DINO PRIOR CALIBRATION FOR GLOBAL CONTEXT MODULATION IN MEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical image segmentation requires local details and global context, but unselective global information aggregation may introduce interference in regions with weak boundaries, low contrast, or similar tissues. In medical images, foreground targets and background tissues often have similar feature distributions, whereas DINO priors are mainly learned from natural images with clearer differences between objects and backgrounds. Consequently, directly introducing these priors may cause the model to attend to background regions with similar appearances or weaken its response to the true target and its boundaries. To address this transfer problem, we calibrate DINO priors with target-specific cues from a coarse segmentation prediction and supplement them with input image features. We propose DPCNet, a DINO Prior Calibration Network for medical image segmentation. DINO Prior Calibration uses coarse segmentation logits to calibrate multi-scale DINO priors. DPCNet then introduces the calibrated DINO priors into GECM as an independent global branch. The calibrated priors also dynamically regulate the structural frequency and selective dependency branches, which supplement the DINO prior with complementary global information from the input image. Experiments on multiple public medical image segmentation datasets demonstrate the effectiveness of DPCNet.

1 INTRODUCTION

Medical image segmentation is a key task in computer-aided diagnosis. It aims to accurately locate organs, lesions, or tissue structures from complex medical images. Compared with natural images, medical images often suffer from low contrast between targets and backgrounds, blurred boundaries, and similar appearance between foreground tissues and surrounding backgrounds (Hesamian et al., 2019; Litjens et al., 2017; Asgari Taghanaki et al., 2020). Therefore, an effective segmentation model needs local texture and boundary details together with semantic information that is selective for the current target.

DINO-based vision models learn rich semantic representations from large-scale natural images, and recent studies have transferred these representations as global priors for two-dimensional medical image segmentation (Caron et al., 2021; Siméoni et al., 2026; Li et al., 2025a; Yang et al., 2026). This transfer introduces a specific domain gap: natural-image pretraining often separates objects from backgrounds more clearly than medical images do, while medical targets and surrounding tissues can share intensity, texture, and structural patterns. A transferred prior can therefore mix target evidence with background responses, as illustrated in Fig. 1(a). The central challenge is consequently to make the spatial response of a generic DINO feature correspond to the target in the current medical image, rather than to add an unconditioned global signal.

Fine-tuning or adapting vision foundation models, including DINO, to medical segmentation has been explored to reduce this domain gap (Li et al., 2025b; Cekmeceli et al., 2024; Wang et al., 2025). Such adaptation increases training and inference cost, however, and does not by itself guarantee that responses outside the current target are suppressed, as illustrated in Fig. 1(a). These observations motivate target-specific calibration of a frozen prior: the model should use the segmentation state of the current image to decide which transferred DINO responses are reliable.

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.

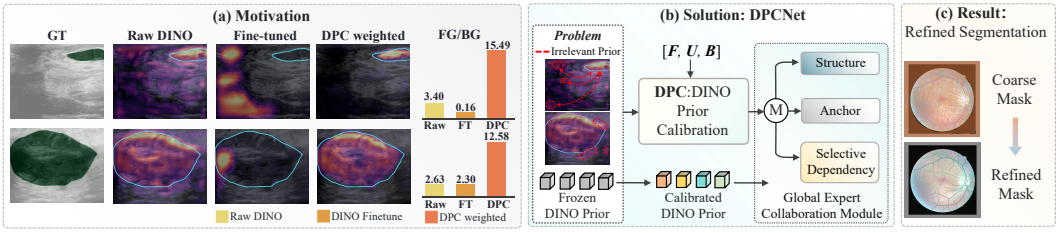


Figure 1: Overview and motivation of DPCNet. (a) Raw and fine-tuned DINO priors may retain responses outside the target region, whereas DPC-weighted responses are more concentrated on the target in these examples. FG/BG denotes the ratio between the mean response inside the ground-truth foreground and that in the far-background region. (b) DPCNet calibrates frozen DINO priors and integrates them with complementary global information from the input image. (c) The calibrated priors support coarse-to-fine segmentation refinement.

DINO priors contain useful generic visual information, but their transfer requires a signal that identifies the current target without using its ground-truth mask. Our core insight is to use a first-stage coarse prediction as this target-specific calibration signal. We propose DPCNet, a two-stage segmentation framework in which DINO Prior Calibration (DPC) derives foreground, uncertainty, and boundary cues from the coarse logits and uses them to adjust multi-scale frozen DINOv3 priors. The calibrated priors become more concentrated on the current target and its boundary, while responses from irrelevant background regions are reduced, as illustrated in Fig. 1(a,b). GECM then uses the calibrated prior as one selective global branch and organizes complementary global information from the input image through anchored context, structural frequency, and selective dependency branches. Accordingly, DPCNet is scoped to target-specific calibration and selective transfer of frozen natural-image features; the global-context components serve this calibration objective rather than defining a separate general-purpose context module.

Our contributions are summarized as follows:

- We formulate target-specific calibration of transferred DINO priors as the central problem and propose DPCNet, a two-stage framework that combines calibrated priors with complementary global information from the input image.
- We design a DINO Prior Calibration module that uses foreground, uncertainty, and boundary information from first-stage coarse segmentation logits to calibrate frozen multi-scale DINOv3 priors for the current medical target, without fine-tuning the DINOv3 backbone.
- We design GECM to introduce calibrated DINO priors as an independent global branch and use them to regulate global information from the input image.
- We validate DPCNet through main experiments, visualization analysis, and controlled ablations that separate the gains from two-stage refinement, raw DINO transfer, and prior calibration.

2 Related Work

Context Modeling in Medical Image Segmentation. U-Net and nnU-Net combine local details with semantic features through encoder–decoder paths and skip connections (Ronneberger et al., 2015; Isensee et al., 2021). nnWNet further incorporates global context modeling into the U-Net framework (Zhou et al., 2025), while Transformer- and state-space-based models capture long-range dependencies for medical image segmentation (Chen et al., 2024; Ruan et al., 2025). These approaches construct global representations primarily from internal network features and focus on enlarging the context range or strengthening long-range interaction.

Vision Foundation Model Priors for Medical Segmentation. Promptable foundation models provide another route to medical image segmentation. SAM produces masks from prompts (Kirillov et al., 2023), and MedSAM adapts this framework to medical images (Ma et al., 2024).

DINO-based self-supervised models learn structure-aware visual representations from large-scale natural images and can provide priors for dense prediction (Caron et al., 2021; Siméoni et al., 2026).

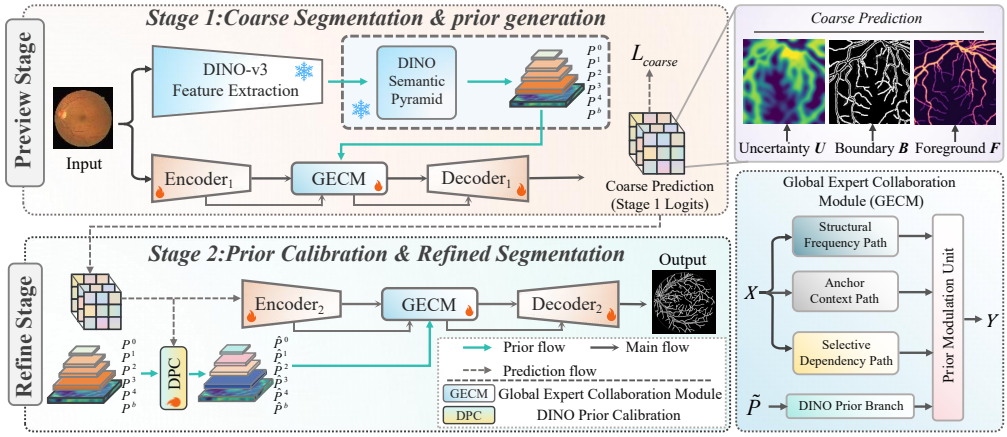


Figure 2: Architecture of the proposed DPCNet. The network uses a two-stage segmentation framework. DINO Prior Calibration calibrates frozen DINOv3 priors with first-stage coarse logits, and GECM uses the priors to coordinate complementary global expert branches.

Recent work transfers frozen DINO representations through spatial guidance, dense-feature adaptation, or multi-scale decoding (Liang et al., 2026; Li et al., 2025a; Yang et al., 2026). DPCNet instead calibrates multi-scale DINO priors using the current image's coarse prediction before global-context modeling.

Selective Fusion and Global Prior Modulation. Many segmentation methods introduce multi-scale fusion, attention fusion, or gating mechanisms to integrate features from different levels or types. Pyramid pooling and ASPP-like methods enhance target representation with multi-scale context (Zhao et al., 2017; Chen et al., 2017). Non-local and attention modules strengthen distant feature interaction through explicit relation modeling (Wang et al., 2018). Gated fusion methods such as GFF use dynamic weights to adjust the contribution of different feature branches (Li et al., 2020). These methods show that selective fusion can reduce information redundancy caused by fixed fusion.

3 METHOD

3.1 OVERVIEW

Medical image segmentation requires global context modeling to preserve complete target structures. However, existing methods that introduce global information without selection may bring target-related information and irrelevant background patterns into the segmentation process at the same time. DPCNet addresses this issue by using DINO priors as general visual priors and calibrating them with the coarse prediction of the segmentation network. The calibrated priors are then introduced through GECM, where they are combined with structural details and contextual information from the input image for refined segmentation.

As shown in Fig. 2, DPCNet contains three main components: a two-stage U-Net segmentation network, a DINO Prior Calibration module (DPC), and a Global Expert Collaboration Module (GECM). The first stage produces coarse segmentation logits, which provide foreground, uncertainty, and boundary information for DPC. DPC uses this information to calibrate the multi-scale DINO priors and make them better match the current medical segmentation target. The second stage receives the calibrated priors for refined segmentation. GECM introduces DINO priors into the network and combines them with complementary global information.

3.2 TWO-STAGE NETWORK DESIGN

Given an input image x, DPCNet consists of two U-Net segmentation stages. A frozen DINOv3 backbone (Siméoni et al., 2026) extracts four intermediate representations from x and converts them into a multi-scale prior pyramid $\mathcal{P} = \{P^s\}_{s\in\mathcal{S}}$. The first-stage U-Net uses the raw DINO priors for initial global context modeling. Specifically, the four representations are projected and fused in a top-down manner to construct five global priors $\{P^0, P^1, P^2, P^3, P^4\}$. The deepest DINO rep-

resentation is also separately projected to form a bottleneck prior P^b. Hence, $\mathcal{S} = \{0, 1, 2, 3, 4, b\}$ contains five global scales and one bottleneck scale. The six tensors are resized to the corresponding U-Net resolutions before calibration.

In the first-stage U-Net, GECM receives the encoder feature together with the raw DINO prior. Its output is passed to the decoder through the U-Net skip connection, allowing the first-stage U-Net to use the general global information provided by DINO for initial segmentation. The purpose of the first-stage U-Net is to produce a coarse prediction that reflects the current medical target. This prediction provides the foreground, uncertainty, and boundary information required by DPC to calibrate the raw DINO priors.

The first-stage U-Net decoder finally produces a full-resolution feature, which is passed through a 1×1 convolutional prediction head to obtain the coarse segmentation logits z_1. DPC uses the information derived from z_1 to calibrate the prior pyramid and obtain $\hat{\mathcal{P}}$.

The second stage receives the first-stage context through two paths: the image is concatenated with the full-resolution first-stage decoder feature at the input, and at each encoder scale the downsampled second-stage feature is concatenated with the corresponding decoder-side GECM output of the first stage before processing. The refined stage therefore builds on the global context already established by the first stage, while DPC updates only the prior supplied to GECM. In the second-stage U-Net, GECM receives the calibrated DINO priors together with the intermediate features transferred from the first-stage U-Net. It uses the calibrated priors for global context modeling, and the second-stage U-Net decoder produces the refined segmentation $\hat{y}$.

The raw priors support initial context modeling, and the resulting coarse prediction calibrates the priors for the second stage.

3.3 DINO Prior Calibration

DPC uses foreground, uncertainty, and boundary information from coarse logits to calibrate multi-scale DINO priors. Foreground information identifies target-related regions, uncertainty information describes ambiguous areas, and boundary information preserves structural transitions.

Given the first-stage coarse logits z_1, DPC derives a foreground response F, an uncertainty map U, and a boundary map B. Here, F indicates the predicted foreground probability, U highlights uncertain regions, and B describes transitions between foreground and background:

$$\begin{aligned} F &= \sigma(z_1), \qquad U = 1 - |2F - 1|, \\ B &= \mathrm{clip}(|\nabla_x F| + |\nabla_y F|, 0, 1). \end{aligned} \tag{1}$$

The foreground response identifies likely target regions, the uncertainty map is incorporated to describe ambiguous regions, and the boundary map captures spatial transitions that are important for medical segmentation. DPC further computes a prompt gate:

$$Q = \mathrm{clip}\left(|2F - 1|^{\rho}(1 + \beta B), 0, 1\right), \tag{2}$$

where ρ controls how sharply the reliability weight varies with coarse-prediction confidence, and β increases the gate near predicted boundaries. We set $\rho = 1$ and $\beta = 1$ in all experiments.

For each scale s, DPC concatenates the foreground, uncertainty, and boundary maps into a prompt signal $S = [F, U, B]$. The signal is resized to the resolution of P^s and projected to its channel space, yielding A^s. Here, U informs the learned refinement of ambiguous regions, while Q limits the update magnitude when the coarse prediction is unreliable. The boundary map B provides structural information to both the prompt signal and the reliability gate. Based on A^s, DPC generates a spatial gate modulated by Q and a residual refinement response from the raw prior:

$$G^s = \mathrm{Gate}_s(A^s) \odot Q, \qquad R^s = \mathrm{Refine}_s([P^s, A^s]). \tag{3}$$

The calibrated prior is obtained through residual updating:

$$\hat{P}^s = P^s + \eta \boldsymbol{\lambda}_s G^s \odot R^s. \tag{4}$$

Here, η is a warm-up factor and $\boldsymbol{\lambda}_s$ is a learnable channel-wise residual scale. The prompt signal S informs the learned correction, while Q uses prediction confidence and boundary strength to limit

updates in unreliable regions. Because the gate is symmetric in prediction confidence, it acts as a reliability gate rather than a foreground selector. A controlled coarse-prompt intervention analysis in the supplementary quantifies refinement behavior under boundary shifts, high-confidence false positives, and missed targets, identifying the conditions under which calibration is most and least reliable.

3.4 Global Expert Collaboration Module

Although calibration makes DINO priors more consistent with the segmentation target, using DINO priors alone for global context is insufficient because they do not contain all structural details and contextual information from the input image. To address this limitation, we propose GECM, which introduces the DINO prior into the network as an independent global branch and uses expert branches to complement it with different types of information from the input image. At each scale, let X denote the input U-Net feature and let $\tilde{P}$ denote the DINO prior. In the first-stage U-Net, GECM receives the raw DINO prior, i.e., $\tilde{P} = P^s$, for global context modeling. In the second-stage U-Net, GECM receives the calibrated prior, i.e., $\tilde{P} = \hat{P}^s$, and integrates it with image-based global information for refined segmentation.

The anchor context branch (ACB) provides the base contextual representation. The structural frequency branch (SFB) projects X into the frequency domain using FFT, applies learned complex weights to different frequency components, and transforms the feature back to the spatial domain. It captures high-frequency texture and boundary details together with low-frequency structural information. The selective dependency branch (SDB) projects the input feature, adds positional information through a depth-wise convolution, and flattens the feature into a sequence. Mamba then models long-range dependencies, and the resulting feature is restored to the spatial domain (Gu & Dao, 2023). The three image-based expert features are denoted as $g_0 = \mathrm{ACB}(X)$, $g_f = \mathrm{SFB}(X)$, and $g_m = \mathrm{SDB}(X)$.

The DINO prior is resized to match the feature resolution and projected into the same channel space, yielding $d = \phi_d(\tilde{P})$. The projected prior d enters GECM as an independent DINO branch. A lightweight gating network uses the input feature X, the projected prior d, and the three branch features to predict γ_f, γ_m, and γ_d, which determine the contributions of the structural frequency, selective dependency, and DINO branches, respectively. The anchor context branch remains the base global response. The final output is computed as

$$\begin{aligned} Y = g_0 &+ \alpha_f \gamma_f \odot g_f + \alpha_m \gamma_m \odot g_m \\ &+ \alpha_d \gamma_d \odot d. \end{aligned} \tag{5}$$

The learnable residual scales α_f, α_m, and α_d are initialized to zero, so GECM starts from the anchor branch and learns the contributions of the frequency, dependency, and DINO responses. Implementation details, the controlled two-stage comparison, and the additional DPC/GECM analyses are provided in Supplementary Sections 1–8.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate DPCNet on six medical image segmentation benchmarks. ISIC-2017, BUSI, DRIVE, and CREMI cover dermoscopic lesion segmentation, breast ultrasound lesion segmentation, retinal vessel segmentation, and electron microscopy structure segmentation, respectively (Codella et al., 2018; Al-Dhabyani et al., 2020; Staal et al., 2004; CREMI Challenge, 2016). ISIC-2017 follows its official split, which contains 2,000 training images, 150 validation images, and 600 test images. DRIVE follows its official split with 20 training images and 20 test images, and all metrics are computed inside the field-of-view (FOV) mask. CREMI follows the volume-wise A+B→C protocol to avoid slice-level leakage. BUSI contains 647 lesion images and is evaluated using stratified 5-fold cross-validation. For endoscopic polyp segmentation, we follow the widely used PraNet-style 90/10 partition (Fan et al., 2020) of CVC-ClinicDB and Kvasir-SEG (Bernal et al., 2015; Jha et al., 2020): 900 Kvasir-SEG and 550 CVC-ClinicDB images are used for training, while the remaining 100 and 62 images are used for testing, respectively.

Method	Venue	Params	MACs	ISIC-2017				BUSI			
		(M)	(G)	Jaccard↑	Dice↑	95HD↓	ASD↓	Jaccard↑	Dice↑	95HD↓	ASD↓
nnU-Net	Nat. Methods'21	46.33	466.25	75.50 ± 0.39	84.05 ± 0.35	17.70 ± 0.97	7.76 ± 0.75	71.53 ± 2.77	79.30 ± 2.82	20.73 ± 3.78	8.19 ± 2.08
EMCAD	CVPR'24	26.76	10.58	74.55 ± 0.48	83.43 ± 0.22	19.01 ± 1.35	8.25 ± 0.92	61.95 ± 2.21	72.02 ± 2.20	30.86 ± 3.57	10.99 ± 1.38
VM-UNet	ACM TOMM'25	44.27	7.56	75.81 ± 0.42	84.28 ± 0.40	18.19 ± 1.16	8.04 ± 0.33	68.15 ± 1.96	77.04 ± 1.66	22.39 ± 3.37	8.60 ± 1.08
nnWNet	CVPR'25	7.04	42.81	75.62 ± 0.31	84.00 ± 0.34	18.01 ± 1.74	7.85 ± 0.54	70.29 ± 1.60	78.09 ± 1.67	21.19 ± 2.29	8.27 ± 0.95
UTANet	AAAI'25	24.28	67.06	76.27 ± 0.54	84.71 ± 0.39	18.33 ± 1.40	7.62 ± 0.88	67.99 ± 2.51	76.51 ± 2.39	22.06 ± 2.13	8.44 ± 1.09
CENet	MICCAI'25	33.39	12.84	76.78 ± 0.35	85.13 ± 0.31	18.77 ± 0.99	7.82 ± 0.77	64.92 ± 2.44	74.68 ± 2.33	26.93 ± 4.01	9.65 ± 1.14
SegDINO	MICCAI'26	24.90	80.30	76.85 ± 0.45	85.19 ± 0.20	16.30 ± 1.04	7.06 ± 0.52	71.14 ± 1.77	78.46 ± 1.55	22.17 ± 3.16	8.22 ± 1.14
DPCNet (Ours)	–	14.48	68.71	**77.52** ± 0.38	**85.67** ± 0.24	**15.68** ± 1.13	**6.85** ± 0.31	**72.61** ± 1.98	**80.04** ± 1.91	**18.78** ± 3.97	**7.42** ± 1.60

Method	Venue	Params	MACs	DRIVE				CREMI			
		(M)	(G)	Jaccard↑	Dice↑	95HD↓	ASD↓	Jaccard↑	Dice↑	95HD↓	ASD↓
nnU-Net	Nat. Methods'21	46.33	466.25	68.61 ± 0.92	81.06 ± 0.81	6.30 ± 0.60	2.03 ± 0.23	76.37 ± 0.32	85.82 ± 0.32	8.35 ± 1.39	2.24 ± 0.27
EMCAD	CVPR'24	26.76	10.58	63.62 ± 0.87	77.74 ± 0.28	12.32 ± 0.47	2.18 ± 0.39	72.39 ± 0.21	83.18 ± 0.20	11.52 ± 0.37	2.75 ± 0.11
VM-UNet	ACM TOMM'25	44.27	7.56	58.64 ± 1.02	73.88 ± 0.64	10.82 ± 0.86	2.06 ± 0.29	73.97 ± 0.56	84.20 ± 0.19	10.70 ± 0.63	3.04 ± 0.35
nnWNet	CVPR'25	7.04	42.81	68.78 ± 0.50	81.11 ± 0.49	4.87 ± 1.14	1.92 ± 0.27	77.19 ± 0.48	86.23 ± 0.26	6.26 ± 0.90	1.32 ± 0.26
UTANet	AAAI'25	24.28	67.06	62.99 ± 0.74	77.26 ± 0.36	7.37 ± 0.87	1.56 ± 0.39	74.12 ± 0.30	84.35 ± 0.26	10.71 ± 1.07	3.33 ± 0.48
CENet	MICCAI'25	33.39	12.84	55.30 ± 0.95	71.19 ± 0.61	16.86 ± 0.65	3.20 ± 0.14	74.18 ± 0.52	84.32 ± 0.27	11.91 ± 1.31	3.67 ± 0.36
SegDINO	MICCAI'26	24.90	80.30	54.44 ± 0.49	69.76 ± 0.37	17.54 ± 1.15	3.70 ± 0.20	76.23 ± 0.61	85.53 ± 0.48	8.61 ± 1.04	2.51 ± 0.50
DPCNet (Ours)	–	14.48	68.71	**70.19** ± 0.73	**82.46** ± 0.50	**3.30** ± 0.62	**0.84** ± 0.13	**77.66** ± 0.41	**86.57** ± 0.50	**6.18** ± 0.97	**1.31** ± 0.11

Table 1: Quantitative comparison on ISIC-2017, BUSI, DRIVE, and CREMI. Params denotes trainable parameters (M), and MACs denotes computational cost (G). DPCNet achieves the highest mean overlap scores and competitive boundary accuracy across different imaging modalities.

Method	Venue	Params	MACs	CVC-ClinicDB					
		(M)	(G)	Jaccard↑	Dice↑	Pre.↑	Rec.↑	95HD↓	ASD↓
nnU-Net	Nat. Methods'21	46.33	466.25	87.59 ± 0.17	91.76 ± 0.20	94.12 ± 0.63	93.03 ± 0.88	6.92 ± 1.22	2.04 ± 0.51
EMCAD	CVPR'24	26.76	10.58	68.04 ± 0.59	76.68 ± 0.74	76.14 ± 1.41	83.09 ± 1.58	45.65 ± 1.61	17.69 ± 0.67
VM-UNet	ACM TOMM'25	44.27	7.56	80.56 ± 0.38	86.78 ± 0.68	86.29 ± 1.16	93.49 ± 1.24	13.69 ± 1.30	4.53 ± 0.46
nnWNet	CVPR'25	7.04	42.81	85.89 ± 0.49	90.16 ± 0.53	**94.15** ± 0.76	91.28 ± 1.04	8.11 ± 0.70	2.75 ± 0.29
UTANet	AAAI'25	24.28	67.06	67.77 ± 0.55	72.63 ± 0.58	93.33 ± 0.82	74.01 ± 0.67	21.83 ± 1.30	7.80 ± 0.40
CENet	MICCAI'25	33.39	12.84	59.82 ± 0.95	68.54 ± 0.72	62.27 ± 0.99	84.86 ± 0.57	64.38 ± 1.99	26.81 ± 0.48
SegDINO	MICCAI'26	24.90	80.30	88.37 ± 0.25	92.05 ± 0.20	91.99 ± 0.95	92.13 ± 0.78	7.96 ± 0.98	2.71 ± 0.50
DPCNet (Ours)	–	14.48	68.71	**88.50** ± 0.27	**92.71** ± 0.59	93.98 ± 0.40	**93.94** ± 0.91	**6.55** ± 0.68	**2.00** ± 0.31

Method	Venue	Params	MACs	Kvasir					
		(M)	(G)	Jaccard↑	Dice↑	Pre.↑	Rec.↑	95HD↓	ASD↓
nnU-Net	Nat. Methods'21	46.33	466.25	89.41 ± 0.33	93.51 ± 0.48	95.67 ± 0.49	92.67 ± 0.82	28.56 ± 1.61	9.05 ± 0.56
EMCAD	CVPR'24	26.76	10.58	88.95 ± 0.83	93.53 ± 0.65	95.75 ± 0.69	92.08 ± 1.64	23.57 ± 2.63	8.34 ± 0.93
VM-UNet	ACM TOMM'25	44.27	7.56	87.94 ± 0.73	92.62 ± 0.35	94.70 ± 1.20	92.24 ± 1.80	30.43 ± 1.57	10.27 ± 0.63
nnWNet	CVPR'25	7.04	42.81	88.40 ± 0.95	92.82 ± 0.67	95.95 ± 0.83	91.88 ± 1.50	27.19 ± 1.92	8.55 ± 0.46
UTANet	AAAI'25	24.28	67.06	84.15 ± 1.19	90.00 ± 0.67	94.33 ± 0.52	89.11 ± 1.05	40.86 ± 3.61	11.37 ± 1.20
CENet	MICCAI'25	33.39	12.84	83.64 ± 0.49	89.78 ± 0.55	88.81 ± 0.87	93.13 ± 1.32	38.47 ± 2.05	11.09 ± 0.87
SegDINO	MICCAI'26	24.90	80.30	88.95 ± 0.55	92.51 ± 0.42	96.23 ± 0.23	92.17 ± 0.74	27.55 ± 3.11	8.46 ± 0.75
DPCNet (Ours)	–	14.48	68.71	**90.75** ± 0.20	**94.18** ± 0.42	**96.35** ± 0.31	**93.86** ± 0.58	**22.73** ± 1.74	**7.62** ± 0.57

Table 2: Quantitative comparison on CVC-ClinicDB and Kvasir. DPCNet achieves the highest mean performance on most reported metrics for endoscopic polyp segmentation.

Implementation details. DPCNet is implemented in PyTorch on nnU-Netv2 and trained on a single NVIDIA GeForce RTX 5090 GPU. We use a frozen DINOv3 ViT-S/16 backbone and extract the outputs of Transformer blocks $\{2, 5, 8, 11\}$. These four features are reshaped and fused into five global prior scales, while the deepest feature is separately projected as a bottleneck prior; DPC therefore processes six prior tensors in total. Dataset-specific preprocessing, patch size, batch size, and data augmentation follow the automatically generated 2D nnU-Netv2 plans. Single-channel images are repeated to three channels before DINOv3 inference. All parameter counts reported for the compared methods and DPCNet in the main tables refer to trainable parameters. DPCNet has 14.48M trainable parameters. The frozen DINOv3 ViT-S/16 backbone contributes 21.60M additional parameters at inference, resulting in 36.08M total parameters. The frozen backbone receives no gradient updates.

The two stages are trained jointly with $\mathcal{L} = \mathcal{L}_{\text{refined}} + \mu\mathcal{L}_{\text{coarse}}$, where $\mu = 0.2$. Both losses use Dice and cross-entropy terms. $\mathcal{L}_{\text{refined}}$ is applied to the second-stage output with deep supervision weights $(8, 4, 2, 1, 0)/15$, while $\mathcal{L}_{\text{coarse}}$ supervises the first-stage coarse prediction. The stage-one logits are detached only before entering DPC; the coarse loss still provides gradients to the first stage. We use SGD with Nesterov momentum 0.99, weight decay 3×10^{-5}, an initial learning rate of 5×10^{-3}, and a polynomial learning-rate schedule. DPCNet and all of its ablation variants are

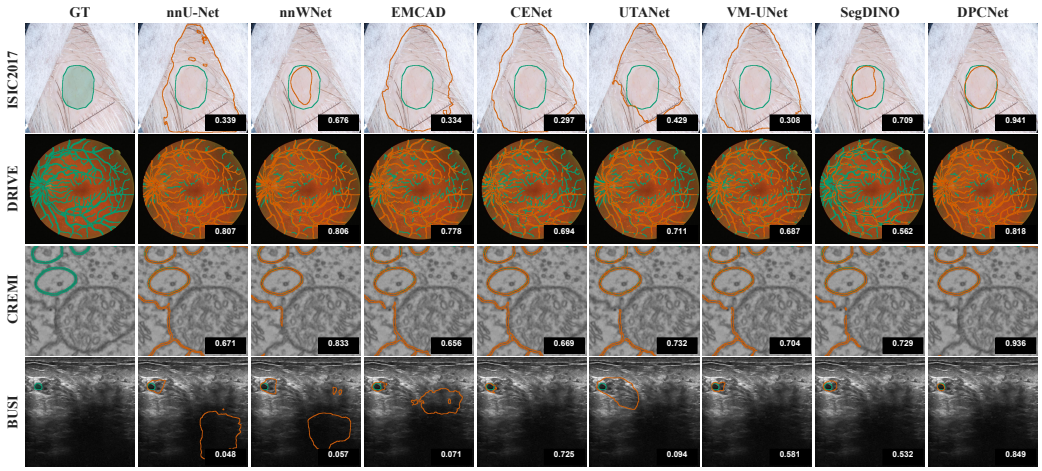


Figure 3: Qualitative comparison of representative segmentation methods. Green contours denote ground truth, orange contours denote model predictions, and the value at the lower right of each prediction is the Dice score. DPCNet shows more consistent target coverage and boundary localization across different imaging scenarios.

trained for 200 epochs. The DPC residual scale is initialized to 0.001, GECM residual scales to 0, and the warm-up reaches 1 after 40 epochs.

Baselines. We compare DPCNet with nnU-Net, nnWNet, EMCAD, CENet, UTANet, VM-UNet, and SegDINO (Isensee et al., 2021; Zhou et al., 2025; Rahman et al., 2024; Bozorgpour et al., 2025; Luo et al., 2025; Ruan et al., 2025; Yang et al., 2026). nnU-Net is a general-purpose medical segmentation baseline, whereas nnWNet is the direct architectural baseline because DPCNet extends its global modeling backbone with DINO prior calibration and GECM.

Evaluation metrics. We report Jaccard, Dice, 95% Hausdorff Distance (95HD), and Average Surface Distance (ASD) on all datasets. Jaccard and Dice measure region overlap, while 95HD and ASD measure boundary error. For the polyp comparison, we additionally report precision and recall. All metrics are computed per image and then averaged over the test set. Higher Jaccard, Dice, precision, and recall are better; lower 95HD and ASD are better.

4.2 Comparison with State-of-the-Art Methods

We compare DPCNet with nnU-Net, nnWNet, EMCAD, CENet, UTANet, VM-UNet, and SegDINO using official implementations, recommended settings, and identical splits. Tables 1 and 2 report mean$\pm$standard deviation over three seeds, except for five-fold BUSI evaluation. DPCNet achieves the best mean Jaccard and Dice across the four benchmarks, the lowest boundary errors on DRIVE and CREMI, and the best results on CVC-ClinicDB and Kvasir. Averaged over the four benchmarks, it improves Jaccard/Dice from $72.97\%/82.36\%$ for nnWNet to $74.50\%/83.69\%$. Figure 3 shows more consistent target coverage and boundary localization.

4.3 DPC and GECM DINO-Branch Analysis

We conduct a component-level analysis on the fixed Kvasir-SEG validation split (880 training and 120 validation images). For DPC, we compare the raw DINO prior P^s, the diagnostic coarse-prompt response $P^s \odot Q^s$, and the final calibrated prior $\hat{P}^s$ across six prior scales, where Q^s is the coarse-prompt gate resized to scale s. For GECM, we compare the projected DINO activation d with its output contribution $\alpha_d \gamma_d \odot d$ across 12 GECM blocks. All measurements are averaged within each case before paired testing, avoiding the use of intermediate features as independent samples.

We compute channel-wise ℓ_2 response maps after alignment with the ground-truth resolution. F/B and B/Bg denote foreground- and boundary-to-far-background activation, respectively, and Leak is the far-background response-energy fraction. Table 3 shows that the coarse-prompt gate increases target and boundary activation while reducing leakage. The final calibrated prior retains higher F/B and B/Bg values and lower leakage than the raw prior. GECM further improves the target

Metric	DPC raw P^s	DPC gate $P^s \odot Q^s$	DPC final $\hat{P}^s$	GECM input d	GECM output $\alpha_d \gamma_d \odot d$
F/B ↑	2.02	9.15†	2.84†	2.53	4.01†
Leak ↓	0.79	0.50†	0.54†	0.76	0.70†
B/Bg ↑	1.38	3.29†	1.86†	1.54	1.65†

Table 3: DINO-response analysis on Kvasir-SEG ($n = 120$). DPC columns are case-wise averages across six prior scales; GECM columns are averaged across 12 blocks. † denotes a Holm-corrected paired test against the corresponding input ($q < 0.05$).

Variant	DPC	F	U	B	Kvasir-SEG Jac.↑	Kvasir-SEG Dice↑	BUSI Jac.↑	BUSI Dice↑
No DINO prior					80.05	86.79	70.35	78.51
Zero DINO prior (full DPC)	✓	✓	✓	✓	80.15	87.01	70.75	78.68
Raw DINO prior					80.39	87.11	70.98	78.89
DPC w/ F	✓	✓			81.08	87.44	71.85	79.20
DPC w/ $F + U$	✓	✓	✓		81.34	87.54	72.53	79.88
DPC w/ $F + B$	✓	✓		✓	81.21	87.53	72.34	79.83
DPC w/ $F + U + B$ (Ours)	✓	✓	✓	✓	**81.59**	**87.83**	**73.55**	**80.12**

Table 4: Ablation of DPC and its coarse-prediction maps on Kvasir-SEG and BUSI. Variant definitions are given in Section 4.4.

localization of its DINO branch, with higher F/B and B/Bg values and lower leakage than its input. Thus, the diagnostic gate and learned correction both contribute to the spatial selection of DINO information before and within GECM.

4.4 Ablation Studies

The ablations are conducted on the official Kvasir-SEG 880/120 split and the fixed BUSI fold-0 split (517 training and 130 held-out images), using Jaccard and Dice. All ablation variants are trained for 200 epochs. Table 4 examines DPC by comparing prior removal ("No DINO prior"), direct use of raw DINO priors, and progressively adding the coarse-prediction maps used for calibration. It also includes a zero-prior control ("Zero DINO prior") that retains the full DPC architecture and trainable parameter count while replacing every DINO prior with zero. This control separates the contribution of DPC pathway capacity from the visual information provided by DINO priors. Tables 5 and 6 examine the expert branches in GECM and its DINO-guided fusion strategy. Tables 7 and 8 compare DPC with direct DINO adaptation and alternative prior injection strategies.

To separate the benefit of DPC from the capacity of coarse-to-fine refinement, the supplementary material reports a controlled two-stage ablation using the same U-Net stages, cross-stage connections, supervision, training schedule, and inference protocol (Supplementary Table S1). On Kvasir-SEG, moving from a single-stage U-Net to the two-stage baseline increases Dice from 85.62 to 86.41, while adding image experts and the GECM gate increases it to 86.79. Adding the raw DINO branch further gives 87.11, whereas replacing the raw prior with the calibrated prior gives 87.83. The corresponding BUSI Dice values are 75.36, 78.17, 78.51, 78.89, and 80.12. Thus, the raw-prior comparison measures the benefit of DINO transfer under the same refinement architecture, and the raw-to-calibrated comparison isolates the additional contribution of target-specific DPC.

As shown in Table 4, the zero-prior control is close to the no-prior variant, with Dice improvements of 0.22 pp on Kvasir-SEG and 0.17 pp on BUSI. Thus, retaining the DPC pathway alone provides limited gains without DINO visual information. Directly using raw DINO priors gives modest improvements. In contrast, full DPC improves Dice by 0.82 pp and 1.44 pp over the zero-prior control on Kvasir-SEG and BUSI, respectively. Foreground conditioning improves over raw-prior use, and incorporating uncertainty and boundary maps further improves the results. The full F, U, and B design achieves the highest mean performance on both datasets, indicating that these maps provide complementary conditioning for learned prior calibration.

As shown in Table 5, the frequency and selective dependency branches complement the DINO prior with image-derived structural information. Using all experts with the DINO branch and DINO-guided gating performs best on both datasets. Table 6 further shows that fixed summation, learned scales, and gating without DINO are weaker than DINO-guided gating. Thus, the calibrated prior is useful both as an independent branch and as guidance for combining image-based expert features.

As shown in Table 7, full fine-tuning and both LoRA settings remain below DPC calibration on both datasets despite updating additional DINO parameters. Table 8 further shows that calibrated priors introduced through GECM outperform alternative injection strategies. These results support calibrating the prior with the first-stage prediction before global context modeling. All ablation variants share the same two-stage backbone, supervision, schedule, and evaluation protocol, and the fixed splits enable paired case-level comparisons at matched training budgets; the supplementary

Setting	A	F	M	D	G	Kvasir-SEG Jac.↑	Kvasir-SEG Dice↑	BUSI Jac.↑	BUSI Dice↑
w/o GECM						80.01	86.41	70.21	78.17
$A + D$	✓			✓		80.79	87.19	70.98	78.53
$A + F + D$	✓	✓		✓		81.04	87.38	71.15	78.72
$A + M + D$	✓		✓	✓		81.12	87.45	71.59	78.86
$A + F + M$	✓	✓	✓			80.13	86.87	70.49	78.43
$A + F + M + D$	✓	✓	✓	✓		81.20	87.58	71.97	79.05
All (Ours)	✓	✓	✓	✓	✓	**81.59**	**87.83**	**73.55**	**80.12**

Table 5: Ablation of GECM expert composition on Kvasir-SEG and BUSI. A, F, M, D, and G denote ACB, SFB, SDB, DINO branch, and DINO-guided gating. Rows with a DINO branch use calibrated priors.

Fusion strategy	Gate	Kvasir-SEG Jac.↑	Kvasir-SEG Dice↑	BUSI Jac.↑	BUSI Dice↑
Fixed sum	–	81.20	87.58	71.97	79.05
Learned scales	–	80.99	87.37	73.28	80.06
Gating w/o DINO	×	81.07	87.13	72.35	79.47
DINO-guided gating (Ours)	✓	**81.59**	**87.83**	**73.55**	**80.12**

Table 6: Ablation of GECM fusion and DINO-guided gating. All variants use calibrated priors; "Gating w/o DINO" omits them only from the gating input.

Variant	Kvasir-SEG Jac.↑	Kvasir-SEG Dice↑	BUSI Jac.↑	BUSI Dice↑	Parameters Train. (M)
DINOFT-NoDPC	77.69	85.47	70.11	77.93	33.20
DINO-LoRA r40	77.76	85.45	70.20	77.82	14.55
DINO-LoRA r64	76.76	84.71	70.69	78.69	16.32
DPC calibration (Ours)	**81.59**	**87.83**	**73.55**	**80.12**	**14.48**

Table 7: Comparison of DPC with direct DINO adaptation strategies on Kvasir-SEG and BUSI.

Injection strategy	Kvasir-SEG Jac.↑	Kvasir-SEG Dice↑	BUSI Jac.↑	BUSI Dice↑
No DINO information	80.05	86.79	70.35	78.51
GECM w/ raw prior	80.39	87.11	70.98	78.89
Encoder fusion	80.16	87.01	70.60	77.74
Decoder fusion	80.57	87.15	71.01	78.68
GECM w/ calibrated prior (Ours)	**81.59**	**87.83**	**73.55**	**80.12**

Table 8: Ablation of DINO injection strategies on Kvasir-SEG and BUSI.

additionally compares recent frozen-DINO transfer methods under shared protocols (Supplementary Section 2.2).

5 Limitations

DPCNet calibrates frozen DINOv3 priors without adapting the representation itself, and the current study covers two-dimensional binary tasks; for modalities far from natural-image statistics, or in volumetric and multi-class settings, the residual domain gap may limit gains. Calibration nevertheless transfers consistently across six diverse modalities and outperforms recent frozen-DINO transfer methods under shared protocols (Supplementary Section 2.2); broader deployment caveats are discussed in Supplementary Section 8. Because calibration is conditioned on the first-stage prediction, severe coarse errors could bias refinement. The coarse-prompt intervention analysis (Supplementary Section 4.4) shows that boundary shifts and high-confidence false positives degrade Dice by at most 1.53 pp, identifies missed targets as the dominant failure mode, and the reliability gate limits updates from unreliable logits. The two-stage design and frozen backbone increase inference cost relative to single-stage baselines, but end-to-end cost remains competitive with recent DINO-transfer methods (68.7 G MACs, 0.48 GB peak memory at batch size one; Supplementary Section 3).

6 Conclusion

We presented DPCNet, a target-specific calibration framework for transferring frozen DINOv3 priors to medical image segmentation. DPCNet uses first-stage coarse logits to calibrate the spatial response of the prior and introduces the calibrated representation into GECM as a selective global branch, complemented by image-derived structural and dependency features. Controlled ablations show that the calibrated prior provides gains beyond the two-stage refinement architecture and raw DINO transfer. The limitations above define the current scope: natural-image pretraining, coarse-prediction dependence, two-stage cost, and two-dimensional inputs. Future work will investigate more domain-aligned priors and volumetric calibration designs.

AI use statement

In this work, we used generative AI tools for language polishing and editing of the manuscript, such as improving grammar, clarity, and wording. We have not used generative AI tools for generating

research ideas, writing or modifying code, analyzing experimental results, creating figures or tables, or producing any scientific content, data, or claims; these tasks are not applicable to this work. We have no further uses with recommended disclosure to report. We have reviewed all AI-assisted text and confirmed that it accurately reflects our work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. doi: 10.1016/j.dib.2019.104863.

Saeid Asgari Taghanaki, Kumar Abhishek, Joseph Paul Cohen, Julien Cohen-Adad, and Ghassan Hamarneh. Deep semantic segmentation of natural and medical images: A review. *Artificial Intelligence Review*, 54(1):137–178, 2020. doi: 10.1007/s10462-020-09854-1.

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. doi: 10.1016/j.compmedimag.2015.02.007.

Afshin Bozorgpour, Sina Ghorbani Kolahi, Reza Azad, Ilker Hacihaliloglu, and Dorit Merhof. Cenet: Context enhancement network for medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2025*, pp. 120–129, 2025. doi: 10.1007/978-3-032-04927-8_12.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9630–9640, 2021. doi: 10.1109/ICCV48922.2021.00951.

Kerem Cekmeceli, Meva Himmetoglu, Guney I. Tombak, Anna Susmelj, Ertunc Erdil, and Ender Konukoglu. Do vision foundation models enhance domain generalization in medical image segmentation? In *Computer Vision – ECCV 2024 Workshops*, volume 15640 of *Lecture Notes in Computer Science*, pp. 185–200. Springer, 2024. doi: 10.1007/978-3-031-91672-4_12. URL https://doi.org/10.1007/978-3-031-91672-4_12.

Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, Matthew P. Lungren, Shaoting Zhang, Lei Xing, Le Lu, Alan Yuille, and Yuyin Zhou. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024. doi: 10.1016/j.media.2024.103280.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587, 2017. URL https://arxiv.org/abs/1706.05587.

Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging, hosted by the international skin imaging collaboration. In *2018 IEEE 15th International Symposium on Biomedical Imaging*, pp. 168–172, 2018. doi: 10.1109/ISBI.2018.8363547.

CREMI Challenge. Circuit reconstruction from electron microscopy images. https://cremi.org/, 2016. Official benchmark website.

Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, volume 12266 of *Lecture Notes in Computer Science*, pp. 263–273. Springer, 2020. doi: 10.1007/978-3-030-59725-2_26.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752, 2023. URL https://arxiv.org/abs/2312.00752.

Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: Achievements and challenges. *Journal of Digital Imaging*, 32(4):582–596, 2019. doi: 10.1007/s10278-019-00227-x.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnu-net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. doi: 10.1038/s41592-020-01008-z.

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, 2020. doi: 10.1007/978-3-030-37734-2_37.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, October 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html.

Haoyue Li, Yifan Gao, Feng Yuan, Xiaosong Wang, and Xin Gao. Dino U-Net: Exploiting high-fidelity dense features from foundation models for medical image segmentation. arXiv preprint arXiv:2508.20909, 2025a. URL https://arxiv.org/abs/2508.20909.

Xiangtai Li, Houlong Zhao, Lei Han, Yunhai Tong, Shaohua Tan, and Kuiyuan Yang. Gated fully fusion for semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11418–11425, 2020. doi: 10.1609/aaai.v34i07.6805.

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, 2025b. URL https://arxiv.org/abs/2509.02379.

Zhuonan Liang, Wei Guo, Jie Gan, Yaxuan Song, Runnan Chen, Hang Chang, and Weidong Cai. LUMOS: Latent universal medical priors for segmentation. arXiv preprint arXiv:2603.01115, 2026. URL https://arxiv.org/abs/2603.01115.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42: 60–88, 2017. doi: 10.1016/j.media.2017.07.005.

Zichen Luo, Xinshan Zhu, Lan Zhang, and Biao Sun. Rethinking u-net: Task-adaptive mixture of skip connections for enhanced medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5874–5882, 2025. doi: 10.1609/aaai.v39i6.32627.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15:654, 2024. doi: 10.1038/s41467-024-44824-z.

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*, pp. 11769–11779, 2024. doi: 10.1109/CVPR52733.2024.01118.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, 2015. doi: 10.1007/978-3-319-24574-4_28.

Jiacheng Ruan, Jincheng Li, and Suncheng Xiang. Vm-unet: Vision mamba unet for medical image segmentation. *ACM Transactions on Multimedia Computing, Communications, and Applications*, pp. 3767748, 2025. doi: 10.1145/3767748.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michaël 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, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=2NlGyqNjns.

J. Staal, M. D. Abramoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, 23(4):501–509, 2004. doi: 10.1109/TMI.2004.825627.

Guankun Wang, Rui Tang, Mengya Xu, Long Bai, Huxin Gao, and Hongliang Ren. Endoarss: Adapting spatially aware foundation model for efficient activity recognition and semantic segmentation in endoscopic surgery. *Advanced Intelligent Systems*, 7(12):e202500288, 2025. doi: 10.1002/aisy.202500288. URL https://doi.org/10.1002/aisy.202500288.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, 2018. doi: 10.1109/CVPR.2018.00813.

Sicheng Yang, Hongqiu Wang, Zhaohu Xing, Sixiang Chen, Qiuxia Yang, Yize Mao, Guang Yang, and Lei Zhu. Segdino: Introducing multi-scale structure into dino for efficient medical image segmentation. arXiv preprint arXiv:2606.17972, 2026. URL https://arxiv.org/abs/2606.17972.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6230–6239, 2017. doi: 10.1109/CVPR.2017.660.

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*, pp. 20852–20862, 2025. doi: 10.1109/CVPR52734.2025.01942.

SUPPLEMENTARY MATERIAL

This supplementary material provides implementation details, controlled experiments, and additional analyses for DPCNet. Unless stated otherwise, all analyses use case-level aggregation so that features from multiple scales or blocks are not treated as independent samples.

1 REPRODUCIBILITY AND IMPLEMENTATION DETAILS

Unless otherwise stated, DPCNet and its ablation variants use a frozen DINOv3 ViT-S/16 backbone with patch size 16 and hidden dimension 384. We extract features from Transformer blocks $\{2, 5, 8, 11\}$. Single-channel images are repeated to three channels before DINO inference, mapped to $[0, 1]$, and normalized with the ImageNet mean and standard deviation. DINOv3 and its input adapter remain frozen throughout training.

Let f_0, f_1, f_2, f_3 denote the extracted features. They are projected by 1×1 convolutions and fused top-down by 3×3 convolutions to construct $\mathcal{P} = \{P^s\}_{s\in\mathcal{S}}$. The five spatial scales have channel dimensions $\{8, 16, 32, 64, 128\}$ at resolutions $\{H, W\}$, $\{H/2, W/2\}$, $\{H/4, W/4\}$, $\{H/8, W/8\}$, and $\{H/16, W/16\}$, respectively. A separate 1×1 projection of f_3 produces a 256-channel bottleneck prior. Thus, $\mathcal{S} = \{0, 1, 2, 3, 4, b\}$ indexes six prior tensors. All resizing uses bilinear interpolation with `align_corners=False`.

The two U-Net stages are trained jointly for 200 epochs with

$$\mathcal{L} = \mathcal{L}_{\text{refined}} + \mu\mathcal{L}_{\text{coarse}}, \qquad \mu = 0.2. \tag{6}$$

Both terms use Dice and cross-entropy losses. The refined loss uses deep-supervision weights $(8, 4, 2, 1, 0)/15$. Coarse logits are detached only on the DPC path; the coarse loss still supervises the first stage. We use SGD with Nesterov momentum 0.99, weight decay 3×10^{-5}, initial learning rate 5×10^{-3}, and the nnU-Net polynomial schedule. We set $\rho = 1$, $\beta = 1$, initialize $\boldsymbol{\lambda}_s$ to 0.001, and use a 40-epoch linear warm-up. GECM residual scales are initialized to zero.

The reported 14.48M parameters are trainable. The frozen DINOv3 backbone adds 21.60M parameters at inference, giving 36.08M total parameters.

For repeated main comparisons, we report the mean and standard deviation over three random seeds. BUSI is evaluated by five-fold cross-validation, and its reported standard deviation is computed across folds. For paired comparisons, each test case is treated as one statistical unit; intermediate feature maps or scales are not treated as independent samples.

1.1 RUN CONFIGURATION RECORD

For each run, we record the dataset split, nnU-Net plan, input patch size, batch size, augmentation configuration, random seed, checkpoint identifier, checkpoint-selection rule, software versions, and execution command. The selected checkpoint has the highest validation Dice under the corresponding validation protocol. This record makes the DINO input resolution and dataset-specific preprocessing explicit.

1.2 CROSS-STAGE FEATURE TRANSFER

The second stage receives intermediate features from the first-stage decoder in addition to the calibrated DINO priors. Let $X^0_{1,d}$ denote the full-resolution output of the first-stage decoder, and let $\{G^s_{1,d}\}_{s=1}^4$ denote the outputs of the decoder-side GECMs at the four lower-resolution decoder scales. The encoder-side GECM outputs are used by the first-stage decoder through its U-Net skip connections; they are not transferred directly to the second stage.

At the second-stage input, the image feature and $X^0_{1,d}$ are concatenated along the channel dimension. At each subsequent encoder scale $s \in \{1, 2, 3, 4\}$, the downsampled second-stage feature is concatenated with $G^s_{1,d}$ before it is processed by the corresponding second-stage encoder block and GECM. In compact form, the transfer is written as

$$\begin{aligned} X^0_{2,e} &= E^0_2\big([x, X^0_{1,d}]\big), \\ X^s_{2,e} &= E^s_2\big([\mathrm{Down}(X^{s-1}_{2,e}), G^s_{1,d}]\big), \qquad s = 1, \ldots, 4, \end{aligned} \tag{7}$$

Setting	Stages	Image experts	DINO branch	GECM gate	DPC	Kvasir-SEG		BUSI	
						Jac.↑	Dice↑	Jac.↑	Dice↑
Single-stage U-Net	1	–	None	–	–	78.53	85.62	68.23	75.36
Two-stage U-Net	2	–	None	–	–	80.01	86.41	70.21	78.17
Two-stage + image experts + gate	2	✓	None	✓	–	80.05	86.79	70.35	78.51
Two-stage + full GECM + raw DINO	2	✓	Raw P^s	✓	–	80.39	87.11	70.98	78.89
DPCNet (Ours)	2	✓	Calibrated $\hat{P}^s$	✓	✓	81.59	87.83	73.55	80.12

Table 9: Controlled two-stage ablation. Kvasir-SEG uses the fixed 880/120 split, and BUSI uses the fixed fold-0 split. Image experts denote ACB, SFB, and SDB. The two-stage baseline removes GECM and DINO priors. The image-expert setting retains ACB, SFB, and SDB without DINO priors while retaining the GECM gate. The raw-DINO setting uses the complete GECM with uncalibrated DINO priors, whereas DPCNet uses calibrated priors. "–" denotes an omitted component; DPC is additionally not applicable to the single-stage model.

Method	DINO use	Kvasir		BUSI		Reported Params.
		Jac.↑	Dice↑	Jac.↑	Dice↑	(M)
LUMOS	Frozen spatial guidance	86.69	90.02	70.92	77.36	7.38
Dino U-Net	Frozen dense transfer	85.35	89.47	69.70	74.92	18.14
SegDINO	Frozen multi-scale decoding	88.95	92.51	71.14	78.46	24.90
DPCNet (Ours)	Coarse-calibrated prior + GECM	**90.75**	**94.18**	**72.61**	**80.04**	14.48

Table 10: Comparison with recent methods that transfer frozen DINO representations. Kvasir follows the joint CVC-ClinicDB and Kvasir-SEG training protocol, whereas BUSI uses 5-fold cross-validation. Reported parameter counts follow the conventions of the corresponding implementations and should not be interpreted as a normalized end-to-end inference-cost comparison.

where $[\cdot,\cdot]$ denotes channel-wise concatenation, $\mathrm{Down}(\cdot)$ is the stage-two downsampling operation, and E_2^s denotes the encoder computation at scale s. The concatenated features are processed by the existing stage-two convolutions; no element-wise addition is used for cross-stage fusion.

No stop-gradient operation is applied to these transferred features. Therefore, the refined segmentation loss propagates through $X_{1,d}^0$ and $\{G_{1,d}^s\}_{s=1}^4$ into the first-stage decoder and its preceding first-stage blocks. The only detached first-stage quantity is z_1 when it enters DPC. The coarse loss continues to supervise the first-stage prediction directly.

2 Additional Controlled Experiments

2.1 Controlled Two-Stage Ablation

Table 9 separates the capacity of coarse-to-refine processing from the contributions of image experts, the DINO branch, and DPC. All two-stage rows use the same U-Nets, cross-stage connections, coarse supervision, loss, schedule, and inference procedure. DPC is not applicable to the single-stage row because it uses the first-stage logits to calibrate priors for a second stage. The last two rows use the full GECM and differ only in whether the DINO branch receives raw P^s or calibrated $\hat{P}^s$ priors.

2.2 Comparison with Recent DINO Transfer Methods

We compare DPCNet with recent methods that transfer frozen DINO representations: LUMOS, Dino U-Net, and SegDINO. All methods use the same Kvasir-SEG and BUSI splits, preprocessing, and evaluation protocol whenever their official implementations permit it. Table 10 reports the resulting comparison under these shared protocols.

3 Inference Cost

Table 11 reports end-to-end inference cost, including the frozen DINOv3 backbone, at batch size one.

Method	Trainable Params. (M)	Total Params. (M)	Total MACs (G)	Peak Memory (GB)	Latency (ms)
nnU-Net	46.33	46.33	466.25	0.40	14.57
nnWNet	7.04	7.04	42.81	0.34	40.64
SegDINO	24.90	24.90	80.30	0.80	23.74
DPCNet (Ours)	14.48	36.08	68.71	0.48	29.27

Table 11: End-to-end inference cost at batch size one. Total MACs, memory, and latency include the frozen DINOv3 backbone. The input size is 512×512 for nnU-Net, nnWNet, SegDINO, and DPCNet. All measurements are conducted on a single NVIDIA H20 GPU after warm-up.

Coarse-conditioning variant	Kvasir Jac.	Kvasir Dice
Direct raw DINO prior	80.39	87.11
Foreground map F only	81.08	87.44
Foreground confidence $F \odot Q$	80.87	87.43
Reliability gate Q only	80.64	87.32
Full DPC with F, U, and B	81.59	87.83

Table 12: Controls for coarse-conditioned prior refinement on the fixed Kvasir-SEG split. Direct raw DINO prior bypasses DPC, whereas the remaining rows use the indicated coarse-conditioning signal with the same two-stage backbone, GECM, and training schedule.

4 Detailed DINO Prior Calibration

4.1 Coarse Prediction Maps and Reliability Gate

Let $z_1 \in \mathbb{R}^{B\times 1\times H\times W}$ be the first-stage coarse logits. DPC constructs

$$\begin{aligned} F &= \sigma(z_1), & U &= 1 - |2F - 1|, \\ B &= \mathrm{clip}(D_x F + D_y F, 0, 1), & S &= [F, U, B], \end{aligned} \tag{8}$$

where $D_x F_{i,j} = |F_{i,j} - F_{i,j-1}|$ and $D_y F_{i,j} = |F_{i,j} - F_{i-1,j}|$; missing entries are zero-padded. The reliability gate is

$$Q = \mathrm{clip}\left(|2F - 1|^{\rho}(1 + \beta B), 0, 1\right). \tag{9}$$

The uncertainty map U describes ambiguous regions for learned refinement, while Q limits strong updates from unreliable coarse logits. Since Q is symmetric for confident foreground and background predictions, it is a reliability gate rather than a foreground selector.

4.2 Scale-Wise Residual Calibration

For each $s \in \mathcal{S}$, S and Q are resized to the spatial size of P^s. DPC computes

$$\begin{aligned} A^s &= \mathrm{GELU}(\mathrm{BN}(W_p^s * S^s)), \\ G^s &= \sigma(W_g^s * A^s + b_g^s) \odot Q^s, \\ R^s &= \mathrm{GELU}(\mathrm{BN}(W_r^s * [P^s, A^s])), \\ \hat{P}^s &= P^s + \eta_e \boldsymbol{\lambda}_s \odot G^s \odot R^s. \end{aligned} \tag{10}$$

W_p^s and W_r^s are bias-free 3×3 convolutions, and W_g^s is a 1×1 convolution with bias. The sigmoid gate is independent for each channel. $\boldsymbol{\lambda}_s$ is channel-wise and $\eta_e = \min(1, e/40)$. The signed response R^s can correct the prior in either direction.

4.3 Alternative Coarse-Conditioning Controls

The reliability gate Q is not a foreground selector. To distinguish the contribution of the learned DPC correction from simpler prompt choices, we define the controls in Table 12. All variants retain the same two-stage backbone, GECM, and training schedule; only the coarse-conditioning signal is changed.

4.4 Coarse-Prompt Intervention Analysis

We perform controlled interventions on the coarse prompt supplied to DPC. Only the stage-one logits entering the DPC bridge are replaced; all model weights, transferred features, and other

Coarse-prompt condition	Coarse Dice (%)	Kvasir Jac. (%)	Kvasir Dice (%)	Δ Dice (pp)
Predicted coarse logits	85.01	81.59	87.83	0.00
Boundary-shift perturbation	79.88	81.14	87.43	-0.40
Missed-target perturbation	1.20	68.93	75.95	-11.88
High-confidence false-positive perturbation	74.94	80.06	86.30	-1.53
Oracle F, U, and B maps	98.27	82.17	89.43	+1.60

Table 13: Coarse-prompt intervention analysis on the Kvasir-SEG fixed validation split ($n = 120$). Each condition replaces only the coarse logits entering the DPC bridge; all model weights and other inference computations are fixed. Δ Dice is measured relative to predicted coarse logits. This table tests prompt effects with fixed cross-stage features; it is not a comparison of models with different architectural components.

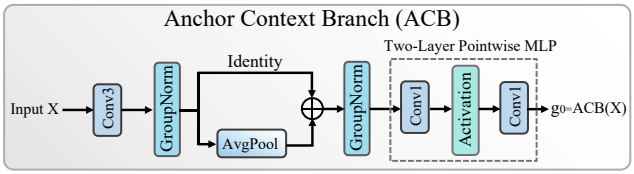


Figure 4: Detailed structure of the anchor context branch (ACB).

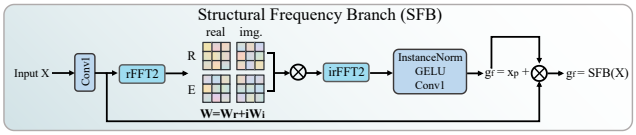


Figure 5: Detailed structure of the structural frequency branch (SFB).

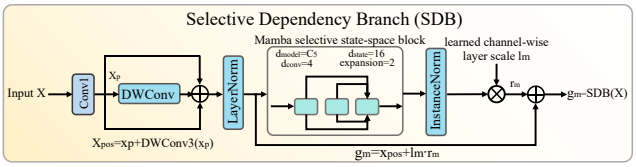


Figure 6: Detailed structure of the selective dependency branch (SDB).

inference computations remain fixed. We test boundary-shift, missed-target, and high-confidence false-positive perturbations, together with an oracle prompt constructed from ground-truth-derived foreground, uncertainty, and boundary maps.

As shown in Table 13, the refined prediction changes with the prompt condition. Boundary shifts decrease Dice by 0.40 pp, while missed-target and false-positive prompts decrease Dice by 11.88 and 1.53 pp, respectively. In contrast, the oracle prompt improves Dice by 1.60 pp. The larger decrease under the missed-target condition indicates that DPC depends on the target information provided by the coarse prompt. Since this intervention retains the image pathway and cross-stage feature transfer, it does not attribute the full performance gain to DPC alone. The controlled two-stage ablation in Table 9 separately compares raw and calibrated DINO priors under the same two-stage backbone and GECM configuration.

5 Detailed GECM Structure

At scale s, GECM receives U-Net feature $X \in \mathbb{R}^{B\times C_{\text{in}}\times H_s\times W_s}$ and prior $\tilde{P}^s$. It uses raw P^s in stage one and calibrated $\hat{P}^s$ in stage two. The three image-based expert features are

$$g_0 = \text{ACB}(X), \qquad g_f = \text{SFB}(X), \qquad g_m = \text{SDB}(X). \tag{11}$$

Figures 4, 5, and 6 show the detailed computations of the ACB, SFB, and selective dependency branch (SDB), respectively.

Anchor context branch. ACB is the original WNet global block. It uses a 3×3 projection, GroupNorm, a 3×3 average-pooling residual response, a second GroupNorm layer, and a two-layer 1×1 MLP with expansion ratio 4.

Algorithm 1 Forward pass of DPCNet.

Require: Image x, frozen DINOv3 backbone, two U-Net stages
Ensure: Refined logits z_2
1: Extract DINOv3 features from blocks $\{2, 5, 8, 11\}$ and build raw priors $\mathcal{P}$.
2: Run stage one with raw P^s in GECM and obtain z_1.
3: Construct F, U, B, Q from $\mathrm{stopgrad}(z_1)$ using Eqs. equation 8–equation 9.
4: Compute $\hat{\mathcal{P}}$ using Eq. equation 10 and transfer stage-one features.
5: Run stage two with $\hat{P}^s$ in GECM and return z_2.

Structural frequency branch. SFB projects X to C_s channels and applies an orthonormal 2D real FFT. Learnable real and imaginary weights are interpolated to the current frequency grid and applied to the spectrum. An inverse real FFT restores a real-valued map, followed by InstanceNorm, GELU, and a 1×1 convolution. No padding is applied before the FFT.

Selective dependency branch. SDB uses a 1×1 projection, depth-wise 3×3 positional encoding, sequence flattening, Mamba, reshaping, and InstanceNorm. Mamba uses $d_{\mathrm{state}} = 16$, convolution width 4, and expansion factor 2.

DINO branch and fusion. Let $d = \phi_d(\tilde{P}^s)$ and $x_r = \phi_x(X)$. The controller first forms an intermediate feature h, then predicts the branch gates:

$$\begin{aligned} h &= \mathrm{GELU}(\mathrm{BN}(W_1 * [x_r, d, g_0, g_f, g_m])) \,, \\ [\gamma_f, \gamma_m, \gamma_d] &= \sigma(W_2 * h + b_2) \,, \\ Y &= g_0 + \alpha_f \gamma_f \odot g_f + \alpha_m \gamma_m \odot g_m \\ &\quad + \alpha_d \gamma_d \odot d. \end{aligned} \tag{12}$$

The controller maps $5C_s$ to $\max(C_s/2, 4)$ and then to $3C_s$ channels using 1×1 convolutions, batch normalization, GELU, and sigmoid. The channel-wise scales α_f, α_m, and α_d are initialized to zero.

6 End-to-End Forward Procedure

7 DPC Response Analysis

The response statistics used in the main paper are defined below. The gate-only response $P^s \odot Q^s$ characterizes the spatial selection induced by the coarse-prompt gate, whereas the final-prior response measures $\hat{P}^s$ after the signed residual update.

For a response tensor $T^s \in \mathbb{R}^{C_s \times H_s \times W_s}$, we first form a nonnegative response map $a_T^s = \mathrm{Resize}(\|T^s\|_2)$ by taking the channel-wise ℓ_2 norm and bilinearly resizing it to the ground-truth resolution. No response threshold, channel normalization, or per-image rescaling is applied. Let M and V denote the binary ground-truth mask and valid image region, respectively. With a three-pixel morphological radius, the target interior I is the eroded mask $\mathrm{Erode}_3(M) \cap V$ (or $M \cap V$ when erosion is empty), the boundary band is $D = (\mathrm{Dilate}_3(M) \setminus \mathrm{Erode}_3(M)) \cap V$, and the far background is $O = V \setminus \mathrm{Dilate}_3(M)$. Thus, far background consists of valid pixels more than three pixels from the target boundary.

For each scale and case, we compute

$$\begin{aligned} \mathrm{F/B}(a) &= \frac{\mathrm{mean}_I(a) + \epsilon}{\mathrm{mean}_O(a) + \epsilon}, \\ \mathrm{B/Bg}(a) &= \frac{\mathrm{mean}_D(a) + \epsilon}{\mathrm{mean}_O(a) + \epsilon}, \\ \mathrm{Leak}(a) &= \frac{\sum_{p \in O} a(p)}{\sum_{p \in V} a(p) + \epsilon}, \end{aligned} \tag{13}$$

where $\epsilon = 10^{-8}$. Metrics are first averaged over the six prior scales or the 12 GECM blocks within each case; paired tests then use cases as the statistical units.

For the diagnostic gate-only response, Q^s denotes the coarse-prompt gate resized to scale s, and the measured tensor is $P^s \odot Q^s$. This quantity reflects the multiplicative selection imposed by Q^s; the final-prior response instead measures $\hat{P}^s$ after the signed residual update in Eq. equation 10.

8 Limitations and Responsible Use

DPCNet is evaluated on public two-dimensional binary medical image segmentation tasks. The current evidence does not establish performance on multi-class, three-dimensional, cross-device, or clinical deployment settings. Because DPC conditions prior refinement on the first-stage prediction, severe coarse-prediction errors may affect later refinement. The reported perturbation analysis examines selected controlled conditions, but does not establish robustness to all failure modes. The model is not intended to make clinical decisions without expert review. Dataset annotation practices, population coverage, acquisition devices, and computational cost may also limit transfer to new settings.