any component added to the bridge is fitted to source-specific appearance and must nevertheless behave sensibly on unseen centres. The design problem is therefore not which models to use, but how to connect them under a constrained supervision budget—*what information the bridge should carry*, and *where that information should enter the promptable model*. Section 3.2 analyses these two questions and derives the requirements they impose; Section 3.3 presents the architecture that meets them.

3.2 Designing the Bridge: An Information-Bottleneck View

The information bottleneck principle motivates preserving task-relevant information while restricting representation capacity (Tishby et al., 2001). We apply this perspective to the interface between DINOv3 and SAM 3. Let D denote DINOv3 features, Y the target mask, $m = h(D)$ the spatial prompt, and $R = r_\eta(D)$ the transferred semantic features. Our analysis assumes fixed model parameters and a fixed domain distribution.

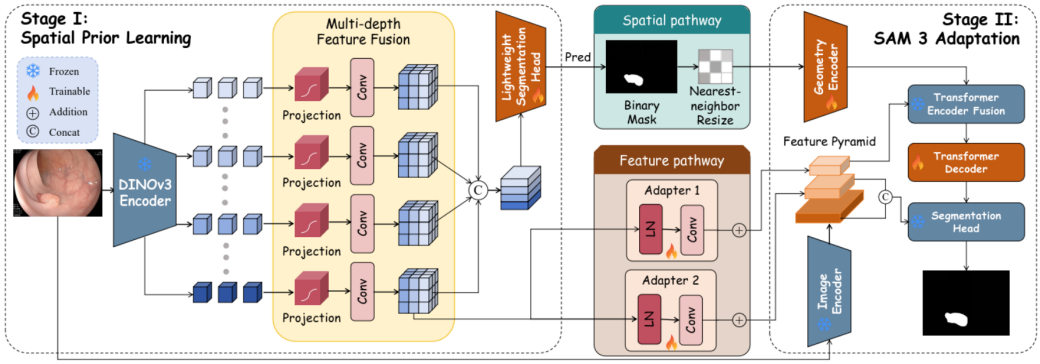

Figure 1: **Overview of our proposed method DualBridge.** The frozen DINOv3 encoder supplies multi-depth features to a source-trained segmentation head and its deepest patch tokens to two SAM 3 feature adapters. The thresholded head output enters SAM 3 as a binary mask prompt; the projected tokens enter the two deepest feature-pyramid levels as gated residuals, while the finest level remains on SAM 3's native pathway.

Spatial prompts provide compact but restricted information. A binary prompt with N entries has an entropy of at most N bits. The data-processing inequality further gives

$$I(Y; m) \leq \min\{I(Y; D), H(m)\}, \qquad H(m) \leq N. \tag{1}$$

Thus, spatial prompts provide a compact task-specific interface, but may omit useful information available in dense features and continuous predictions (Cover & Thomas, 2006).

Complementary information in the feature channel. For a feature channel containing d_R entries at b-bit precision, its representation budget is $C_R = b\, d_R$. Its contribution beyond the prompt satisfies

$$0 \leq I(Y; R \mid m) \leq \min\{I(Y; D) - I(Y; m),\, C_R\}. \tag{2}$$

The bound characterizes the information available beyond the prompt and the channel's representation budget. Under limited source supervision, the complexity of the trainable integration mechanism is also relevant to learning (Bartlett & Mendelson, 2002).

Implication for DualBridge. We retain the compact spatial interface and supplement it with a restricted semantic channel. Spatial prompts enter SAM 3's mask prompt encoder, while selected semantic features enter high-level representations through gated residual connections. The bounds motivate the form and capacity of this interface; the cross-domain experiments and ablations assess its segmentation behavior. Proofs and a discussion of learning complexity under limited source supervision appear in Appendix A.

3.3 Overview of DualBridge

DualBridge adds two small trainable components to the frozen E_ψ and S_θ (Figure 1): a lightweight segmentation head G_ϕ that, followed by thresholding, implements h, and two gated feature adapters with parameters η that implement r_η. The DINOv3 encoder is run once per image and its patch tokens feed both pathways. The *spatial pathway* turns the head prediction into a binary mask prompt for SAM 3's mask prompt encoder. The *feature pathway* projects the deepest DINOv3 tokens into the two deepest levels of SAM 3's feature pyramid as gated residuals, leaving the finest level untouched. SAM 3 then decodes the prompt and the augmented pyramid into the final mask.

Training proceeds in two stages on source labels only. Stage 1 (Section 3.4) trains G_ϕ with E_ψ frozen. Stage 2 (Section 3.5) freezes E_ψ and G_ϕ and trains the two feature adapters, SAM 3's geometry encoder, and low-rank updates to SAM 3's transformer encoder and decoder, with the SAM 3 image backbone frozen.

3.4 Stage 1: Spatial Prior from Multi-Depth DINOv3 Features

The spatial pathway requires a prompt that localizes the target structure, which the frozen encoder cannot provide on its own. Features at different depths of a vision transformer play different roles (Raghu et al., 2021; Amir et al., 2021): shallow blocks mainly encode edges, local texture, and fine detail; intermediate blocks encode regions, shapes, and part relations; and deep blocks encode global semantics and category information. The patch tokens of self-supervised DINO-family models exhibit strong dense correspondence and object-level structure (Caron et al., 2021; Siméoni et al., 2026), but their pretrained semantics are learned from natural images and thus misaligned with medical class boundaries. The frozen encoder therefore requires a trained head, and that head should not rely on the final semantic layer alone. Inspired by Li et al. (2025), the head fuses patch features from four depths so that the boundary and shape cues of shallower blocks complement the semantics of deeper ones.

We write $D^{(i)}$ for the patch tokens output by the i-th block of the frozen encoder and tap four blocks $\mathcal{I} = (i_1, \dots, i_4)$ spread over its depth, with i_4 the deepest tap. Each tap is projected by a small convolutional branch Π_k and resampled to a common grid, and the four branches are aggregated by channel-wise concatenation, so that each depth retains its own channels:

$$A(x) = \left[\Pi_1\left(D^{(i_1)}\right); \dots; \Pi_4\left(D^{(i_4)}\right)\right], \qquad z_\phi(x) = \mathrm{Up}(W_{\mathrm{out}} * A(x)), \tag{3}$$

where W_{out} is a 1×1 convolution producing one logit map per foreground class and Up resamples them to the input resolution. Tap indices and branch widths are given in Appendix B. Each class is supervised independently with the segmentation loss

$$\mathcal{L}_{\mathrm{seg}}(z, y) = \mathcal{L}_{\mathrm{BCE}}(z, y) + \mathcal{L}_{\mathrm{Dice}}(\sigma(z), y), \tag{4}$$

i.e., the sum of binary cross-entropy and soft Dice losses, where σ is the sigmoid function and y is the binary label of that class. The checkpoint that performs best on the source validation split is then frozen, and Stage 2 derives the spatial prompt from it by thresholding the class probability map,

$$m(x) = \mathbf{1}[\sigma(z_\phi(x)) > \tau], \qquad \tau = 0.5, \tag{5}$$

using the same threshold in training and inference.

3.5 Stage 2: Dual-Pathway Transfer into SAM 3

Stage 2 trains only the interfaces at which the two signals enter SAM 3.

Spatial pathway. The binary prompt $m(x)$ of equation 5 is resampled from the head grid to the SAM 3 input grid by nearest-neighbour interpolation, which keeps it binary, and is embedded by SAM 3's pretrained mask prompt encoder. Since the prompt is a coarse, source-trained prior, the geometry encoder is left trainable so that source supervision can calibrate how strongly SAM 3 relies on it.

Feature pathway. Let $\{F_l(x)\}_{l=1}^{3}$ be the three-level feature pyramid of the frozen SAM 3 image backbone, ordered from the finest to the deepest level, with C_l channels and spatial size $H_l \times W_l$. SAM 3 fuses only the deepest level into its transformer encoder, so a residual there reaches both the encoder and the mask-prediction head, whereas a residual at the intermediate level reaches the mask-prediction head alone; the finest level carries the high-resolution detail on which mask sharpness depends. We therefore inject at $\mathcal{L}_{\mathrm{sem}} = \{2, 3\}$ and leave F_1 untouched. The semantic channel is built from the patch tokens $D^{(i_4)}(x)$ of the deepest Stage-1 tap, which are already computed for the head, through an independent adapter per level:

$$R_l(x) = \mathrm{Resize}_{H_l \times W_l}\left[W_l\ \mathrm{LN}\left(D^{(i_4)}(x)\right) + b_l\right], \tag{6}$$

$$\widetilde{F}_l(x) = F_l(x) + \sigma(\gamma_l) \odot R_l(x), \qquad l \in \mathcal{L}_{\mathrm{sem}}, \tag{7}$$

and $\widetilde{F}_1 = F_1$. Here LN is layer normalization over the DINOv3 channel dimension, (W_l, b_l) is a trainable 1×1 convolution from the DINOv3 width to C_l, Resize is bilinear interpolation to the pyramid grid, $\gamma_l \in \mathbb{R}^{C_l}$ is a per-channel gate, and $\odot$ broadcasts over spatial positions; $\eta = \{W_l, b_l, \gamma_l\}_{l \in \mathcal{L}_{\mathrm{sem}}}$.

Layer normalization is needed because D and F_l come from different pretraining objectives and differ in activation scale by roughly an order of magnitude. The 1×1 projection is the only mixing the channel performs, so the transferred content is the DINOv3 representation itself rather than a new source-trained feature extractor, and the per-channel gate lets the model open the channel on some feature dimensions while keeping others at their native values. We initialize γ_l to a large negative value, so that $\sigma(\gamma_l) \approx 0$ and training starts close to SAM 3's native pathway, but keep the default initialization of W_l: a zero W_l would make $R_l \equiv 0$ and remove the gradient of γ_l at the first step, whereas the small gate alone already yields the near-identity start.

Training and inference. SAM3 processes the embedded prompt together with $\{\widetilde{F}_l(x)\}$ and produces candidate mask logits with scores; the highest-scoring candidate $q_\theta(x)$ is trained with $\mathcal{L}_{\text{seg}}(q_\theta(x), y)$ of equation 4. Multiple foreground classes are handled as independent binary problems, each with its own prompt and SAM3 forward pass, while R_l is class-agnostic and computed once per image. The DINOv3 encoder, the Stage-1 head, and the SAM3 image backbone stay frozen; the trainable set consists of the two adapters η, SAM3's geometry encoder (including the mask prompt branch), and low-rank (LoRA) updates (Hu et al., 2021) to the linear layers of SAM3's transformer encoder and decoder, whose base weights are fixed. Inference follows the same computation path: one DINOv3 forward pass yields the prompt and $\{R_l\}$, and one prompted SAM 3 forward pass yields the final logits, whose highest-scoring candidate is resized to the original image size and thresholded at 0.5. No parameter is updated at test time.

4 Experiments

4.1 Datasets and Implementation Details

The fundus segmentation benchmark follows the five-domain setting of VPTTA (Chen et al., 2024): A (RIM-ONE-r3 (Fumero et al., 2011)), B (REFUGE training set (Orlando et al., 2020)), C (ORIGA (Zhang et al., 2010)), D (REFUGE validation/test sets (Orlando et al., 2020)), and E (Drishti-GS (Sivaswamy et al., 2014)), containing 159, 400, 650, 800, and 101 images, respectively. B and D are separate partitions of the same REFUGE challenge. Each image is annotated with the optic disc (OD) and optic cup (OC).

The polyp segmentation dataset consists of four public datasets collected from different medical centres, denoted as domain A (BKAI-IGH-NeoPolyp (Ngoc Lan et al., 2021)), B (CVC-ClinicDB (Bernal et al., 2015)), C (ETIS-LaribPolypDB (Silva et al., 2014)), and D (Kvasir-Seg (Jha et al., 2020)), which contain 1000, 612, 196, and 1000 images, respectively.

We report the Dice similarity coefficient (DSC) for polyp segmentation and the mean of optic-disc and optic-cup DSC for fundus segmentation. Scores are multiplied by 100 in all tables. For each source domain, we report the value averaged over all its target-domain images; the overall average is then computed by averaging across source domains with equal weight.

Implementation details. We train one model per source domain; target images and labels are used neither for training nor for model selection. All experiments are conducted on a single NVIDIA L40 GPU (46 GB memory). DINOv3 ViT-L/16 (256×256 input) and SAM3 (1008×1008 input) stay frozen. Stage 1 trains only the segmentation head, taking taps at every sixth block; its source-selected checkpoint is then frozen, and Stage 2 trains the mask geometry encoder, rank-8 transformer LoRA modules, and the two gated adapters, which inject into the two deepest SAM3 pyramid levels with their gates initialized to $\sigma(\gamma_l) \approx 0.007$. Both stages use binary cross-entropy plus soft Dice, AdamW with base learning rate 10^{-4} and weight decay 10^{-4}, batch size 1, gradient clipping at 1.0, and at most 20 epochs; the adapters use a $30\times$ learning-rate multiplier, and the prompt threshold is $\tau = 0.5$ in training and inference.

4.2 Comparison with State-of-the-Arts

We compare DualBridge with test-time adaptation (TTA), conventional domain generalization (DG), and VFM-based DG methods in Table 1; Appendix C describes the baselines. TTA results are taken from Chen et al. (2024); DG and VFM-based methods are rerun with the authors' code under our configuration. Each column A–E reports a model trained on that source and evaluated on the

Table 1: Cross-domain segmentation results. Each letter identifies the training source; the score averages over images from all other domains. Avg. is the unweighted mean across sources. Fundus scores average optic-disc and optic-cup Dice. Best and second-best values in each column are **bold** and underlined, respectively.

Method	Venue	Polyp					Fundus					
		A	B	C	D	Avg.	A	B	C	D	E	Avg.
TTA Method.												
TENT	ICLR'21	74.86	67.51	17.79	73.55	58.43	73.07	78.66	71.94	46.81	70.20	68.13
CoTTA	CVPR'22	76.46	66.77	71.39	70.71	71.33	75.39	75.98	69.14	53.99	70.40	68.98
DLTTA	TMI'22	76.27	66.58	63.72	71.20	69.44	75.11	78.85	73.89	51.64	69.71	69.84
DUA	CVPR'22	78.93	66.84	76.53	86.24	77.13	72.28	76.59	70.13	56.17	71.38	69.31
SAR	ICLR'23	76.48	66.45	71.46	70.41	71.20	74.55	77.71	70.78	55.40	71.72	70.03
DomainAdaptor	CVPR'23	77.48	70.82	71.96	76.89	74.29	74.50	76.39	71.81	56.78	70.55	70.01
VPTTA	CVPR'24	81.00	76.87	77.58	86.39	80.46	73.91	79.36	74.51	56.51	75.35	71.93
DG Method.												
CSDG	TMI'22	79.55	69.93	61.71	77.69	72.22	80.36	84.50	77.75	83.12	55.39	76.22
SLAug	AAAI'23	80.24	71.39	68.34	78.31	74.57	72.49	67.62	76.96	65.09	64.30	69.29
Mamba-Sea	TMI'25	70.40	65.80	64.55	72.18	68.23	82.59	86.52	85.68	78.76	81.91	83.09
VFM Based DG Method.												
SAM-Med	Arxiv'24	53.14	59.48	52.51	62.13	56.81	59.44	58.52	71.87	47.33	63.76	60.18
DAPSAM	MICCAI'24	78.97	71.68	74.00	77.98	75.66	81.82	83.97	**87.41**	86.12	82.74	84.41
Rein	CVPR'24	87.97	81.60	78.02	84.22	82.95	79.24	85.45	84.14	86.00	80.55	83.08
MFuser	CVPR'25	88.46	83.66	81.74	85.03	84.72	84.54	85.02	83.18	84.27	81.54	83.71
SAM2-UNeXT	Arxiv'25	88.46	82.57	80.85	82.97	83.71	84.24	85.55	75.32	78.17	83.85	81.43
Causal-Tune	AAAI'26	87.10	79.38	78.20	82.23	81.73	78.55	**86.57**	82.19	84.94	84.37	83.32
DualBridge (Ours)	–	**89.22**	**86.12**	**87.23**	**87.92**	**87.62**	**84.66**	86.39	86.89	**86.93**	**85.10**	**86.00**

remaining domains. TTA methods additionally use unlabeled target images to update the model at test time. Appendix G reports a separate DINOv3-L backbone replacement control for Rein and Causal-Tune.

Main results. DualBridge obtains the highest source-averaged score on both tasks: 87.62 on polyp and 86.00 on fundus. These exceed the strongest baselines, MFuser (84.72) and DAPSAM (84.41), by 2.90 and 1.59 points, respectively. On polyp, DualBridge leads for every source; its largest margin occurs on the smallest source, ETIS-LaribPolypDB (domain C: 87.23 versus 81.74, a 5.49-point gain). On fundus, it leads on sources A, D, and E; it trails the best score by 0.18 points on B and 0.52 points on C. These gaps are all small; therefore, the fundus results show a stronger average.

Robustness to the choice of source domain. A single-source method should not depend on a fortunate choice of training centre. The per-source scores of DualBridge span only 3.10 points on polyp, versus at least 6.72 points for every competitor in Table 1. On fundus, its span is 2.27 points; the smallest competitor span is 3.00 points. Its *worst* source-domain result (86.12% on polyp, 84.66% on fundus) exceeds the *best average* of any compared method on the same task. Thus, the mean improvement is accompanied by a relatively narrow spread across the evaluated training sources.

Qualitative comparison. Figure 2 illustrates target-domain polyp cases with missed lesions, leakage, or fragmented masks in competing predictions. DualBridge better preserves the lesion region in these examples, including cases with weak lesion contrast or similar-looking surrounding mucosa. These visual comparisons complement the Dice scores by showing where mask boundaries differ; the source-averaged comparison is reported in Table 1. Additional cross-domain qualitative results, including successful fundus cases and failure cases on both fundus and polyp images, are shown in Appendix E.

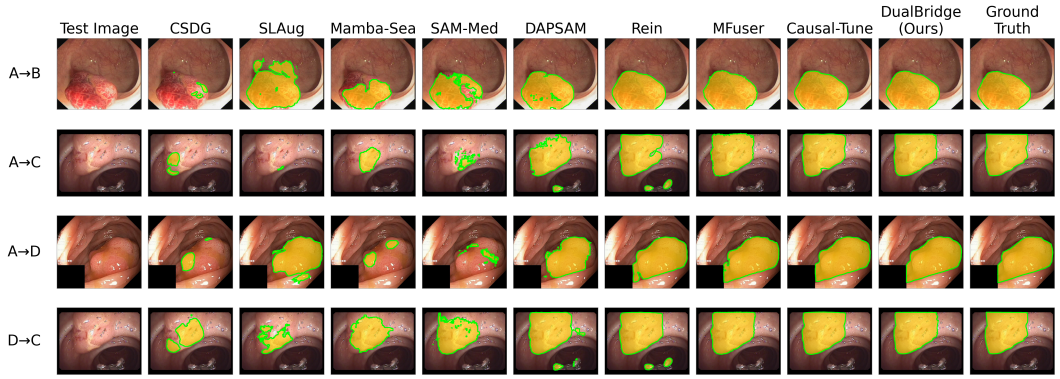

Figure 2: Qualitative comparison on unseen polyp domains. Rows show target images and masks from models trained on different source domains; columns show the input, ground truth, competing methods, and DualBridge. The examples illustrate missed lesions, leakage into surrounding tissue, and fragmented predictions.

4.3 Ablation Study

All ablations follow the protocol of Section 4.2; **Avg.** is the unweighted mean of the two tasks.

The two pathways are complementary. Without guidance, SAM3 reaches 86.23% on polyp but only 69.74% on fundus, whereas the Stage-1 DINOv3 head alone gives 75.32% and 84.32% (Table 4): the two models fail on different tasks. The spatial path lifts fundus by 15.93 points (Table 2); the feature path adds a further 0.66 on polyp and 0.33 on fundus. The larger spatial-path effect on fundus shows the value of explicit localization for that task, while the smaller feature-path gains motivate the content controls below.

The feature path helps through content, not perturbation. Table 3 keeps the adapter, parameter count, and training schedule fixed while changing the injected signal. Magnitude-matched noise yields an Overall score of 86.11, 0.70 points below the real residual and 0.21 below the spatial-path-only result. Spatially shuffled tokens yield 85.62, 1.19 points below the real residual and 0.70 below the spatial-path-only result. The real residual also leads both controls on each task separately: by 0.60–1.39 points on polyp and 0.80–0.99 points on fundus. Thus, the gain is visible in both task averages and depends on feature content and spatial alignment. The shuffled control is especially informative because it preserves the transferred tokens while disrupting their correspondence with SAM3's image features.

Both stages are necessary. Stage 2 improves on the Stage-1 head by 12.30 points on polyp and 1.68 on fundus (Table 4). Figure 3 shows why. On polyp, the prompt fires on distractors, captures only a fragment of the lesion or floods the surrounding mucosa (Dice 0.21–0.30), yet SAM3 recovers the correct object (0.92–0.97): the prompt serves as a localization hint, not as a mask to be refined. On fundus, the prompt already locates the disc but mis-sizes the cup (0.36–0.62); Stage 2 corrects the cup (0.80–0.86) and leaves the disc intact, which is why the fundus gain is smaller.

4.4 Trainable Parameters

DualBridge keeps both backbones frozen and trains 17.17 M parameters: 6.39 M in Stage 1 and 10.78 M in Stage 2, including 0.53 M in the feature adapters. This is approximately one tenth of MFuser's trainable parameter count (Table 5). Rein, Causal-Tune, DAPSAM, and SAM2-UNeXT train fewer parameters but score at least 3.79 points lower Avg.; MFuser trains more parameters and scores 2.59 points lower. The fixed-parameter controls in Table 3 further isolate the effect of transmitted feature content.

Table 2: Contribution of the spatial and feature paths. Avg. is the mean of the two task averages.

Configuration	Polyp	Fundus	Avg.
No guidance	86.23	69.74	77.99
+ Spatial path	86.96	85.67	86.32
+ Feature path (ours)	**87.62**	**86.00**	**86.81**

Table 3: Feature-content controls. Only the injected signal changes; the adapter architecture and trainable parameter count are fixed.

Injected signal	Polyp	Fundus	Avg.
Real residual (ours)	**87.62**	**86.00**	**86.81**
Magnitude-matched noise	87.02	85.20	86.11
Spatially shuffled tokens	86.23	85.01	85.62

Table 4: Stage-wise results. Stage 1 is the DINOv3 segmentation head; Stage 1+2 includes the prompted SAM3 stage.

Configuration	Polyp	Fundus	Avg.
Stage 1	75.32	84.32	79.82
Stage 1 + 2	**87.62**	**86.00**	**86.81**

Table 5: Trainable parameters.

Method	Params	Polyp	Fundus	Avg.
Rein	7.55 M	82.95	83.08	83.02
Causal-Tune	7.55 M	81.73	83.32	82.53
DAPSAM	11.67 M	75.66	84.41	80.04
MFuser	170.06 M	84.72	83.71	84.22
SAM2-UNeXT	6.00 M	83.71	81.43	82.57
Ours	17.17 M	**87.62**	**86.00**	**86.81**

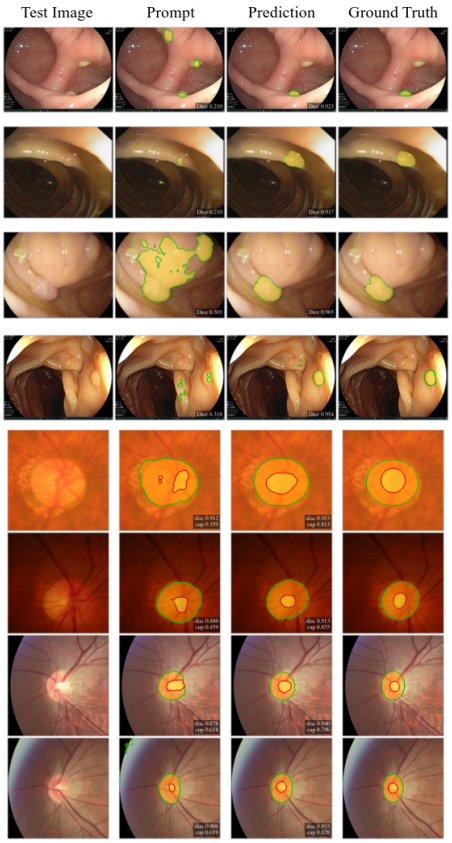

Figure 3: Target-domain examples comparing the Stage-1 spatial prompt with the final Stage-2 prediction (top: polyp; bottom: fundus).

5 Conclusion

We presented DualBridge for single-source domain-generalized medical image segmentation, using an information-bottleneck perspective to guide the interface between DINOv3 and SAM3. A DINOv3-derived mask prompt provides compact spatial guidance, while gated adapters transfer complementary dense features into SAM3's high-level representations. This design leaves SAM3's finest-resolution pathway on its native route and trains only the segmentation head and cross-model interfaces using source-domain labels. Without target-domain adaptation, DualBridge achieves source-averaged Dice scores of 87.62 on polyp and 86.00 on fundus. Ablations show that the spatial prompt drives most of the gain, while spatially aligned feature content provides an additional improvement.

Limitations. We have so far evaluated DualBridge only on two-dimensional medical images and have not analyzed its performance on three-dimensional medical images. In addition, our study focuses on cross-institutional datasets and does not examine cross-modality generalization. Extending the interface to three dimensions and validating it across a broader range of modalities and institutions remain future work.

AI USE STATEMENT

We used a generative AI tool to assist with English wording and editing of the limitations statement. The authors reviewed and revised the resulting text and take responsibility for the final manuscript, its claims, and all reported results.

ETHICS STATEMENT

This study uses publicly available, previously released fundus and polyp image datasets and does not collect new patient data. We follow the original datasets' terms of use. Since errors in medical image segmentation could affect downstream clinical decisions, the reported cross-institutional results should not be taken as evidence of clinical readiness; further validation is required before any clinical use.

REPRODUCIBILITY STATEMENT

Section 3.4 and Section 3.5 describe the two training stages and model interfaces. The dataset composition, source-to-target evaluation protocol, training settings, and metrics are reported in the Datasets and Implementation Details subsection of Section 4; the ablation protocol is described in Section 4.3. The appendix gives the Stage-1 head architecture and the assumptions and derivation behind the information-theoretic analysis.

REFERENCES

Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021.

Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2245–2264, 2025.

Lev Ayzenberg, Raja Giryes, and Hayit Greenspan. ProtoSAM for automated one shot medical image segmentation using foundational models. *Scientific Reports*, 15(1):1, 2025.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of machine learning research*, 3(Nov):463–482, 2002.

Jorge Bernal, F. Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99–111, 2015. ISSN 0895-6111.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International Conference on Learning Representations*, volume 2026, pp. 138846–138923, 2026.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Ziyang Chen, Yongsheng Pan, Yiwen Ye, Mengkang Lu, and Yong Xia. Each test image deserves a specific prompt: Continual test-time adaptation for 2d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11184–11193, 2024.

Zihan Cheng, Jintao Guo, Jian Zhang, Lei Qi, Luping Zhou, Yinghuan Shi, and Yang Gao. Mambasea: A mamba-based framework with global-to-local sequence augmentation for generalizable medical image segmentation. *IEEE Transactions on Medical Imaging*, 44(9):3741–3755, 2025.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, 2nd edition, 2006.

Francisco Fumero, Silvia Alayón, José L. Sanchez, Jose Sigut, and M. Gonzalez-Hernandez. Rim-one: An open retinal image database for optic nerve evaluation. In *2011 24th International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 1–6, 2011.

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. *arXiv preprint arXiv:2106.09685*, 2021.

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 *MultiMedia Modeling*, pp. 451–462, Cham, 2020. Springer International Publishing.

Jiaxu Jiang, Sen Lei, and Heng-Chao Li. Enhancing sam prompt generation with dinov3 representations for referring medical image segmentation. In *2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI)*, pp. 1–4. IEEE, 2026.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Yuheng Li, Yizhou Wu, Yuxiang Lai, Mingzhe Hu, and Xiaofeng Yang. Meddinov3: How to adapt vision foundation models for medical image segmentation? *arXiv preprint arXiv:2509.02379*, 2025.

Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment Anything with One Shot Using All-Purpose Feature Matching. In *International Conference on Learning Representations*, 2024.

M. Jehanzeb Mirza, Jakub Micorek, Horst Possegger, and Horst Bischof. The norm must go on: Dynamic unsupervised domain adaptation by normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14765–14775, 2022.

Phan Ngoc Lan, Nguyen Sy An, Dao Viet Hang, Dao Van Long, Tran Quang Trung, Nguyen Thi Thuy, and Dinh Viet Sang. Neounet : Towards accurate colon polyp segmentation and neoplasm detection. In *Advances in Visual Computing*, pp. 15–28, Cham, 2021. Springer International Publishing.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations*, 2023.

Shuaicheng Niu, Guohao Chen, Deyu Chen, Yifan Zhang, Jiaxiang Wu, Zhiquan Wen, Yaofo Chen, Peilin Zhao, Chunyan Miao, and Mingkui Tan. Adapt in the wild: Test-time entropy minimization with sharpness and feature regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel van Keer, Deepti R. Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical Image Analysis*, 59:101570, 2020. ISSN 1361-8415.

Cheng Ouyang, Chen Chen, Surui Li, Zeju Li, Chen Qin, Wenjia Bai, and Daniel Rueckert. Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(4):1095–1106, 2022.

Dzung L Pham, Chenyang Xu, and Jerry L Prince. Current methods in medical image segmentation. *Annual Review of Biomedical Engineering*, 2(1):315–337, 2000.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12159–12168. IEEE, 2021.

Hamza Rasaee, Taha Koleilat, and Hassan Rivaz. Grounding dino-us-sam: Text-prompted multiorgan segmentation in ultrasound with lora-tuned vision–language models. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 72(10):1414–1425, 2025.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9(2):283–293, 2014.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Herve Jegou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026.

Jayanthi Sivaswamy, S. R. Krishnadas, Gopal Datt Joshi, Madhulika Jain, and A. Ujjwaft Syed Tabish. Drishti-gs: Retinal image dataset for optic nerve head(onh) segmentation. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pp. 53–56, 2014.

Zixian Su, Kai Yao, Xi Yang, Kaizhu Huang, Qiufeng Wang, and Jie Sun. Rethinking data augmentation for single-source domain generalization in medical image segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2366–2374, 2023.

Naftali Tishby, Fernando Pereira, and William Bialek. The information bottleneck method. *Proceedings of the 37th Allerton Conference on Communication, Control and Computation*, 49, 07 2001.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Hanhui Wang, Huaize Ye, Yi Xia, and Xueyan Zhang. Leveraging sam for single-source domain generalization in medical image segmentation. *arXiv preprint arXiv:2401.02076*, 2024.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022.

Zhikai Wei, Wenhui Dong, Peilin Zhou, Yuliang Gu, Zhou Zhao, and Yongchao Xu. Prompting segment anything model with domain-adaptive prototype for generalizable medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 533–543. Springer, 2024a.

Zhixiang Wei, Lin Chen, Yi Jin, Xiaoxiao Ma, Tianle Liu, Pengyang Ling, Ben Wang, Huaian Chen, and Jinjin Zheng. Stronger, fewer, & superior: Harnessing vision foundation models for domain generalized semantic segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28619–28630. IEEE, 2024b.

Xinyu Xiong, Zihuang Wu, Lei Zhang, Lei Lu, Ming Li, and Guanbin Li. Sam2-unext: an improved high-resolution baseline for adapting foundation models to downstream segmentation tasks. *arXiv preprint arXiv:2508.03566*, 2025.

Hongzheng Yang, Cheng Chen, Meirui Jiang, Quande Liu, Jianfeng Cao, Pheng Ann Heng, and Qi Dou. Dltta: Dynamic learning rate for test-time adaptation on cross-domain medical images. *IEEE Transactions on Medical Imaging*, 41(12):3575–3586, 2022.

Jee Seok Yoon, Kwanseok Oh, Yooseung Shin, Maciej A Mazurowski, and Heung-Il Suk. Domain generalization for medical image analysis: A review. *Proceedings of the IEEE*, 112(10):1583–1609, 2024.

Jian Zhang, Lei Qi, Yinghuan Shi, and Yang Gao. Domainadaptor: A novel approach to test-time adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18971–18981, 2023.

Xin Zhang and Robby T. Tan. Mamba as a bridge: Where vision foundation models meet vision language models for domain-generalized semantic segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14527–14537, 2025.

Yin Zhang, Yongqiang Zhang, Yaoyue Zheng, Bogdan Raducanu, and Dan Liu. Causal-tune: Mining causal factors from vision foundation models for domain generalized semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12916–12924, 2026.

Zequn Zhang, Yun Jiang, Yunnan Wang, Baao Xie, Wenyao Zhang, Yuhang Li, Zhen Chen, Xin Jin, and Wenjun Zeng. Exploring contrastive pre-training for domain connections in medical image segmentation. *IEEE Transactions on Medical Imaging*, 44(4):1686–1698, 2025.

Zhuo Zhang, Feng Shou Yin, Jiang Liu, Wing Kee Wong, Ngan Meng Tan, Beng Hai Lee, Jun Cheng, and Tien Yin Wong. Origa-light: An online retinal fundus image database for glaucoma analysis and research. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pp. 3065–3068, 2010.

Wei Zhuo, Zhiyue Tang, Wufeng Xue, Hao Ding, Junkai Ji, and Linlin Shen. Uino-fss: Unifying representation learning and few-shot segmentation via hierarchical distillation and mamba-hypercorrelation. *IEEE Transactions on Image Processing*, 2026.

A Additional Analysis of Cross-Model Information Transfer

Assumptions and representation budgets. We consider a fixed domain distribution and fixed model parameters. The spatial prompt $m = h(D)$ and semantic channel $R = r_\eta(D)$ are deterministic functions of DINOv3 features. All logarithms are base two. A binary prompt with N entries has at most 2^N possible values, giving $H(m) \le N$. Similarly, a transmitted feature tensor with d_R entries at b-bit precision satisfies $H(R) \le b\, d_R$. These upper bounds describe the capacity of the two transmitted representations at finite precision. The precision assumption describes their numerical storage.

Proof of the complementary-information bound. Because (m, R) is a deterministic function of D, the data-processing inequality gives

$$I(Y; m, R) \le I(Y; D). \tag{8}$$

The chain rule then yields

$$I(Y; R \mid m) = I(Y; m, R) - I(Y; m) \le I(Y; D) - I(Y; m). \tag{9}$$

Furthermore,

$$I(Y; R \mid m) \le H(R \mid m) \le H(R) \le b\, d_R. \tag{10}$$

Combining these inequalities with non-negativity proves equation 2. The information gap concerns the full mapping from D to m, including segmentation-head prediction and thresholding.

Accounting for the receiving model. SAM 3 also receives information through its native visual representation F, defined before semantic feature injection. The additional information supplied by R is therefore more precisely characterized by

$$0 \le I(Y; R \mid F, m) \le \min\{I(Y; D \mid F, m),\, b\, d_R\}. \tag{11}$$

The conditional form accounts for information already present in SAM 3's native representation F. It characterizes the additional information available to the receiving model; the decoder's use of that information is examined empirically.

Finite-sample learning and integration complexity. Let $\mathcal{G}$ denote the class of predictors induced by the trainable integration mechanism, and let $\mathcal{L} = \ell \circ \mathcal{G}$ be its loss class. For i.i.d. source samples and losses bounded in $[0, 1]$, a standard empirical Rademacher bound gives, with probability at least $1 - \delta$, simultaneously for all $g \in \mathcal{G}$,

$$\mathcal{R}_s(g) \le \widehat{\mathcal{R}}_s(g) + 2\widehat{\mathfrak{R}}_n(\mathcal{L}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}. \tag{12}$$

This classical result motivates attention to the complexity of the learned integration mechanism under limited supervision (Bartlett & Mendelson, 2002). For nested function classes, restricting the class cannot increase its empirical Rademacher complexity. Architecture-specific comparisons require an explicit characterization of the corresponding function classes.

B Stage-1 Head Architecture

This section details the branches Π_k of equation 3. From each tapped hidden state we discard the class and register tokens and reshape the 256 patch tokens into a map $D^{(i_k)}(x) \in \mathbb{R}^{1024\times 16\times 16}$; with $D^{(0)}$ the patch embedding, $\mathcal{I} = (5, 11, 17, 23)$ taps every sixth block and excludes the final block. Each tap passes through a 1×1 convolution P_k from 1024 to a depth-dependent width c_k, with $(c_1, \ldots, c_4) = (256, 512, 1024, 1024)$ from the shallowest to the deepest tap, and then through a bias-free 3×3 convolution Q_k to a common width of 128:

$$A_k(x) = Q_k\Big(P_k\Big(D^{(i_k)}(x)\Big)\Big) \in \mathbb{R}^{128\times 16\times 16}, \qquad k = 1, \ldots, 4. \tag{13}$$

The shallowest branch is upsampled by a learned 4×4 transposed convolution with stride 4, which sets the common 64×64 grid, and the other three are bilinearly interpolated to it:

$$\Pi_1\Big(D^{(i_1)}\Big) = \mathrm{ConvT}_{4\times 4,\, s=4}(A_1)\,, \qquad \Pi_k\Big(D^{(i_k)}\Big) = \mathrm{Up}_{64\times 64}(A_k)\,, \quad k = 2, 3, 4. \tag{14}$$

The four branches are concatenated with equal weight, without per-branch gating, and no normalization or nonlinearity is applied inside the head. Because all taps of a plain ViT share one spatial resolution, this head differs from the original DPT decoder (Ranftl et al., 2021), which resamples the taps to four resolutions and fuses them progressively with residual refinement units; here all taps are brought to one grid and fused in a single step. The Dice term of equation 4 is $\mathcal{L}_{\mathrm{Dice}}(p, y) = 1 - \big(2\sum_i p_i y_i + \epsilon\big)/\big(\sum_i p_i + \sum_i y_i + \epsilon\big)$ with $\epsilon = 10^{-6}$, where i indexes pixels.

C Compared Methods

We summarize the methods compared in Table 1 . The first group adapts to unlabeled target images at inference; the other two groups are evaluated without target-domain training. These descriptions concern the original methods, while the numerical results in the tables follow the experimental settings of this paper.

Test-time adaptation. TENT (Wang et al., 2021) minimizes prediction entropy on target images and updates the affine parameters of normalization layers online. CoTTA (Wang et al., 2022) extends adaptation to a continual stream of changing target domains: an exponential-moving-average teacher supplies augmented predictions, while stochastic restoration of student weights helps retain the source model's knowledge. DLTTA (Yang et al., 2022) adjusts the test-time learning rate dynamically according to the discrepancy between each target image and the source-domain distribution, thereby varying the update strength across images. DUA (Mirza et al., 2022) updates batch-normalization statistics from incoming unlabeled samples with a dynamic momentum schedule, without gradient-based parameter updates. SAR (Niu et al., 2023) improves the stability of entropy-based adaptation by selecting reliable, low-entropy samples and using sharpness-aware optimization, with a recovery mechanism for model collapse. DomainAdaptor (Zhang et al., 2023) adapts the feature normalization and prediction process to the target domain at test time. VPTTA (Chen et al., 2024) performs image-specific visual prompt tuning: a learnable prompt adjusts the input for each test image while the segmentation network remains frozen, using source-domain feature statistics as the adaptation signal.

Single-source domain generalization. CSDG (Ouyang et al., 2022) uses a causal view of image formation to separate anatomy-related information from domain-specific appearance, and trains for robustness to interventions on the latter. SLAug (Su et al., 2023) generates diverse source-domain views through shape and low-frequency appearance augmentation, encouraging the segmenter to preserve anatomy across plausible acquisition changes. Mamba-Sea (Cheng et al., 2025) introduces global-to-local sequence augmentation in a Mamba-based segmentation framework to improve generalization to unseen medical-image domains. These methods use labeled source data to learn transferable segmentation models and require no target-image updates at deployment.

Foundation-model-based segmentation. SAM-Med (Wang et al., 2024) adapts the Segment Anything Model to medical-image segmentation, using its pretrained image representation and promptable mask prediction. DAPSAM (Wei et al., 2024a) uses domain-adaptive prototypes to prompt SAM for segmentation across unseen medical-image domains. Rein (Wei et al., 2024b) inserts lightweight trainable components into a frozen visual foundation model to refine its intermediate representations for downstream segmentation. MFuser (Zhang & Tan, 2025) fuses complementary features from CLIP and DINOv2 for generalizable medical-image segmentation. CausalTune (Zhang et al., 2026) uses causal guidance to tune a pretrained visual representation toward domain-invariant segmentation cues. Unlike the test-time adaptation methods above, these approaches do not optimize on unlabeled target images during inference.

D Dataset Details

The benchmark combines retinal photographs collected for optic-nerve-head analysis with colonoscopy images collected for polyp detection and segmentation. We use their pixel-level annotations for two segmentation tasks: optic disc/cup delineation and polyp delineation. Table 6 lists the image counts in our data manifests. Domain letters are task-specific. Following the five-domain fundus setting of VPTTA (Chen et al., 2024), REFUGE training images form domain B and its

Table 6: Dataset origins and image counts in our manifests. Train and Test indicate the within-domain manifest partitions; domain letters identify the source used in each single-source experiment.

Task/domain	Dataset	Train	Test	Total
Fundus A	RIM-ONE-r3 (Fumero et al., 2011)	99	60	159
Fundus B	REFUGE (Orlando et al., 2020)	320	80	400
Fundus C	ORIGA (Zhang et al., 2010)	500	150	650
Fundus D	REFUGE-Validation/Test (Orlando et al., 2020)	640	160	800
Fundus E	Drishti-GS (Sivaswamy et al., 2014)	50	51	101
Polyp A	BKAI-IGH-NeoPolyp (Ngoc Lan et al., 2021)	800	200	1000
Polyp B	CVC-ClinicDB (Bernal et al., 2015)	478	134	612
Polyp C	ETIS-LaribPolypDB (Silva et al., 2014)	152	44	196
Polyp D	Kvasir-SEG (Jha et al., 2020)	800	200	1000

validation/test images form domain D. These are partitions of one challenge, while the remaining fundus domains come from other named datasets. In each single-source experiment, labels from the chosen source domain are available for training; the other domains supply evaluation images.

Fundus images and annotations. The fundus datasets were developed for analysis of the optic nerve head, particularly the optic disc (OD) and optic cup (OC) used in glaucoma assessment. Each color photograph is paired with OD and OC delineations. We convert these annotations into two binary segmentation masks, with the cup nested inside the disc, and compute Dice separately for the two structures before taking their mean. This preserves the distinction between the relatively visible disc margin and the subtler cup boundary in the reported fundus score.

Fundus domains. RIM-ONE-r3 (A) is an optic-nerve-head image collection used in glaucoma-related research; our manifest contains 159 photographs with disc and cup annotations (Fumero et al., 2011). REFUGE was released as a glaucoma challenge with fundus images and optic-disc/cup segmentation labels (Orlando et al., 2020). Following VPTTA (Chen et al., 2024), we designate its 400 training images as B and its 800 validation/test images as D. ORIGA (C) contributes 650 images from an online retinal image archive assembled for glaucoma analysis, with OD and OC labels (Zhang et al., 2010). Drishti-GS (E) contributes 101 fundus photographs with expert disc and cup delineations for glaucoma-oriented evaluation (Sivaswamy et al., 2014). The datasets vary in image appearance and annotation context; we use their shared OD/OC labels to test transfer across the specified domains. For preprocessing, we extract an 800×800 region centered on the OD before the model-specific resizing described in Section 4.

Polyp images and annotations. The polyp datasets originate from colonoscopy imaging, where lesion localization and boundary delineation support computer-assisted detection and assessment. Each selected image is paired with a pixel-level polyp mask, which we use as a binary foreground label. BKAI-IGH-NeoPolyp (A) was introduced for identifying and segmenting neoplastic and non-neoplastic polyps; our benchmark uses 1000 of its annotated images for binary polyp segmentation (Ngoc Lan et al., 2021). CVC-ClinicDB (B) provides 612 annotated colonoscopy frames used for polyp detection and segmentation benchmarking (Bernal et al., 2015). ETIS-LaribPolypDB (C) provides 196 annotated frames for evaluating polyp-detection methods, making it the smallest polyp source here (Silva et al., 2014). Kvasir-SEG (D) extends the Kvasir endoscopy collection with 1000 images and manually annotated polyp masks for segmentation evaluation (Jha et al., 2020). Together, these sources vary in lesion size, appearance, illumination, and surrounding mucosa. We evaluate binary polyp masks with Dice under the same source-to-target protocol as the fundus task.

E Additional Qualitative Results

Figures 4 and 5 show successful cross-domain fundus cases. In these examples, DualBridge's optic-disc and optic-cup boundaries are close to the annotations despite changes in image appearance across source and target datasets. Figures 6 and 7 show cases where DualBridge still makes errors. The fundus examples include an oversized optic cup, while the polyp examples include missed or

over-segmented lesions. These examples complement the aggregate Dice scores in Table 1 and illustrate remaining failure modes.

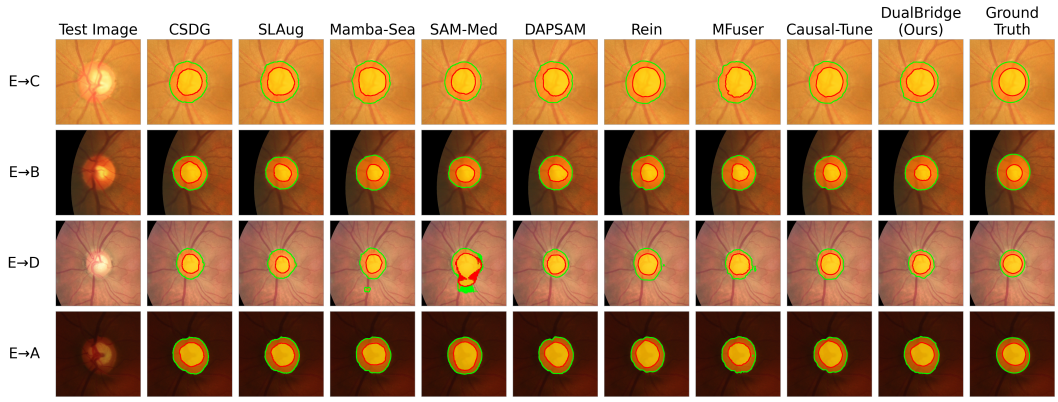

Figure 4: Successful cross-domain fundus examples with source domain E and target domains C, B, D, and A (top to bottom). Columns show the test image, baseline predictions, DualBridge, and ground truth. The overlaid contours indicate optic-disc and optic-cup boundaries.

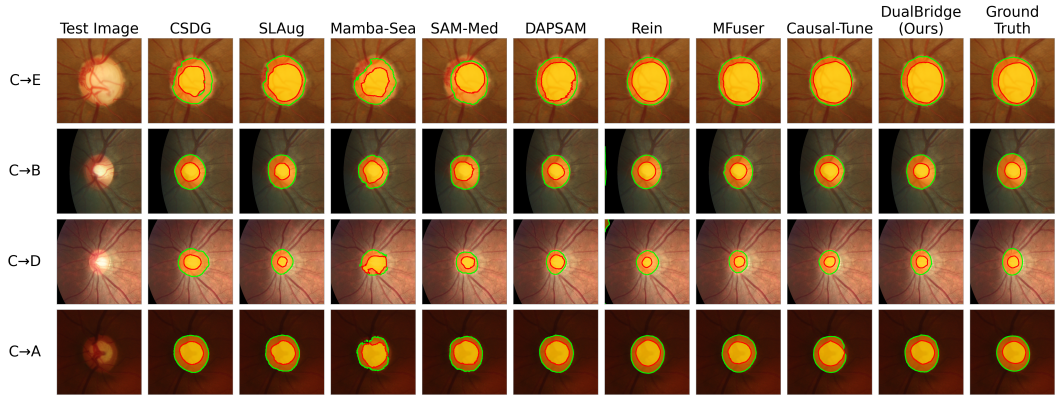

Figure 5: Successful cross-domain fundus examples with source domain C and target domains E, B, D, and A (top to bottom). The layout and boundary overlays follow Figure 4.

F Early Gate and Gradient Dynamics

We examined early Stage-2 optimization using the deepest Stage-1 tap ($i_4 = 23$), with $\gamma_l = -5$ at initialization and the same adapter learning-rate multiplier as in the main experiments. We used one polyp source (ETIS-LaribPolypDB; 152 training and 44 source-validation images) and one fundus source (Drishti-GS; 50 training and 51 source-validation images), each with seed 0. Table 7 reports the channel-mean gate $\sigma(\gamma_l)$ and the first-step gradients measured after backpropagation but before global gradient clipping. The projection gradient is the combined ℓ_2 norm of the 1×1 convolution's weight and bias gradients.

The gate and projection gradients are nonzero at the first optimizer step for both sources and both injected levels. The mean gates rise during the first ten steps and continue to rise through epoch 3. On the source validation splits, polyp Dice increases from 0.6600 at epoch 1 to 0.7123 at epoch 2 and 0.7287 at epoch 3; fundus mean disc/cup Dice increases from 0.7825 to 0.8171 and 0.8845. These measurements show that the small initial gate does not prevent early gradient flow or optimization in these two runs.

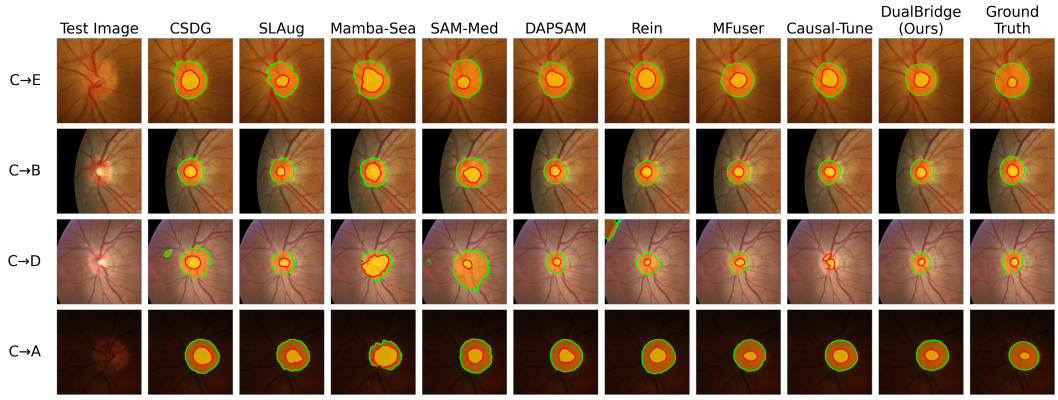

Figure 6: Fundus failure examples with source domain C and target domains E, B, D, and A (top to bottom). DualBridge segments the optic disc but overestimates the optic cup in several cases; the ground-truth column provides the reference boundaries.

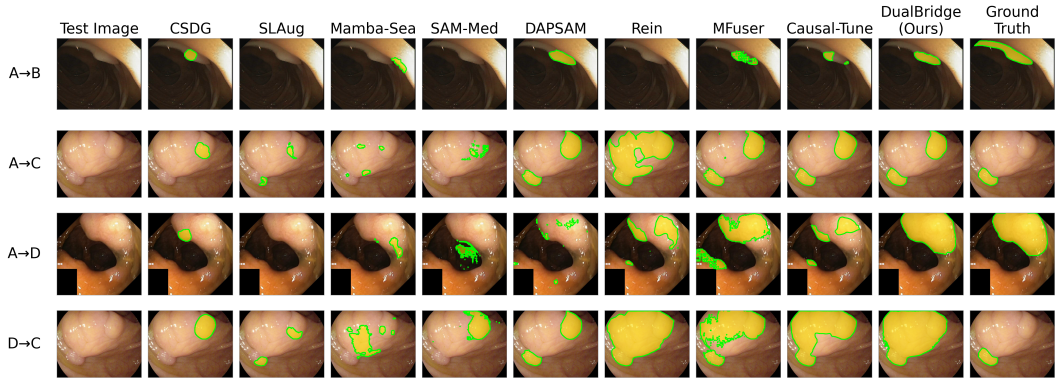

Figure 7: Polyp failure examples for source-to-target pairs A→B, A→C, A→D, and D→C (top to bottom). The examples include missed lesions and masks that extend into surrounding tissue; the final column shows the ground truth.

Table 7: Early Stage-2 gate dynamics with DINOv3 tap 23. Gate values are channel means; gradient norms are from the first optimizer step, before clipping.

Source	Level	Init.	Step 10	Epoch 1	Epoch 3	$\|\nabla_\gamma L\|_2$	$\|\nabla_{W,b} L\|_2$
Polyp (ETIS)	-2	0.00669	0.00676	0.00721	0.00764	2.35×10^{-4}	1.48×10^{-2}
Polyp (ETIS)	-1	0.00669	0.00675	0.00713	0.00778	9.87×10^{-4}	5.75×10^{-2}
Fundus (Drishti)	-2	0.00669	0.00673	0.00705	0.00731	5.74×10^{-4}	3.27×10^{-2}
Fundus (Drishti)	-1	0.00669	0.00672	0.00691	0.00706	2.02×10^{-3}	1.12×10^{-1}

Table 8: DINO backbone replacement controls. Columns A–E denote training sources, as in Table 1. Avg. is the unweighted mean over source domains, computed before rounding.

Task	Method / backbone	A	B	C	D	E	Avg.
Polyp	Rein / DINOv2-L	87.97	81.60	78.02	84.22	–	82.95
Polyp	Rein / DINOv3-L	86.92	81.49	81.24	84.86	–	83.63
Polyp	Causal-Tune / DINOv2-L	87.10	79.38	78.20	82.23	–	81.73
Polyp	Causal-Tune / DINOv3-L	86.69	82.37	82.26	83.98	–	83.82
Fundus	Rein / DINOv2-L	79.24	85.45	84.14	86.00	80.55	83.08
Fundus	Rein / DINOv3-L	82.21	86.08	79.15	86.80	81.49	83.15
Fundus	Causal-Tune / DINOv2-L	78.55	86.57	82.19	84.94	84.37	83.32
Fundus	Causal-Tune / DINOv3-L	83.40	86.32	81.83	86.78	82.58	84.18

G Backbone Replacement Controls

To assess how much of the comparison with DINO-based methods depends on backbone generation, we replace the DINOv2-L backbone of Rein and Causal-Tune with DINOv3-L and retain their original rows as references. Table 8 follows the source-to-target protocol of Table 1; source-domain scores average over target images, and Avg. weights source domains equally. These replacements control the backbone version within each baseline.

The DINOv3-L replacement raises Causal-Tune's average by about 2.1 points on polyp and 0.86 points on fundus, and Rein's by 0.68 and 0.07 points, respectively. DualBridge's reported averages remain higher than the DINOv3-L variants: by 3.80 and 1.82 points relative to Causal-Tune, and 3.99 and 2.85 points relative to Rein, on polyp and fundus respectively.