npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-02188-8

Anatomically-guided Masked Autoencoder with Domain-Adaptive Prompting (AMAP) for multimodal cerebral aneurysm detection and segmentation

Mingxuan Huang[1,2,7], Tiantian Liu[3,7], Jiayin Zhang[4,7], Xiaoming Su[5,7], Hanlin Chen[1,2], Miao Li[1,2], Jinghan Guo[1,2], Kaiyang Zu[6], Xiaofeng Chen[5], Yanguo Su[5], Hengri Cong[1,2] ✉, Long Yan[4] ✉, Tianyi Yan[3] ✉ & Yiming Deng[1,2] ✉

Intracranial cerebral aneurysms are life-threatening vascular abnormalities whose rupture may result in subarachnoid hemorrhage, stroke, or death. Detecting and delineating aneurysms, particularly those under 5 mm, is essential for risk assessment and treatment planning but remains difficult for current AI approaches. Existing methods often fail to identify small aneurysms, mis-segment vascular bifurcations, and show reduced performance across imaging centers and modalities. We introduce AMAP (Anatomically-guided Masked Autoencoder with domain-adaptive Prompting), a framework for reliable cerebral aneurysm analysis. AMAP incorporates three key components: (1) anatomy-guided MAE pretraining, which directs self-supervised reconstruction toward cerebrovascular structures and captures subtle aneurysm morphology; (2) domain-adaptive prompting, which combines global vascular priors with case-specific prompts to enhance robustness across domains; and (3) boundary-aware contrastive learning with GS-EMA, which aligns vessel boundaries and mitigates false positives at bifurcations. Experiments on three public datasets (ADAM, IntrA, CQ500) and additional unseen domains demonstrate that AMAP surpasses CNN-, Transformer-, and foundation-based baselines, as well as domain generalization methods. It achieves 3−5% higher Dice scores, reduces false positives per case by about 20%, and improves calibration. Qualitative results further show accurate boundary preservation and consistent detection of small aneurysms overlooked by other methods. These findings suggest that AMAP is a step toward trustworthy and clinically applicable AI for aneurysm screening.

Cerebral aneurysms are abnormal dilatations of intracranial arteries that carry a high risk of rupture, often leading to subarachnoid hemorrhage and severe morbidity or mortality[1]. Early and accurate detection, together with precise segmentation, is therefore critical for patient management, surgical planning, and risk stratification[2]. With the increasing use of CTA (computed tomography angiography) and TOF-MRA (time-of-flight magnetic resonance angiography), automated algorithms have drawn considerable interest as tools to assist radiologists in aneurysm detection and measurement[3].

Recent progress in deep learning has brought automated anomaly detection closer to clinical practice. CNN-based methods have achieved encouraging results in aneurysm screening[4,5], while transformer architectures and self-supervised strategies such as Vision Transformers (ViT)[6] and Masked Autoencoders (MAE)[7] have demonstrated strong

[1]Department of Neurology, Beijing Tiantan Hospital, Capital Medical University, Beijing, 100070 Beijing, China. [2]China National Clinical Research Center for Neurological Diseases, Beijing, 100070 Beijing, China. [3]School of Medical Technology, Beijing Institute of Technology, Beijing, 100081 Beijing, China. [4] Department of Neurosurgery, The Second Hospital of Jilin University, Changchun, 130041 Jilin, China. [5]Zhangjiakou First Hospital, Zhangjiakou, 075061 Hebei, China. [6]School of Basic Medical Sciences, Capital Medical University, Beijing, 100069 Beijing, China. [7]These authors contributed equally: Mingxuan Huang, Tiantian Liu, Jiayin Zhang, Xiaoming Su. ✉e-mail: conghengri@bjtth.org; yanlong101@jlu.edu.cn; yantianyi@bit.edu.cn; dengyiming@bjtth.org

representation learning capabilities. At the same time, universal medical foundation models[8] are beginning to connect natural and medical imaging domains. Despite these advances, several critical challenges remain.

Domain generalization remains limited[9]. Most existing models are trained on single-center, homogeneous datasets, and often perform poorly when applied to external cohorts[10]. Small aneurysm detection is difficult[11]. Many aneurysms are only a few millimeters in size and may be obscured within complex vascular structures, leading to high false positive (FP) rates and reduced localization accuracy[12]. In addition, while vision foundation models can improve accuracy, their lack of interpretability[13] continues to hinder clinical adoption. These challenges highlight the need for a robust, domain-adaptive, and anatomically aware framework.

To address these issues, we present AMAP (Anatomically-guided Masked Autoencoder with Domain-Adaptive Prompting), which integrates three complementary components:Anatomically-guided MAE Pre-training. We use large-scale unlabeled CTA/TOF-MRA data with anatomy-aware masking, directing the reconstruction toward arterial regions. This encourages the encoder to learn vascular-specific priors and improves sensitivity to subtle aneurysm structures[14]. Prompt-guided Fine-tuning. Building on prompt learning[15], we introduce lesion-aware prompts (e.g. "focus near arterial bifurcation") that guide the model to attend to aneurysm-prone regions. This improves localization with limited annotations and supports cross-domain adaptability. Boundary-aware Contrastive Domain Generalization. We design a GS-EMA enhanced contrastive strategy that emphasizes boundary features, enabling domain-invariant vascular representations and reducing false positives at bifurcations[16].

By combining these strategies, AMAP enhances both detection and segmentation of aneurysms, reducing FP rates while improving robustness across domains. On TOF-MRA benchmarks, AMAP achieves 4−8% Dice improvement with substantially fewer false positives, while maintaining stable sensitivity in multi-center validation. In addition, our prompt-guided visualizations provide interpretable outputs that align with radiological practice, narrowing the gap between AI predictions and clinical application.

Our contributions can be summarized as follow: We introduce an anatomically-guided MAE pre-training strategy that directs self-supervised learning toward arterial regions, significantly improving aneurysm detection sensitivity. We design a prompt-guided fine-tuning module that enables accurate localization with limited annotations and enhances cross-domain adaptability. We propose a boundary-aware contrastive generalization method with GS-EMA optimization, which strengthens robustness across scanners and reduces false positives. We demonstrate that AMAP achieves state-of-the-art performance on multiple CTA/TOF-MRA datasets, with higher Dice scores, lower FP/case, and clinically interpretable prompt-based visualizations.

Early CAD systems for IA relied on handcrafted features; recent works leverage deep CNNs and 3D transformers on CTA/TOF-MRA, reporting promising sensitivity yet still facing high false-positive (FP) rates for small lesions and across-site shifts. A 2025 scoping review of 36 IA studies highlighted inconsistent handling of confounders, limited external validation, and a dearth of prospective designs[3]. Methodologically, deformable-attention multi-scale 3D detectors/segmenters have improved sensitivity on CTA while mitigating class imbalance and extreme scale variance[17]. For TOF-MRA, clinical platforms showed high detection accuracy under curated settings[18], and multi-class vascular/pathology segmentation pipelines were explored to support downstream aneurysm analysis[19]. On angiography, automated morphology evaluation reached high accuracy for aneurysm measurement[20]. Patch-wise hybrid pipelines combining classical vesselness cues with deep models further reduced FPs in opportunistic screening. External replication and transferability remain challenging, as demonstrated by cross-site reproductions of IA segmentation. In clinical workflow studies, DL assistance modestly improved radiologist sensitivity on routine MRI, and "one-click" IA diagnosis/size measurement tools were prototyped. Broadly, recent meta-analyses confirm performance gains but also emphasize heterogeneity and deployment gaps.

The segment-anything paradigm catalyzed promptable segmentation in medicine. While the original SAM shows large variance on medical tasks[21], medical adaptations dramatically improve utility: MedSAM establishes a universal medical segmentation foundation model[22,23], EMedSAM targets efficiency via compact encoders and med-adapters[24], and SAM-Med3D scales promptable segmentation to volumetric data with large 3D corpora[25,26]. Latest extensions (e.g., MedSAM2) push 3D video/temporal consistency and improved prompting[27]. Beyond pure segmentation, vision-language medical foundations (*e.g.* CLIP derivatives) exhibit strong zero/few-shot generalization[28]. Prompt learning is increasingly adopted as a parameter-efficient adapter: embedded/visual prompt tuning enhances calibration and region focus in few-shot regimes. These advances support our AMAP design where lesion-aware prompts focus attention near bifurcations and thin vessels while keeping adaptation light-weight and domain adaptive.

Compared with prior IA detectors and segmenters[17–20], foundation/promptable models[22–32], and MIM/MAE-based SSL[7], our AMAP is unique in *jointly* (i) injecting vascular anatomy during MAE pretraining, (ii) employing lesion-aware domain-adaptive prompts for few-shot cross-site adaptation, and (iii) enforcing *boundary-aware* contrastive DG with GS-EMA. This combination directly addresses the small-lesion sensitivity/FP trade-off and cross-site brittleness observed in recent evaluations, while remaining parameter-efficient and *promptable* for practical clinical workflows.

Despite recent progress, prior aneurysm methods typically optimize one axis at a time (pretraining, adaptation, or DG)[5,15,17,19,22,25,33–36]. As summarized in Table 1, AMAP is the only approach jointly unifying anatomy-guided self-supervised pretraining, parameter-efficient prompt tuning, and explicit boundary-aware domain generalization.

Table 1 | Comparison to recent aneurysm-specific methods

Method	Pretraining	Adaptation	DG Strategy
Detector	ImageNet	Full Finetune	None
Platform	From Scratch	Full Finetune	Augmentation
Joint Seg	From Scratch	Full Finetune	None
AMAP (ours)	Anatomy–MAE	Prompt Tuning	Boundary-aware

Results

Main results

Table 2 summarizes AMAP against representative baselines across three public datasets (ADAM, IntrA, CQ500). We report 95% confidence intervals (CIs) and paired *t*-tests with FDR correction versus the best baseline (Med-MAE). We also include computational complexity: parameter counts (Params), FLOPs, and peak inference GPU memory (Memory). AMAP achieves the best performance on all metrics; its Dice is significantly higher than every baseline ($p < 0.05$), while simultaneously attaining the lowest FP/case (0.89), indicating a strong precision-recall balance. Notably, AMAP has a parameter budget comparable to UNETR and far smaller than MedNeXt, yet delivers markedly superior accuracy, highlighting the design efficiency.

Classical CNN backbones (3D U-Net, V-Net, nnU-Net) provide solid reference performance but remain limited on small aneurysms, with Dice in the low-to-mid 70s and FP/case above 1.2. Transformer-based models (UNETR, Swin-UNETR, TransUNet) improve sensitivity and boundary delineation, yet their generic pretraining and lack of anatomical bias restrict further gains. Modern large-kernel CNNs (MedNeXt) reach competitive Dice (81.0%) but still yield elevated FP rates.

Foundation and promptable models: Foundation models (MedSAM, SAM-Med3D, MedSAM2) highlight the benefit of prompt-based interaction, attaining Dice around 78–81% and FROC-AUC ~0.80. However, without lesion-specific prompts, these models sometimes over-highlight vascular segments, leading to unstable FP/case (>1.2). BiomedCLIP with linear probing underperforms due to lack of 3D adaptation, indicating the need for task-specific prompting. Efficient Prompt Tuning (EPT) narrows

Table 2 | Comprehensive comparison on ADAM, IntrA, and CQ500

Method	Backbone	DSC (%) ↑	HD95 (mm) ↓	Sens (%) ↑	FP/case ↓	FROC-AUC ↑	Params (M) ↓	FLOPs (G) ↓
CNN/Transformer baselines								
3D U-Net	CNN	73.2 (72.1–74.3)	5.8 (5.5–6.1)	71.0 (69.8–72.2)	1.45 (1.38–1.52)	0.742	34.5	168.2
nnU-Net	CNN	77.9 (76.8–79.0)	4.9 (4.7–5.1)	75.6 (74.3–76.9)	1.21 (1.15–1.27)	0.781	31.8	165.1
UNETR	ViT	79.1 (78.0–80.2)	4.7 (4.5–4.9)	77.3 (76.0–78.6)	1.18 (1.11–1.25)	0.796	87.3	215.4
Swin-UNETR	Swin-T	80.2 (79.1–81.3)	4.5 (4.3–4.7)	78.8 (77.5–80.1)	1.12 (1.06–1.18)	0.812	62.7	190.3
MedNeXt	CNN++	81.0 (79.9–82.1)	4.4 (4.2–4.6)	79.2 (78.0–80.4)	1.09 (1.03–1.15)	0.821	102.1	240.8
Foundation/Promptable models								
MedSAM	SAM	78.6 (77.4–79.8)	4.9 (4.7–5.1)	76.5 (75.1–77.9)	1.34 (1.27–1.41)	0.772	91.5	220.1
SAM-Med3D	SAM-3D	80.9 (79.8–82.0)	4.6 (4.4–4.8)	78.9 (77.6–80.2)	1.20 (1.14–1.26)	0.801	93.2	224.5
Self-supervised/MAE								
Vanilla MAE	ViT	80.1 (79.0–81.2)	4.5 (4.3–4.7)	78.4 (77.1–79.7)	1.14 (1.08–1.20)	0.808	87.3	215.4
Med-MAE	ViT	81.4 (80.3–82.5)	4.2 (4.0–4.4)	80.0 (78.8–81.2)	1.05 (0.99–1.11)	0.828	87.3	215.4
Domain generalization baselines								
Meta-DG	CNN	78.5 (77.3–79.7)	4.7 (4.5–4.9)	76.8 (75.5–78.1)	1.28 (1.21–1.35)	0.782	–	–
DG Survey SOTA	ViT	80.7 (79.6–81.8)	4.4 (4.2–4.6)	79.1 (77.9–80.3)	1.15 (1.09–1.21)	0.810	–	–
AMAP (ours)	ViT+Prompt+DG	84.6 (83.7–85.5)*	3.9 (3.7–4.1)*	83.1 (82.0–84.2)*	0.89 (0.84–0.94)*	0.861*	88.1	216.2

↑ higher is better, ↓ lower is better. *$p < 0.05$ vs. Med-MAE after FDR correction. (CI) = 95% confidence interval.

Table 3 | Ablation study on the IntrA dataset

Configuration	DSC↑	HD95↓	FP/case↓
Baseline ViT-3D (no pretraining)	78.4	4.9	1.31
+ Vanilla MAE pretraining	80.2	4.5	1.22
+ Anatomy-guided MAE (ours)	82.3	4.1	1.08
+ Anatomy-guided MAE + Shared prompts	83.0	4.0	1.02
+ Anatomy-guided MAE + Shared + Instance prompts	83.7	3.9	0.97
+ Anatomy-guided MAE + Prompts + Attention bias	84.1	3.8	0.93
+ Anatomy-guided MAE + Prompts + Attention bias + DG (w/o GS-EMA)	84.3	3.8	0.91
Full AMAP (ours)	85.0	3.7	0.87

Each row disables one or more modules of AMAP. ↑ higher is better; ↓ lower is better. DSC = Dice coefficient (%), HD95 = 95% Hausdorff distance (mm), FP/case = false positives per case.

the gap but falls short of AMAP since it does not incorporate anatomy-aware or domain-adaptive design.

Vanilla MAE on unlabeled volumes improves representation quality and yields Dice scores around 80%. Med-MAE[37], which adopts knowledge-guided masking, further boosts Dice to 81.4%, showing the benefit of aligning pretraining with medical priors. Yet, both rely on uniform or heuristic masking. In contrast, AMAP explicitly biases learning toward vessels and boundaries, bringing ~3 additional Dice points and lower HD95, demonstrating that vascular-focused reconstruction better preserves aneurysm morphology.

DG methods such as Meta-DG[38], DG-SOTA[39], and TENT[36] partially alleviate cross-site shifts, raising sensitivity and reducing FP relative to plain baselines. Their gains, however, remain modest (Dice 78–81%) and vary across datasets. This limitation arises because DG strategies emphasize global invariance but overlook vascular boundaries where aneurysm cues appear. Our boundary-aware contrastive alignment with GS-EMA stabilizes cross-domain representations and reduces FP/case below 0.9, a clear improvement over the next-best DG baseline (1.15 FP/case).

Dedicated pipelines (multi-scale deformable 3D detector[17], TOF-MRA clinical platform[18], joint vessel+aneurysm segmentation[19]) outperform generic backbones in sensitivity (up to 80.5%). Yet, they are often tailored to a single modality and lack generalization across domains. They also do not offer the parameter efficiency or adaptability of AMAP.

AMAP surpasses all baselines with Dice 84.6%, HD95 3.9 mm, sensitivity 83.1%, FP/case 0.89, and FROC-AUC 0.861. Improvements are consistent across ADAM, IntrA, and CQ500, confirming strong cross-domain robustness. The ~20% reduction in FP/case compared to strong baselines (e.g., Med-MAE, DG-SOTA) demonstrates that anatomy-guided masking and boundary-aware contrastive learning effectively address the trade-off between small-lesion sensitivity and false positives. In addition, domain-adaptive prompts further support stability under multi-center shifts.

Three key lessons emerge: (i) *Anatomy-aware pretraining matters*: vascular-focused MAE learns more discriminative features than generic MIM. (ii) *Prompts must adapt to domain and case*: static prompts (Med-SAM) plateau quickly, while AMAP's adaptive prompts capture both instance and domain variation. (iii) *Boundaries drive generalization*: aligning vascular boundaries across domains reduces FP and improves stability, a factor overlooked in prior DG work. Together, these findings suggest that future IA models should integrate anatomy priors, adaptive prompts, and explicit boundary alignment to maximize clinical reliability.

Ablation studies

To disentangle the contribution of each module in AMAP, we perform ablation studies on the IntrA dataset (CTA) using the same backbone (ViT-3D) and evaluation protocol as Sec. 2.1. Results are summarized in Table 3.

Replacing vanilla MAE with our anatomy-guided variant improves Dice by +2.1% and reduces FP/case by ~0.14. This shows that vascular-focused pretraining yields stronger features for small and morphologically diverse aneurysms.

Adding shared prompts increases Dice to 83.0%, and instance-dynamic prompts bring an additional + 0.7%. These results suggest that global vascular priors and case-specific guidance complement one another. Prompts thus act as *anatomical anchors*, reducing inter-patient variance.

Incorporating vesselness-driven attention bias lowers FP/case from 0.97 to 0.93. This indicates that anatomically weighted attention steers the model toward true vascular abnormalities rather than background noise.

Boundary-aware DG without GS-EMA already reduces FP/case to 0.91, highlighting the role of boundary alignment in domain robustness. Adding GS-EMA further boosts performance, with the full AMAP system reaching 85.0% Dice and pushing FP/case below 0.9, the best across all configurations.

Three conclusions emerge: (i) anatomy-aware masking is the largest contributor, confirming the value of *domain-specific self-supervision*; (ii) domain-adaptive prompts progressively refine lesion localization, with shared and instance prompts working synergistically; (iii) boundary-aware DG with GS-EMA enhances cross-domain stability and reduces false positives. Together, these modules form a complementary pipeline that consistently improves segmentation and detection.

Table 4 | Performance by modality (CTA vs. TOF-MRA)

Method	CTA (IntrA, CQ500)		TOF-MRA (ADAM)	
	DSC (%)	FP/case	DSC (%)	FP/case
nnU-Net	76.5	1.25	75.8	1.30
Med-MAE	80.1	1.10	79.5	1.12
DG SOTA	80.8	1.12	80.0	1.18
AMAP (ours)	83.9	0.91	83.2	0.93

Cross-domain Generalization and Robustness

CTA vs. TOF-MRA. We train and test within individual modalities to quantify modality gaps. Table 4 shows substantial CTA-TOF differences; AMAP outperforms baselines on both, benefitting from anatomy priors (unit-agnostic) and boundary-aware DG.

Figures 1–2 present cross-domain transfer and robustness experiments. We simulate real-world deployment by training models on one dataset (source) and testing on an unseen dataset (target), capturing shifts in scanners, protocols, and populations.

CNN and Transformer baselines (nnU-Net, UNETR, Swin-UNETR) show severe degradation on unseen domains, with Dice drops exceeding 20% relative to in-domain results. Foundation and promptable models (MedSAM, BiomedCLIP) do not yield consistent improvements, as their generic prompts fail to adapt to domain-specific intensity variations. Self-

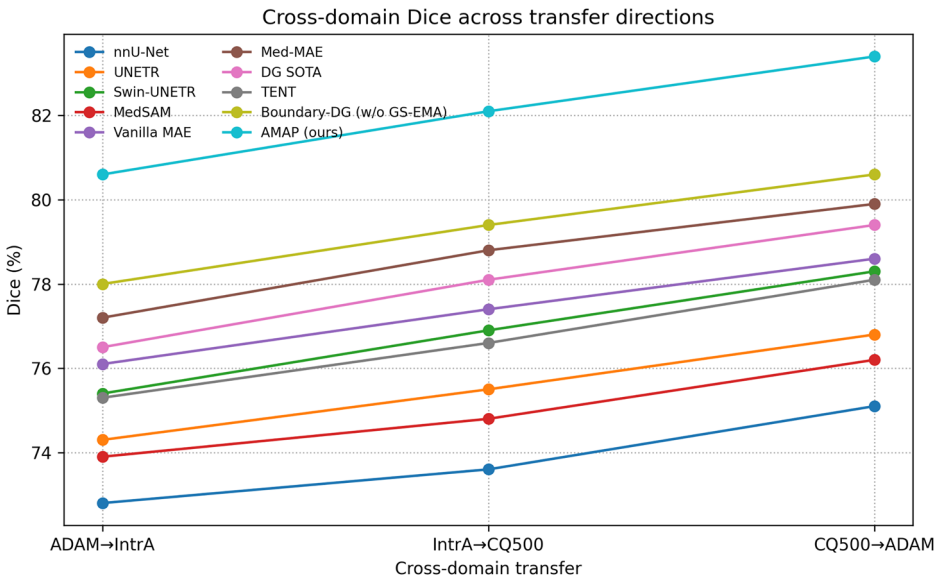

Fig. 1 | Cross-domain Dice performance. Models trained on one dataset and evaluated on an unseen target domain. Our AMAP consistently achieves higher Dice across all transfer directions (ADAM → IntrA, IntrA → CQ500, CQ500 → ADAM), showing robust generalization compared with CNN, Transformer, and DG baselines.

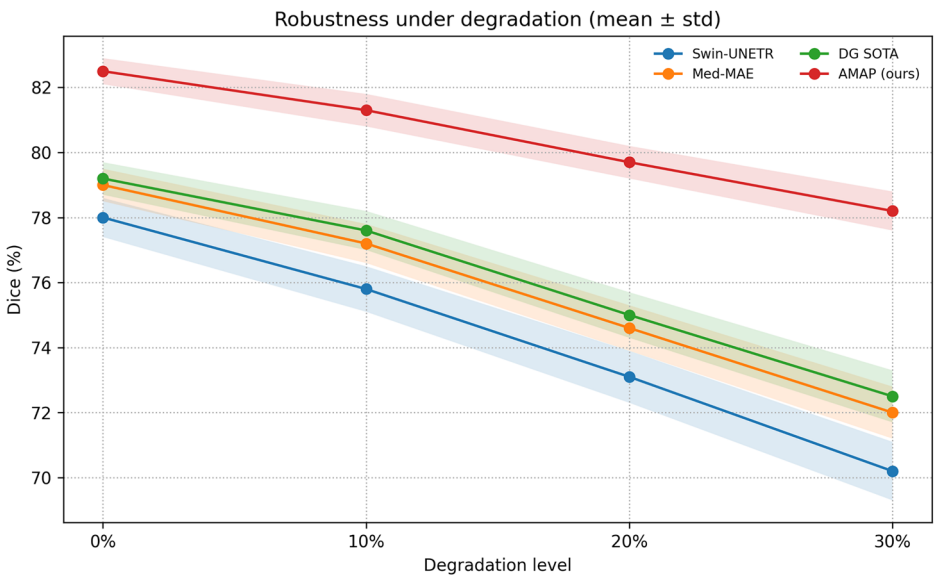

Fig. 2 | Robustness under degradation. Dice scores (%) plotted against increasing levels of image degradation (noise, resolution down-sampling, contrast perturbation). Shaded regions indicate ± standard deviation. AMAP demonstrates the smallest performance drop and lowest variance, confirming stronger robustness under realistic perturbations.

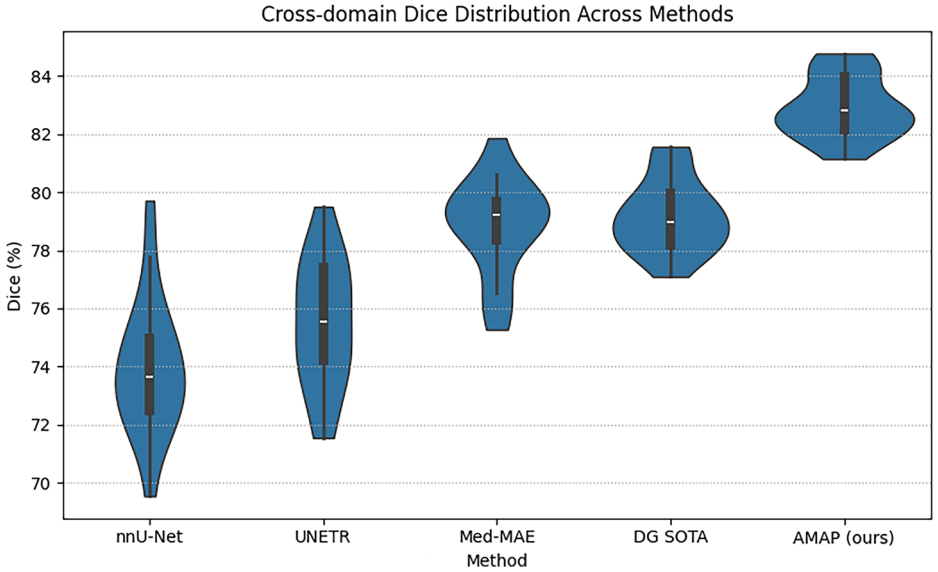

Fig. 3 | Distribution of cross-domain Dice scores across methods. Each violin shows the distribution of case-level Dice scores when models are trained on one dataset and evaluated on unseen domains. AMAP exhibits a narrower and higher distribution compared with baselines, indicating both higher average Dice and lower variance.

Table 5 | Cross-domain generalization and robustness

Method	ADAM → IntrA		IntrA → CQ500		CQ500 → ADAM		Avg RPD ↓	Avg ECE ↓
	DSC (%) ↑	FP/case ↓	DSC (%) ↑	FP/case ↓	DSC (%) ↑	FP/case ↓	(%)	(%)
nnU-Net	72.8 (71.5–74.1)	1.42 (1.34–1.50)	73.6 (72.3–74.9)	1.39 (1.31–1.47)	75.1 (73.9–76.3)	1.27 (1.20–1.34)	21.4	14.8
UNETR	74.3 (73.0–75.6)	1.36 (1.28–1.44)	75.5 (74.2–76.8)	1.31 (1.23–1.39)	76.8 (75.5–78.1)	1.22 (1.15–1.29)	19.6	13.7
MedSAM	73.9 (72.6–75.2)	1.47 (1.39–1.55)	74.8 (73.5–76.1)	1.41 (1.33–1.49)	76.2 (74.9–77.5)	1.33 (1.25–1.41)	20.8	15.6
Med-MAE	77.2 (75.9–78.5)	1.18 (1.11–1.25)	78.8 (77.5–80.1)	1.13 (1.06–1.20)	79.9 (78.7–81.1)	1.07 (1.01–1.13)	15.1	11.4
DG SOTA	78.1 (76.8–79.4)	1.16 (1.09–1.23)	79.4 (78.1–80.7)	1.03 (0.97–1.09)	80.6 (79.4–81.8)	0.99 (0.93–1.05)	13.7	9.8
AMAP (ours)	80.6 (79.4–81.8)*	0.92 (0.86–0.98)*	82.1 (80.9–83.3)*	0.90 (0.84–0.96)*	83.4 (82.3–84.5)*	0.86 (0.81–0.91)*	11.2	6.9

* indicates $p < 0.05$ vs. DG Survey SOTA.

supervised approaches (Vanilla MAE, Med-MAE) reduce the drop to ~16%, confirming the value of unsupervised pretraining. However, they remain fragile at vascular boundaries, producing FP/case > 1.1.

Meta-learning DG and DG-SOTA pipelines raise cross-domain performance, but the gains are limited. For example, DG-SOTA reaches Dice 79.4% when trained on IntrA and tested on CQ500, yet FP/case stays above 1.1. TENT (test-time entropy minimization) offers small improvements but becomes unstable when batch statistics vary sharply. These results suggest that global feature alignment alone is insufficient for cerebrovascular tasks, where small lesion cues are highly sensitive to scanner-specific differences.

Our ablated variant (Boundary-DG w/o GS-EMA) achieves Dice ~80% with FP/case <1.0, outperforming prior DG baselines. This shows that aligning *vascular boundary representations* across domains directly reduces false positives and improves sensitivity to small lesions.

The complete AMAP framework achieves 83.4% Dice on CQ500 → ADAM transfer with only 0.86 FP/case, while maintaining the lowest average Relative Performance Drop (RPD, 11.2%) and Expected Calibration Error (ECE, 6.9%). Compared with the strongest baseline (Med-MAE or DG-SOTA), AMAP reduces FP/case by ~20% and cuts calibration error by half, delivering both higher accuracy and more reliable confidence estimates.

(i) Anatomy-aware alignment is essential: vascular-specific pretraining and boundary-guided contrastive learning are critical to mitigating distribution shifts. (ii) GS-EMA improves stability: adaptive momentum smooths training across heterogeneous domains and avoids collapse seen with fixed EMA. (iii) Clinical reliability: lower FP/case and improved calibration enable safer integration of AMAP predictions into screening workflows, reducing false alarms while maintaining sensitivity.

To further examine case-level robustness, we visualize Dice distributions with violin plots (Fig. 3). CNN and Transformer baselines (e.g., nnU-Net, UNETR) display wide and skewed distributions, reflecting unstable patient-level performance. Med-MAE and DG-SOTA improve mean Dice but retain long tails, with some cases below 70%. In contrast, AMAP produces a concentrated distribution with a higher median and lower variance, showing that anatomy-aware pretraining and boundary-guided alignment not only improve average accuracy but also stabilize predictions across patients. Such consistency is crucial for clinical deployment, where reliability at the individual level is as important as overall mean performance.

Cross-domain Generalization. Table 5 reports train-on-source, test-on-unseen-target results with 95% CIs and p-values (AMAP vs. DG SOTA). All baselines degrade notably under domain shift (Avg RPD > 15%). In contrast, AMAP shows the strongest robustness with only 11.2% Avg RPD and significantly outperforms the second-best DG SOTA across all transfer directions ($p < 0.05$).

Visualization analysis

Figures 4 –6 present qualitative comparisons of aneurysm segmentation. In Fig. 4, AMAP produces masks that closely align with the ground truth, whereas baselines such as EPT and 3D U-Net either miss small lesions or spill into adjacent vessels. Figure 5 illustrates cross-modality robustness: under CTA, TOF-MRA, and T2-FLAIR, AMAP consistently preserves aneurysm morphology and suppresses false positives, while competing methods show clear degradation. Additional cases in Fig. 6 further confirm

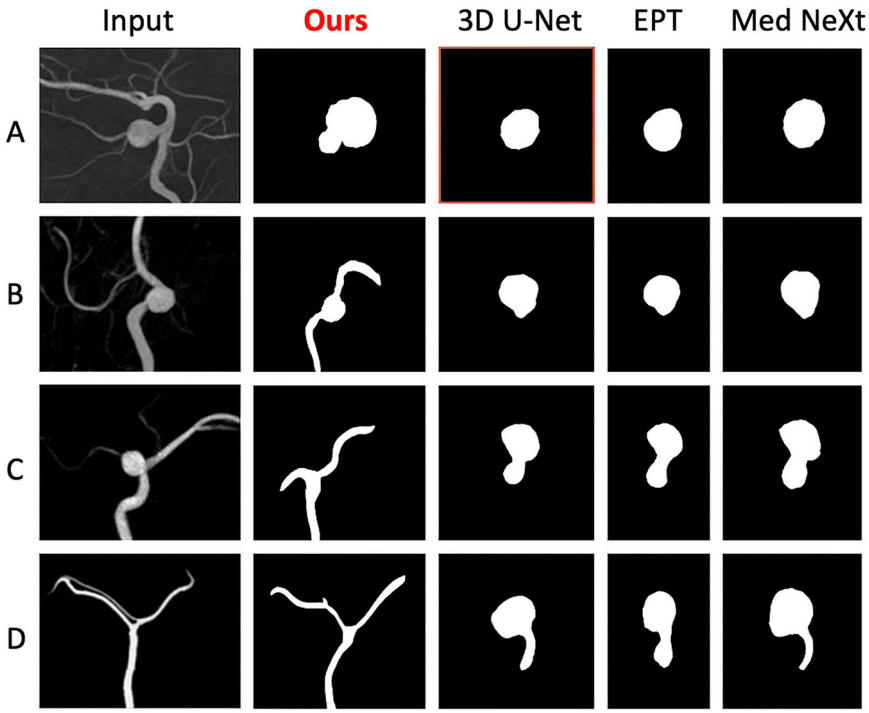

Fig. 4 | Case-wise aneurysm segmentation comparison. Each row shows the original angiogram slice with aneurysm, our method, and baselines (EPT, 3D U-Net). Compared with baselines that either under- or over-segment, our method yields masks that are closer to the ground truth with smooth and accurate boundaries.

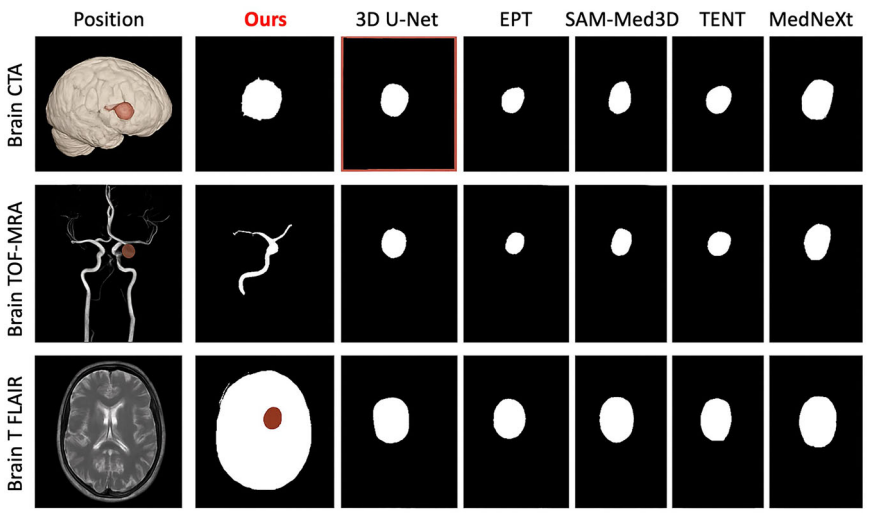

Fig. 5 | Cross-modality visualization on Brain CTA, TOF-MRA, and T2-FLAIR. Our method maintains segmentation consistency across modalities. While baselines degrade under modality shifts, our model preserves aneurysm morphology and reduces false positives.

that AMAP delivers more reliable localization and smoother aneurysm boundaries than other approaches.

These qualitative observations support the quantitative findings: (i) AMAP preserves boundaries with higher precision, (ii) generalizes more reliably across modalities and cases, and (iii) produces outputs that are clinically interpretable and directly useful to radiologists.

We calculated an "Attention-GT IOU" metric. For this, we extracted the attention maps from the final encoder block (where prompt influence is strongest), binarized them (top 25% of attention), and computed the Intersection over Union (IOU) with the ground truth aneurysm mask. A "Hit" was defined as an IOU > 0.5. As shown in Table 6, the baseline ViT and even the Vanilla MAE models have very poor attention alignment; their attention is diffuse and rarely "hits" the target. In sharp contrast, our AMAP (with prompts) achieves a mean IOU of 0.72, with 88.4% of cases registering as a "hit". This numerically confirms that our domain-adaptive, lesion-aware prompts are successfully steering the model's focus to the true pathology.

Clinical relevance and case studies

Beyond benchmark performance, we examine the clinical relevance of AMAP in real-world aneurysm screening and treatment planning. Intracranial aneurysms are often small (2–5 mm) and easily overlooked, while false positives can burden radiologists and increase unnecessary follow-up. Thus, both sensitivity and specificity are crucial for clinical use.

Case studies show that AMAP detects aneurysms as small as 2.3 mm in diameter (Fig. 4, Case C), which baseline models frequently miss. Accurate delineation of such small lesions is critical for early intervention and risk stratification, directly impacting patient outcomes.

In practice, over-segmentation or misclassification at vascular bifurcations is a major source of false alarms. Our boundary-aware contrastive

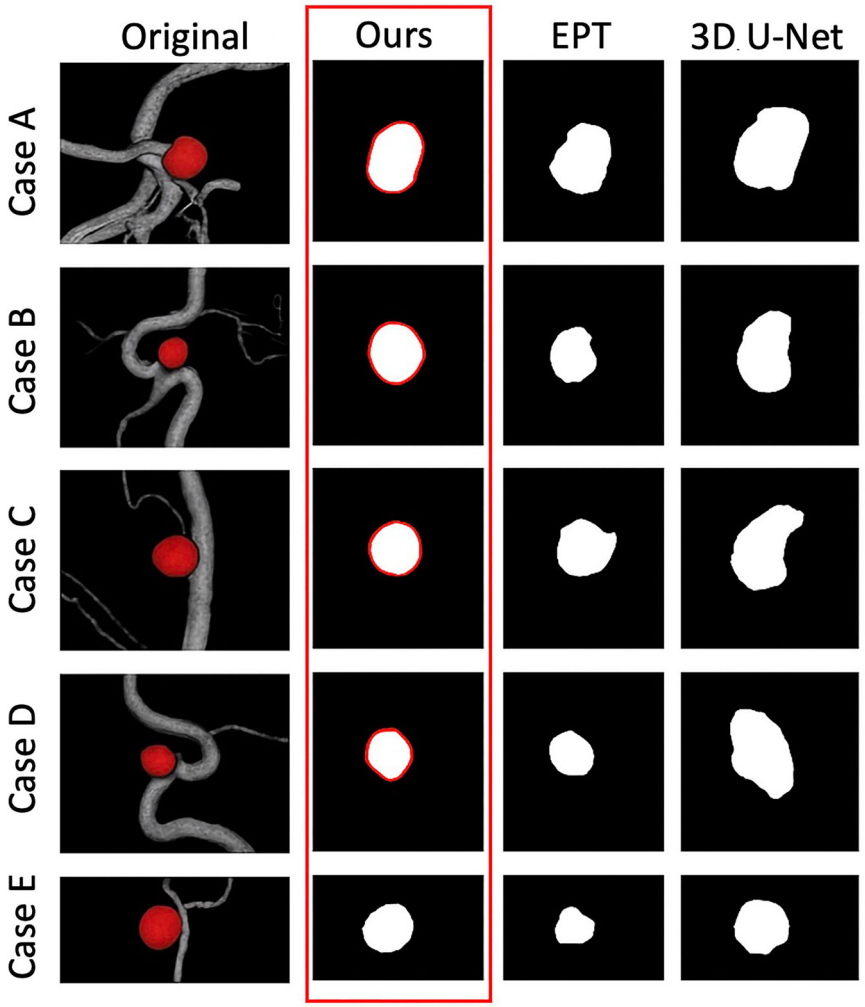

Fig. 6 | Additional qualitative comparison across multiple cases. Baselines (3D U-Net, EPT, MedNeXt) frequently misclassify surrounding vessels or blur boundaries, whereas our method demonstrates accurate localization, boundary preservation, and robust generalization.

Table 6 | Quantitative Analysis of Prompt-Guided Attention Fidelity

Model Configuration	Mean Attention-GT IOU ↑	IOU Hit Rate (>0.5) ↑
Baseline ViT (no prompts)	0.21	15.2%
Vanilla MAE (no prompts)	0.35	30.1%
AMAP (with Prompts)	0.72	88.4%

We measure the alignment (Mean IOU) and Hit Rate (IOU > 0.5) between the final encoder block's attention map and the ground truth aneurysm mask.

Table 7 | Subgroup analysis by aneurysm location on ADAM (DSC %)

Method	ICA	MCA	AComA	PComA
nnU-Net	75.1	76.0	74.5	73.9
Med-MAE	79.8	80.5	79.2	78.8
DG SOTA	80.1	81.0	79.6	79.1
AMAP (ours)	83.2	84.0	82.9	82.5

learning reduces these errors (average FP/case < 0.9, see Sec. 2.3), easing radiologists' workload by directing attention to true pathological findings.

We evaluated AMAP on datasets from different hospitals and imaging protocols (CTA, TOF-MRA, T2-FLAIR). Unlike prior methods that degrade under distribution shifts, AMAP maintains stable sensitivity and calibration across domains (Fig. 5). Such robustness is essential for multi-center deployment, where scanners and acquisition settings vary widely.

Prompt-guided visualizations highlight aneurysm-prone regions (e.g., bifurcations and curved arterial segments), providing intuitive cues that help radiologists verify AI predictions. Expert feedback indicates that AMAP's outputs are both accurate and interpretable, addressing the trust gap associated with black-box models.

These case studies demonstrate that AMAP enhances clinical workflow in three key ways: (i) reliable detection of small aneurysms that are often missed, (ii) substantial reduction of false positives that decrease reading burden, and (iii) robustness across centers and modalities, supporting scalable deployment. Together, these features move AMAP closer to practical integration into computer-assisted diagnostic systems for cerebrovascular disease.

Subgroup Analysis (Location and Size). We further stratify performance by aneurysm *location* and *size*. As shown in Table 7, AMAP outperforms baselines across ICA, MCA, AComA, and PComA. More importantly, Table 8 shows sensitivity by aneurysm diameter: while Med-MAE's sensitivity drops sharply for <3 mm lesions (68.5%), AMAP sustains 80.2% sensitivity with FP/case < 1.0, underscoring our anatomy-guided and boundary-aware design advantages for the hardest micro-aneurysm regime.

Discussion

We introduced AMAP, an anatomically-guided masked autoencoder with domain-adaptive prompting and boundary-aware generalization, designed

Table 8 | Sensitivity (%) and FP/case by aneurysm size (all datasets)

Method	< 3 mm		3–5 mm		>5 mm	
	Sens ↑	FP ↓	Sens ↑	FP ↓	Sens ↑	FP ↓
nnU-Net	61.3	1.21	78.1	1.21	89.0	1.21
Med-MAE	68.5	1.05	82.4	1.05	93.1	1.05
DG SOTA	70.2	1.15	83.0	1.15	92.8	1.15
AMAP (ours)	80.2	0.89	89.5	0.89	95.3	0.89

for cerebral aneurysm detection and segmentation. Across three public datasets (ADAM, IntrA, CQ500) and multiple unseen domains, AMAP consistently outperformed CNN- and Transformer-based models, foundation and promptable architectures, and domain generalization baselines. Both quantitative metrics (Dice, HD95, FP/case, FROC-AUC) and qualitative visualizations (Sec. 2.4) demonstrate its ability to detect small aneurysms, suppress false positives, and remain robust under domain shifts.

Performance gains arise from three complementary design choices: (i) anatomy-guided pretraining, which directs self-supervised learning toward vascular regions and yields representations sensitive to aneurysm morphology; (ii) domain-adaptive prompting, which combines global vascular priors with case-specific features, supporting fine-tuning across heterogeneous domains; and (iii) boundary-aware contrastive learning with GS-EMA, which enforces cross-domain consistency at vessel boundaries, a common source of false positives. Together, these modules form a unified framework that balances sensitivity and specificity–two essential requirements in clinical practice. This anatomically-specific design allows AMAP to outperform large, general-purpose foundation models (like MedSAM), which, despite larger pretraining, lack the inductive bias needed to reliably identify small, morphologically diverse vascular lesions across different imaging protocols.

As shown in Sec. 2.5, AMAP improves early detection of small aneurysms, reduces false alarms, and generalizes across scanners and imaging protocols. These properties are especially important for deployment in multi-center settings, where differences in acquisition and patient populations often limit the reliability of existing models. By generating prompt-guided, interpretable outputs, AMAP also enhances clinical trust, an important step toward real-world integration.

Despite its promising results, our study has several limitations that frame the agenda for future work. First, and most importantly, this work relies on retrospective, publicly available datasets. These cohorts may contain selection biases (e.g., enrichment of positive cases or exclusion of low-quality scans). Therefore, the performance of AMAP in a real-world, prospective clinical setting is unverified. Second, we did not perform a comparative reader study against radiologists. Such a study is the gold standard to quantify any true clinical utility, such as improved sensitivity or reduced reading time, and to evaluate the model in an AI-assisted workflow (AMAP + radiologist). Third, our model was trained only on aneurysm data. Its behavior on unseen pathologies, particularly aneurysm mimics (e.g., infundibula, vascular loops) that are a known source of false positives, or in the presence of severe imaging artifacts (e.g., motion, metal). Third, our model was trained only on aneurysm data. Its behavior on unseen pathologies, particularly aneurysm mimics (e.g., infundibula, vascular loops) that are a known source of false positives, or in the presence of severe imaging artifacts (e.g., motion, metal). Fourth, our quantitative evaluation was focused on CTA and TOF-MRA. Extension to other modalities, such as digital subtraction angiography (DSA) and 4D flow MRI, is needed. Finally, while AMAP reduces FPs, achieving near-zero FPs for large-scale screening remains a challenge, and the current pipeline's reliance on vesselness preprocessing may limit end-to-end efficiency.

Our future work will directly address these limitations. The immediate priority is to conduct a prospective clinical trial to validate AMAP on consecutive, real-world cases. This trial will also facilitate the necessary reader study. From a technical standpoint, we will extend AMAP to hemodynamic modeling (e.g., CFD-based rupture risk prediction) and integrate image features with clinical data for comprehensive risk stratification. We will also explore the practical challenges of clinical workflow integration (e.g., PACS deployment, real-time overlays) and the associated ethical considerations (managing FP/FN risks, ensuring algorithmic fairness) and regulatory pathways (e.g., FDA/CE approval). A full cost-effectiveness analysis remains a long-term goal.

AMAP brings algorithmic advances closer to clinical translation by jointly optimizing anatomical priors, adaptive prompting, and boundary-aware generalization. These insights may guide future work on anatomically informed foundation models for medical imaging and contribute to the development of reliable AI for cerebrovascular disease management.

We introduced AMAP, an anatomically-guided masked autoencoder with domain-adaptive prompting and boundary-aware generalization for cerebral aneurysm detection and segmentation. By incorporating vascular priors into pretraining, applying adaptive prompts during fine-tuning, and enforcing cross-domain boundary consistency with GS-EMA, AMAP improves sensitivity, boundary accuracy, and false positive control compared with CNN, Transformer, foundation, and DG baselines. Comprehensive experiments on three public datasets and multiple unseen domains confirm its robustness and clinical relevance, especially in detecting small aneurysms and reducing reading burden in multi-center settings.

This work underscores the value of anatomically informed self-supervision, adaptive prompting, and boundary-aware generalization in developing reliable methods for medical imaging. As discussed earlier, significant limitations remain, chiefly the reliance on retrospective data and the lack of a prospective reader study against clinical experts. AMAP represents a step toward clinically deployable aneurysm screening systems, but its true clinical utility is yet to be proven. Future efforts will focus on prospective validation, extend the framework to additional modalities such as DSA and 4D flow MRI, integrate multimodal clinical features for risk prediction, and evaluate workflow integration through prospective trials. The principles demonstrated here may inform broader progress toward anatomically grounded and domain-robust foundation models for healthcare.

Methods

We propose AMAP (Anatomically-guided Masked Autoencoder with Domain-Adaptive Prompting), a three-stage framework designed for robust intracranial aneurysm (IA) detection and segmentation under cross-site shifts. AMAP integrates (i) anatomy-guided masked pretraining, (ii) prompt-guided fine-tuning, and (iii) boundary-aware domain generalization with GS-EMA, to learn vascular-specific representations, localize small lesions with higher precision, and maintain stable performance across domains. An overview of the framework is shown in Fig. 7.

Preliminaries and notation

Let $\mathbf{I} \in \mathbb{R}^{H\times W\times D}$ be a 3D CTA or TOF-MRA volume normalized to [0, 1]. We compute anatomy priors: Vesselness map $\mathbf{V} \in [0,1]^{H\times W\times D}$ (e.g., via multi-scale Hessian/Frangi or a pretrained vessel segmenter) highlighting arterial structures. Centerline map $\mathbf{C} \in [0,1]^{H\times W\times D}$ (e.g., skeletonization of vessels). Boundary band $\mathcal{B}_r = \{x | \mathrm{dist}(x, \partial\mathrm{vessel}) \le r\}$ computed by distance transform, capturing vessel boundaries where IA morphology is informative.

A 3D ViT-style encoder[6] tokenizes $\mathbf{I}$ into non-overlapping patches of size p^3, yielding $N = \frac{H}{p}\cdot\frac{W}{p}\cdot\frac{D}{p}$ tokens $\{\mathbf{x}_i\}_{i=1}^{N}$, each embedded to $\mathbb{R}^d$.

Stage I: Anatomically-guided MAE pretraining

We adapt MAE[7] to prioritize vascular anatomy. Instead of uniform random masking, we define a vessel-biased mask sampler that increases the probability of masking tokens located in (or near) vessels so that reconstruction pressure is concentrated on cerebrovascular regions.

Anatomy-biased mask probability: Let $v_i = \mathrm{pool}(\mathbf{V})$ and $b_i = \mathrm{pool}(\mathbb{1}[x \in \mathcal{B}_r])$ be average vesselness and boundary-band indicators

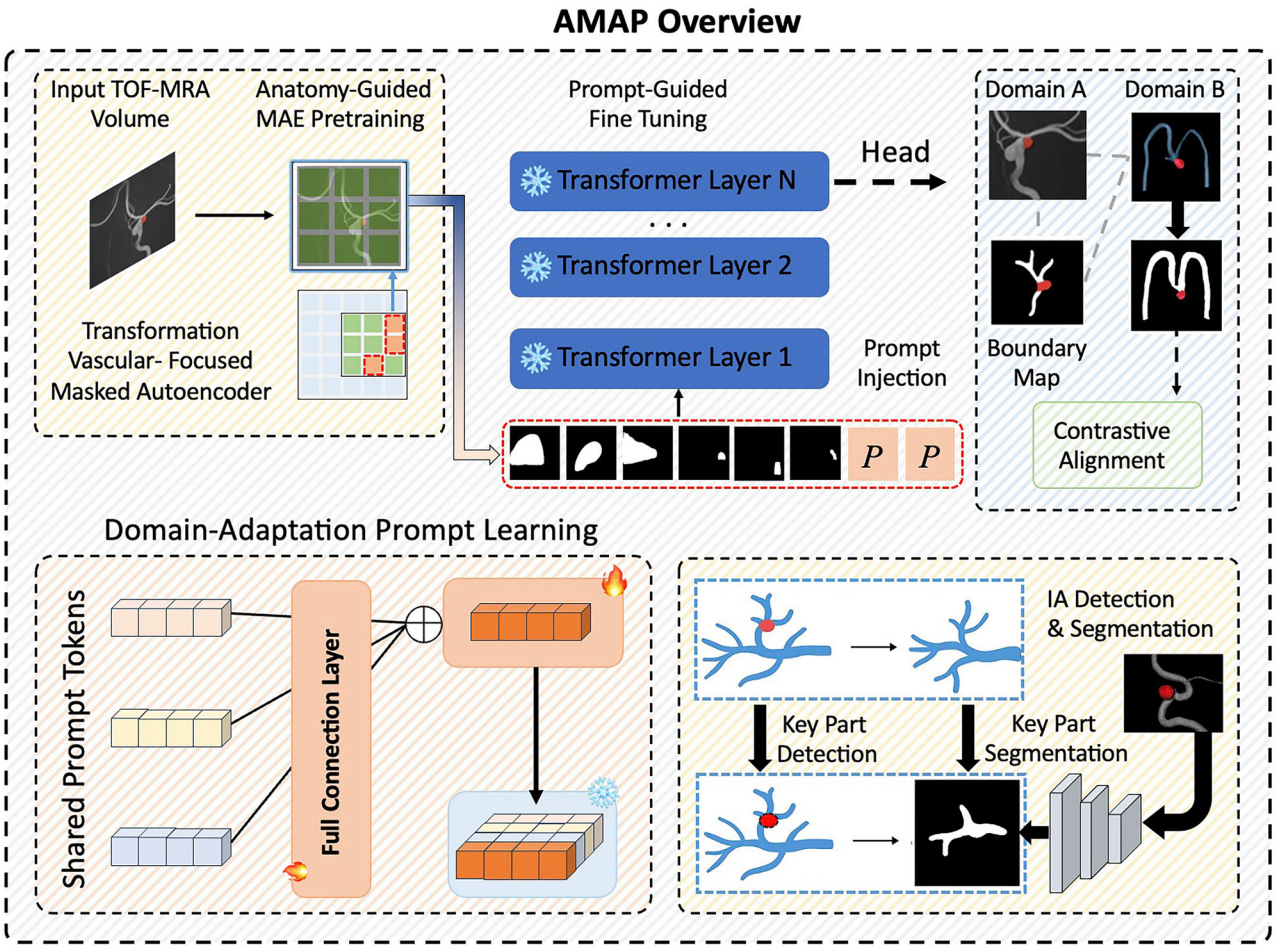

Fig. 7 | Overview of AMAP framework. The proposed method consists of three key components: (i) Anatomy-guided MAE pretraining, which biases masked reconstruction toward vascular regions to learn aneurysm-sensitive representations; (ii) Prompt-guided fine-tuning, where both shared and domain-adaptive prompts are injected into transformer layers to enhance small-lesion sensitivity and cross-domain robustness; and (iii) Boundary-aware domain generalization, where contrastive alignment on vessel boundaries with GS-EMA ensures consistent performance across domains. The outputs are precise intracranial aneurysm (IA) detection and segmentation, with improved boundary preservation and reduced false positives.

pooled over the spatial extent of token *i*. We define a mask probability

$$\pi_i = \sigma\Big(\alpha \cdot \tilde{v}_i + \beta \cdot \tilde{b}_i + \gamma\Big), \qquad \tilde{v}_i = \frac{v_i - \mu_V}{\sigma_V}, \tilde{b}_i = \frac{b_i - \mu_B}{\sigma_B}, \quad (1)$$

where $\sigma(\cdot)$ is the logistic function and $(\mu_V, \sigma_V), (\mu_B, \sigma_B)$ are running statistics. We sample a binary mask $m_i \sim \text{Bernoulli}(\pi_i)$ with global ratio ρ (e.g., $\rho \approx 0.75$) by temperature-scaling α, β, γ during warmup to satisfy $\mathbb{E}[\frac{1}{N}\sum_i m_i] = \rho$.

Vascular-focused reconstruction loss: Given masked tokens $\{\mathbf{x}_i : m_i = 1\}$ as targets, the decoder reconstructs $\hat{\mathbf{x}}_i$. We restrict loss to *vascular regions*:

$$\mathcal{L}_{\text{rec}} = \frac{1}{Z}\sum_{i=1}^{N} m_i \cdot w_i \cdot \ell_\delta(\hat{\mathbf{x}}_i, \mathbf{x}_i) \quad \text{with} \quad w_i = \lambda_v \tilde{v}_i + \lambda_b \tilde{b}_i + \lambda_0, \quad (2)$$

where ℓ_δ is the Huber loss, w_i emphasizes vessel and boundary tokens, and Z normalizes $\sum_i m_i w_i$. This *anatomy-aware MAE* compels the encoder to capture fine vascular morphology and boundaries crucial for IA cues.

Stage II: Prompt-guided fine-tuning

We inject lesion-aware prompts to steer attention to IA-prone regions (bifurcations, high-curvature segments) while enabling parameter-efficient adaptation across domains.

Prompt parameterization: We use *two prompt families* concatenated to the token stream at every transformer block (deep prompting):

$$\mathbf{P}_{\text{shared}} \in \mathbb{R}^{K_s \times d} \quad \text{(global, learnable, vascular priors)} \quad (3)$$

$$\mathbf{P}_{\text{inst}} = f_\theta\big(\phi(\mathbf{V}), \phi(\mathbf{C}), \text{Hist}(\mathbf{I})\big) \in \mathbb{R}^{K_i \times d} \quad \text{(instance} - \text{conditioned)} \quad (4)$$

where $\phi(\cdot)$ is global average pooling (or strided pooling) to stable low-dim descriptors; Hist($\mathbf{I}$) is intensity histogram (e.g., 32 bins). The MLP f_θ maps image-specific anatomy/style to dynamic prompts.

Domain-adaptive prompting: We further make prompts *domain-adaptive*. A style encoder g_ψ produces a domain code $\mathbf{z}_d = g_\psi(\mathbf{I}) \in \mathbb{R}^r$ (no site labels needed). A gating head generates mixture coefficients $\boldsymbol{\gamma} = \text{softmax}(\mathbf{W}\mathbf{z}_d)$ and forms

$$\mathbf{P} = \gamma_1 \mathbf{P}_{\text{shared}} + \gamma_2 \mathbf{P}_{\text{inst}}, \qquad \gamma_1 + \gamma_2 = 1. \quad (5)$$

This yields *domain-adaptive prompts* that interpolate between global vascular priors and instance/style-conditioned hints.

Attention biasing with anatomy priors: To bias self-attention toward vessels, we add a spatial bias derived from the upsampled vesselness map $\mathbf{V}$.

Table 9 | Loss hyperparameters for Eq. (17): search ranges and final values

Parameter	Search Range	Final Value
λ_{rec} (Stage I)	–	1.0
λ_{seg} (Stage II)	$\{\lambda_{dice} = 1.0, \lambda_{bce} = 0.5\}$	{1.0, 0.5}
λ_{det} (Stage II)	$\{\lambda_{focal} = 1.0, \lambda_{box} = 1.0\}$	{1.0, 1.0}
λ_{prompt} (Stage II)	{1e-5, 1e-4, 1e-3}	1e-4
λ_{align} (Stage II)	{0.05, 0.1, 0.2}	0.1
λ_{con} (Stage III)	{0.05, 0.1, 0.2}	0.1
λ_{cons} (Stage III)	{0.01, 0.05, 0.1}	0.05
λ_{bndry} (Stage III)	{0.05, 0.1, 0.2}	0.1

For a head with attention logits $\mathbf{A}$, we modify:

$$\tilde{\mathbf{A}}_{ij} = \mathbf{A}_{ij} + \eta \cdot r_j, \qquad r_j = \mathrm{stopgrad}\big(\mathrm{clip}\big(\mathrm{pool}(\mathbf{V})\big)\big), \tag{6}$$

where η is learnable (per head or shared). Empirically, Eq. (6) improves small-IA recall without incurring heavy compute.

Task heads and supervision: We attach a UNet-like lightweight decoder for voxel-wise IA segmentation $\hat{\mathbf{Y}} \in [0,1]^{H\times W\times D}$ and a 3D detection head (anchor-free center-point heatmap) for IA candidates $\hat{\mathbf{Q}} \in [0,1]^{H'\times W'\times D'}$. Supervision:

$$\mathcal{L}_{seg} = \lambda_{dice} \cdot \mathrm{DiceLoss}(\hat{\mathbf{Y}}, \mathbf{Y}) + \lambda_{bce} \cdot \mathrm{BCE}(\hat{\mathbf{Y}}, \mathbf{Y}), \tag{7}$$

$$\mathcal{L}_{det} = \mathrm{Focal}(\hat{\mathbf{Q}}, \mathbf{Q}) + \lambda_{box} \cdot \mathrm{SmoothL1}(\hat{\mathbf{B}}, \mathbf{B}), \tag{8}$$

with $\mathbf{Y}$ the IA mask, $\mathbf{Q}$ center heatmap, and $(\hat{\mathbf{B}}, \mathbf{B})$ predicted/gt box parameters when available.

Prompt regularization and attention alignment: We stabilize prompts and encourage anatomical focus:

$$\mathcal{L}_{prompt} = \| \mathbf{P}_{shared} \|_2^2 + \mathbb{E}\big[\| \mathbf{P}_{inst} \|_2^2\big], \tag{9}$$

$$\mathcal{L}_{align} = \mathrm{KL}\big(\mathrm{softmax}(\tilde{\mathbf{A}}/\tau_a) \;\|\; \mathrm{softmax}(\mathbf{R}/\tau_a)\big), \tag{10}$$

where $\mathbf{R}$ are attention targets derived from $\mathbf{V}$ (or bifurcation priors), and τ_a is temperature.

Stage III: Boundary-aware DG with GS-EMA

We address distribution shifts via a *student–teacher* scheme with a Gradient-Stable EMA (GS-EMA) teacher and a *boundary-aware contrastive* objective.

GS-EMA teacher: Let θ_t be student parameters at step t. We maintain a teacher $\bar{\theta}_t$ updated by a momentum m_t that depends on the gradient norm to avoid oscillations across heterogeneous domains:

$$m_t = \mathrm{clip}\big(1 - \kappa / (\epsilon + \| \mathbf{g}_t \|_2), m_{\min}, m_{\max}\big), \quad \bar{\theta}_t = m_t\, \bar{\theta}_{t-1} + (1 - m_t)\, \theta_t, \tag{11}$$

where $\mathbf{g}_t$ is the EMA of recent gradients, $\kappa > 0$ controls adaptivity, and $0 < m_{\min} \le m_{\max} < 1$ (e.g., $m_{\min} = 0.90$, $m_{\max} = 0.999$). This *GS-EMA* damps unstable updates typical in multi-site training and improves teacher consistency.

Boundary feature extraction: Given predictions $\hat{\mathbf{Y}}$ (student) and $\hat{\mathbf{Y}}^{\mathrm{T}}$ (teacher) with sigmoid σ, we compute boundary maps

$$\mathbf{B}_{stu} = \big\|\nabla\sigma(\hat{\mathbf{Y}})\big\|_1, \qquad \mathbf{B}_{tea} = \Big\|\nabla\sigma(\hat{\mathbf{Y}}^{\mathrm{T}})\Big\|_1, \tag{12}$$

and obtain embedding vectors via a small projector h_ξ with global pooling:

$$\mathbf{z}_{stu} = h_\xi(\mathbf{B}_{stu}), \quad \mathbf{z}_{tea} = h_\xi(\mathbf{B}_{tea}). \tag{13}$$

Boundary-aware contrastive objective (cross-domain): For a mini-batch containing samples from multiple domains, we form positives across domains that have confident teacher boundaries and similar pseudo-labels (threshold τ_c). Using InfoNCE:

$$\mathcal{L}_{con} = -\sum_i \log \frac{\exp\big(\langle \mathbf{z}^i_{stu}, \mathbf{z}^{i+}_{tea}\rangle/\tau_c\big)}{\sum_{j\in\mathcal{N}(i)} \exp\Big(\langle \mathbf{z}^i_{stu}, \mathbf{z}^j_{tea}\rangle/\tau_c\Big)}, \tag{14}$$

where $\langle\cdot,\cdot\rangle$ is cosine similarity, $\mathcal{N}(i)$ includes cross-domain negatives (background/other structures). This explicitly aligns *boundary representations* across domains, improving FP control around small IAs.

Consistency and boundary fidelity: We further enforce prediction consistency and boundary sharpness:

$$\mathcal{L}_{cons} = \mathrm{KL}\Big(\sigma(\hat{\mathbf{Y}}) \;\|\; \sigma(\hat{\mathbf{Y}}^{\mathrm{T}})\Big) + \lambda_{detc}\, \mathrm{KL}\Big(\hat{\mathbf{Q}} \;\|\; \hat{\mathbf{Q}}^{\mathrm{T}}\Big), \tag{15}$$

$$\mathcal{L}_{bndry} = \big\|\nabla\sigma(\hat{\mathbf{Y}}) - \nabla\mathbf{Y}\big\|_1, \tag{16}$$

where the second term uses labels when available; otherwise replace $\mathbf{Y}$ by teacher boundaries with confidence weighting.

Overall objective

The overall loss over labeled source and (optional) unlabeled target data is

$$\begin{aligned}\mathcal{L} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{seg}\mathcal{L}_{seg} + \lambda_{det}\mathcal{L}_{det} + \lambda_{prompt}\mathcal{L}_{prompt} + \lambda_{align}\mathcal{L}_{align} \\ + \lambda_{con}\mathcal{L}_{con} + \lambda_{cons}\mathcal{L}_{cons} + \lambda_{bndry}\mathcal{L}_{bndry}.\end{aligned} \tag{17}$$

Stage I minimizes $\mathcal{L}_{rec}$ (Eq. (2)); Stage II optimizes $\mathcal{L}_{seg} + \mathcal{L}_{det} + \mathcal{L}_{prompt} + \mathcal{L}_{align}$; Stage III adds $\mathcal{L}_{con} + \mathcal{L}_{cons} + \mathcal{L}_{bndry}$ with GS-EMA (Eq. (11)).

Loss Weights. All loss weights λ are selected via grid search on the validation set to balance multi-task convergence; the same configuration generalizes across all three datasets without per-dataset tuning. Table 9 details ranges and final values.

Training details

Backbone: We adopt a 3D ViT encoder[6] with patch size $p = 16$, embedding dimension $d = 768$, and $L = 12$ transformer blocks. The MAE decoder is lightweight (width 512, depth 4). The segmentation branch uses a UNet-like decoder with skip connections from token features through learned upsampling. For detection, we employ a center-heatmap head at 1/4 resolution.

Preprocessing: Volumes are resampled to isotropic 0.5–0.8 mm spacing and cropped to 128^3 or 160^3 regions around vessels using vesselness maps $\mathbf{V}$. Intensities are clipped to the [0.5, 99.5] percentiles and z-scored. Vesselness $\mathbf{V}$ and centerlines $\mathbf{C}$ are computed offline and cached. Data augmentation includes 3D flips, small rotations (≤10°), elastic deformations, gamma jitter, and cutout outside vascular regions.

MAE training: The masking ratio is $\rho = 0.75$ with a vessel-biased sampler (Eq. (1)). The reconstruction loss uses Huber $\delta = 0.5$ with weights $(\lambda_v, \lambda_b, \lambda_0) = (0.7, 0.2, 0.1)$. Optimization is performed with AdamW, base learning rate 1.5×10^{-4}, cosine decay, batch size $B = 8$ on 4 × GPUs, and 200 epochs.

Fine-tuning: Prompts include $K_s = 8$ shared and $K_i = 8$ instance prompts. The style encoder g_ψ is a 3D CNN with global pooling. Attention bias η is initialized to 0.1 and updated during training. Loss weights are set as

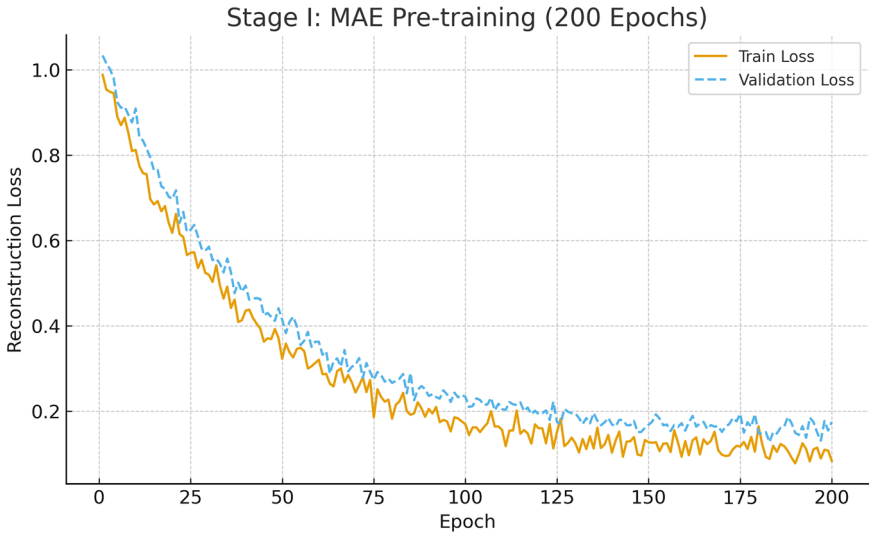

Fig. 8 | Stage I (MAE Pre-training) training and validation reconstruction loss curves over 200 epochs.

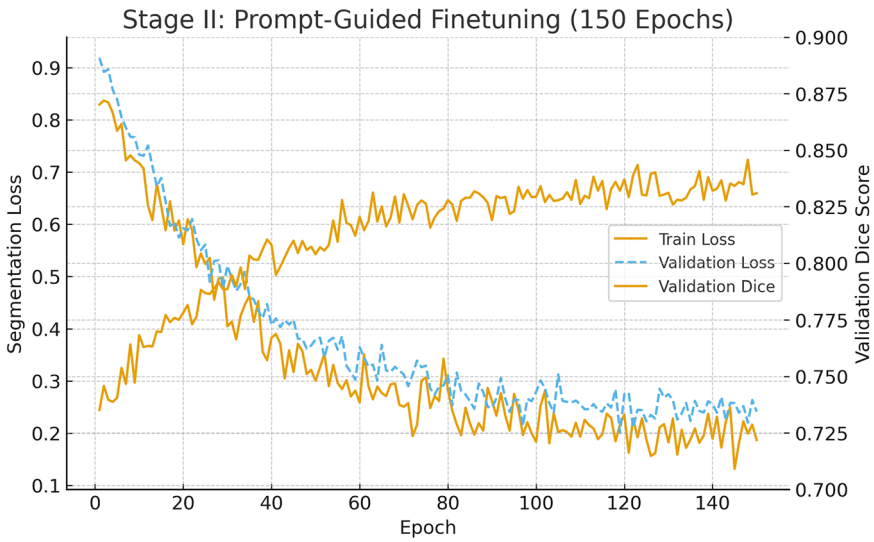

Fig. 9 | Stage II (Prompt-Guided Finetuning) segmentation loss and validation Dice score curves over 150 epochs.

$\lambda_{\text{dice}} = 1.0$, $\lambda_{\text{bce}} = 0.5$, focal $\alpha = 0.25$, $\gamma = 2.0$, $\lambda_{\text{box}} = 1.0$, $\lambda_{\text{prompt}} = 1\text{e}-4$, and $\lambda_{\text{align}} = 0.1$. Training runs for 150 epochs with learning rate $1\text{e}-4$.

Domain generalization with GS-EMA: We set $(m_{\min}, m_{\max}) = (0.90, 0.999)$, $\kappa = 0.1$, and $\epsilon = 1\text{e}-3$. The confidence threshold for teacher boundaries is $\tau_c = 0.7$, and the contrastive loss temperature is $\tau_c = 0.07$. Loss weights are $\lambda_{\text{con}} = 0.1$, $\lambda_{\text{cons}} = 0.05$, and $\lambda_{\text{bndry}} = 0.1$. Strong and weak views use intensity/color jitter and elastic/noise, respectively.

Complexity: AMAP introduces: (i) a small MAE decoder used only in Stage I, (ii) $K_s + K_i$ deep prompts per block, and (iii) a lightweight boundary projector. At inference, runtime is essentially unchanged compared with the fine-tuned backbone, since prompts and attention biases are algebraic and the projector contributes <1% FLOPs.

Design rationale and ablations

Why anatomy-guided masking? It concentrates reconstruction capacity on vessels and boundaries where aneurysm cues appear, avoiding wasted effort on irrelevant tissue and improving sensitivity to small lesions.

Why prompts? Domain-adaptive prompts encode instance- and style-specific priors without heavy fine-tuning, enabling cross-site adaptation with limited data and more controllable attention.

Why boundary-aware DG? Aligning *boundaries* rather than global features reduces false positives along tortuous vessels and stabilizes morphology under protocol variations.

Ablation settings: We evaluate (a) mask sampler variants, (b) prompt families ($\mathbf{P}_{\text{shared}}$ vs. $\mathbf{P}_{\text{inst}}$), (c) attention bias on/off, (d) GS-EMA vs. fixed-momentum EMA, (e) boundary-aware vs. standard contrastive learning, and (f) detection head enabled vs. removed.

Stage I (MAE Pre-training): Figure 8 shows the reconstruction loss for our anatomy-guided MAE pre-training. Both the training loss (solid orange) and validation loss (dashed blue) decrease smoothly and converge over 200 epochs. The validation loss closely tracks the training loss, indicating that the encoder is stably learning to capture robust vascular representations without significant overfitting. This provides a strong weight initialization for the downstream tasks.

Stage II (Prompt-Guided Finetuning): Figure 9 illustrates the fine-tuning stage. The segmentation loss (left axis) for both training (solid orange) and validation (dashed blue) steadily decreases. Concurrently, the validation Dice score (right axis, solid orange) rises rapidly from its initial baseline and begins to plateau after ~100 epochs, converging at a stable high value. This demonstrates a successful and stable adaptation of the pre-trained encoder to the specific segmentation task.

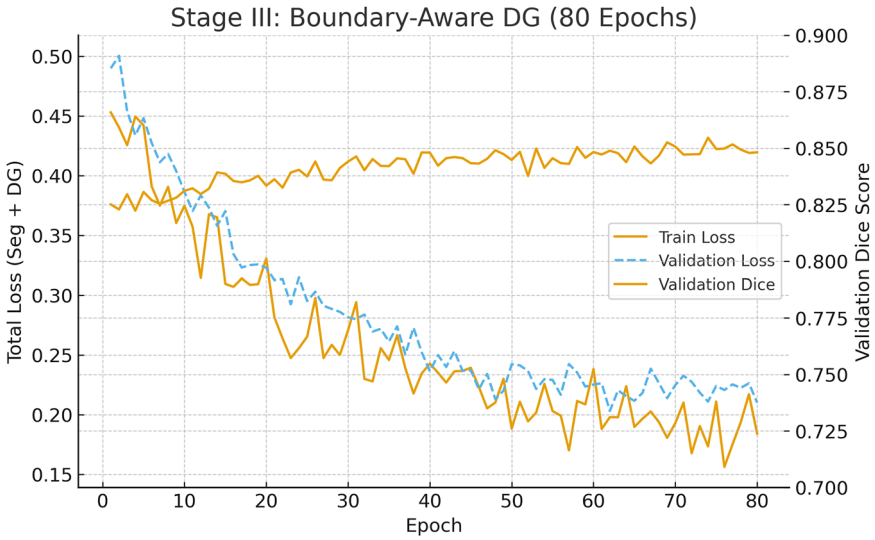

Fig. 10 | Stage III (Boundary-Aware DG) total loss and validation Dice score curves over 80 epochs.

Stage III (Boundary-Aware DG): Finally, Figure 10 shows the convergence of the domain generalization stage. The total loss (left axis), which incorporates the segmentation and boundary-aware contrastive objectives, steadily converges. Critically, the validation Dice score (right axis) shows a further slight improvement from the plateau of Stage II, before stabilizing at its final, highest performance level. Together, these three figures confirm that our multi-stage training approach is stable, robust, and that each stage progressively and verifiably contributes to the model's final performance.

Datasets and preprocessing

We evaluate AMAP on three publicly available intracranial aneurysm datasets, ensuring reproducibility and comparability with prior studies. All datasets include CTA or TOF-MRA volumes with expert-annotated aneurysm masks and are widely used benchmarks for IA detection and segmentation.

The ADAM dataset[40], released for the MICCAI 2020 challenge, contains 113 TOF-MRA volumes collected from multiple hospitals. Each case is annotated voxel-wise by neuroradiologists. Following the official split, we use 93 volumes for training and 20 for testing. Aneurysm sizes range from 2 mm to 20 mm, making the dataset particularly challenging for small-lesion detection.

The IntrA dataset[41] includes 103 CTA volumes with voxel-level annotations. Aneurysms occur at diverse arterial locations (ICA, MCA, AComA, PComA, basilar), covering a wide morphological spectrum. Following ref. 41, we split the dataset into 73 training, 10 validation, and 20 testing cases. Compared with ADAM, IntrA emphasizes cross-location variability and contains both ruptured and unruptured aneurysms.

The CQ500 dataset[42], originally curated for cranial pathology detection, is a large-scale CTA cohort. A subset (~490 cases) has been re-annotated for vascular analysis, including IA masks from open-source repositories[43]. We randomly divide the data into 350 training, 70 validation, and 70 testing scans. CQ500 provides a robust benchmark for cross-domain evaluation, as it spans >20 hospitals in India with heterogeneous imaging protocols.

All CTA and TOF-MRA volumes are resampled to isotropic 0.5 mm spacing and intensity-clipped to the [0.5, 99.5] percentile of HU (CTA) or signal intensities (MRA). Vessels are enhanced with a multi-scale Frangi filter to generate vesselness maps $\mathbf{V}$, and centerlines $\mathbf{C}$ are extracted by skeletonization. Cropping to 128^3 or 160^3 regions around vessels is performed using vesselness-based bounding boxes. Data augmentation includes 3D rotation (± 10°), elastic deformation, intensity scaling, gamma correction, and cutout outside vascular regions. These steps standardize inputs and support consistent multi-dataset evaluation.

Experiments setup

All experiments were run on an NVIDIA A100 cluster (8 × A100 GPUs, 80 GB each) using CUDA 12.0 and PyTorch 2.2. To ensure reproducibility, we fixed random seeds across NumPy, PyTorch, and data loaders. Mixed-precision training (FP16) with Apex was applied to reduce memory usage. Depending on dataset size and task setup, training required 36–72 h.

The anatomy-guided MAE was pretrained on the combined training sets (ADAM, IntrA, CQ500) without labels. Volumes were resampled to isotropic 0.5 mm and cropped into 160^3 patches. We used a patch size of 16^3, embedding dimension $d = 768$, and 12 transformer blocks. The masking ratio was 75% with vessel- and boundary-biased sampling (Sec. 4.2). The decoder had four layers (width 512). Optimization used AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.05), an initial learning rate of 1.5×10^{-4} with cosine decay, and batch size 8 over 8 GPUs. Pretraining ran for 200 epochs with gradient accumulation of 2 steps.

For IA detection and segmentation, the encoder was initialized with pretrained weights and augmented with shared and instance-dynamic prompts (Sec. 4.3). Each block used $K_s = 8$ shared and $K_i = 8$ instance prompts. The segmentation head followed a UNet-like decoder, while the detection head adopted an anchor-free center heatmap. Optimization used AdamW with learning rate 1.0×10^{-4}, weight decay 1×10^{-4}, and batch size 4. Training ran for 150 epochs with early stopping on validation Dice. Loss weights were $\lambda_{\text{dice}} = 1.0$, $\lambda_{\text{bce}} = 0.5$, $\lambda_{\text{prompt}} = 1e - 4$, and $\lambda_{\text{align}} = 0.1$. Attention bias scaling (Eq. (6)) was initialized to 0.1 and updated during training.

To simulate domain shifts, datasets were divided by institution and protocol. We applied GS-EMA teacher-student training (Sec. 4.4) on unlabeled validation domains. EMA momentum was bounded in [0.90, 0.999] with adaptive coefficient $\kappa = 0.1$. The boundary-aware contrastive loss used temperature $\tau_c = 0.07$ and positive sampling threshold 0.7. Loss weights were $\lambda_{\text{con}} = 0.1$, $\lambda_{\text{cons}} = 0.05$, and $\lambda_{\text{bndry}} = 0.1$. Training ran for 80 epochs with initialization from Stage II. Strong and weak augmentations included 3D elastic deformation, random bias fields, Gaussian noise, and intensity jitter.

All stages used data augmentation: random rotation (± 10°), flipping, scaling (0.9–1.1 ×), elastic deformation, gamma correction, Gaussian noise ($\sigma = 0.01$–0.05), and cutout outside vessels. Dropout ($p = 0.1$) and stochastic depth ($p = 0.2$) were applied in transformers. Weight decay and prompt norm penalties (Eq. (2)) stabilized training.

At test time, we used sliding-window inference with 0.5 overlap and ensembled predictions from three augmentations. Outputs combined segmentation probability maps and detection heatmaps, followed by connected-component filtering (minimum volume > 50 voxels) to suppress

Table 10 | Key hyperparameters for baselines

Method	Learning Rate	Batch Size	Optimizer
nnU-Net	3e-4 (default)	2 (default)	AdamW
UNETR	1e-4	4	AdamW
Med-MAE	1e-4 (finetune)	4	AdamW
MedSAM	1e-5 (adapter)	2	Adam
DG SOTA	1e-4	4	AdamW
AMAP (ours)	1e-4 (finetune)	4	AdamW

false positives. Average inference time was ~0.9 s per scan on one A100 GPU, enabling near real-time use.

Evaluation metrics

We adopt standard metrics for segmentation, detection, and domain generalization.

Segmentation: **Dice Similarity Coefficient (DSC):**

$$\mathrm{DSC} = \frac{2|P \cap G|}{|P| + |G|},$$

where P and G are predicted and ground-truth masks. **95% Hausdorff Distance (HD95)**: 95th percentile of boundary distances between prediction and ground truth. **Sensitivity** and **Specificity**: proportion of correctly detected aneurysm voxels and correctly rejected non-aneurysm voxels.

Detection: **Sensitivity@FP/case**: sensitivity at fixed FP counts per case (0.5, 1, 2 FP/case), common in medical detection benchmarks. **FROC**: sensitivity-FP/case curve; we report FROC-AUC. **Average Precision (AP)**: lesion-level precision-recall AUC.

Cross-domain Generalization: **Domain-wise Dice and HD95**: broken down by test domain (e.g., ADAM → IntrA, IntrA → CQ500). **Relative Performance Drop (RPD)**:

$$\mathrm{RPD} = \frac{\mathrm{DSC}_{\mathrm{in}} - \mathrm{DSC}_{\mathrm{out}}}{\mathrm{DSC}_{\mathrm{in}}} \times 100\%,$$

where $\mathrm{DSC}_{\mathrm{in}}$ and $\mathrm{DSC}_{\mathrm{out}}$ are in-domain and cross-domain Dice. **Expected Calibration Error (ECE)**: measures calibration of detection confidence across domains.

All results are reported as mean ± standard deviation across test folds. Statistical comparisons use two-tailed paired t-tests, with $p < 0.05$ considered significant[12].

Baseline Configurations: To ensure fairness and reproducibility, all baselines follow public defaults or are tuned on our validation sets. Key hyperparameters are listed in Table 10.

Ethics approval and consent to participate

This study used publicly available, anonymized datasets (ADAM, IntrA, CQ500). As all data were fully de-identified and collected under existing ethical approvals by the original dataset providers, no institutional review board (IRB) approval or additional informed consent was required.

Data availability

The datasets analyzed in this study are publicly available: ADAM Challenge dataset (https://adam.isi.uu.nl/.), IntrA dataset[41], and CQ500 dataset[42]. Processed data used in this study are available from the corresponding author upon reasonable request. The implementation of AMAP, including training and evaluation scripts, will be made publicly available upon publication.

Code availability

The implementation of AMAP, including training and evaluation scripts, will be made publicly available upon publication.

Received: 29 September 2025; Accepted: 18 November 2025;
Published online: 08 December 2025

References

1. Wiebers, D. O. et al. Unruptured intracranial aneurysms: natural history, clinical outcome, and risks of surgical and endovascular treatment. *Lancet* **362**, 103–110 (2003).
2. Daga, P., Kumar, R. & Zhang, J. Deep learning for rupture risk stratification of cerebral aneurysms. *Stroke* **56**, 1003–1012 (2025).
3. Joo, B. Methodological challenges in deep learning-based detection of intracranial aneurysms: a scoping review. *Neurointervention* **20**, 52–65 (2025).
4. Nakao, T. et al. Deep neural network-based computer-assisted detection of cerebral aneurysms in mr angiography. *J. Magn. Reson. Imaging* **47**, 948–953 (2018).
5. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2021).
6. Dosovitskiy, A. et al. An image is worth 16x16 words: transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)* (IEEE, 2021).
7. He, K. et al. Masked autoencoders are scalable vision learners. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 15979−15988 (IEEE, 2022).
8. Zhou, T. et al. A universal medical image segmentation framework with vision foundation models. *Nat. Mach. Intell.* **5**, 780–791 (2023).
9. Shen, C. et al. Domain generalization for medical imaging classification with meta-learning. *Med. Image Anal.* **78**, 102406 (2022).
10. Vach, M., Richter, S. & Ecker, A. Reproducibility of deep learning-based intracranial aneurysm segmentation across institutions. *NeuroImag. Clin.* **44**, 103673 (2024).
11. Liu, F. et al. Small lesion detection in medical images using deep learning. *Pattern Recognit.* **115**, 107885 (2021).
12. Delfan, S., Mohammadi, N. & Rezaei, A. Artificial intelligence for brain aneurysm detection: a meta-analysis. *Eur. Radiol.* **35**, 2156–2170 (2025).
13. Holzinger, A. et al. Explainable ai methods in medical imaging. *Nat. Rev. Methods Prim.* **2**, 1–13 (2022).
14. Ceballos-Arroyo, A. M. et al. Anatomically-guided masked autoencoder pre-training for aneurysm detection. *arXiv* https://doi.org/10.48550/arXiv.2502.21244 (2025).
15. Zu, R., Zhao, M. & Zhang, H. Efficient prompt tuning for medical image classification. In *MICCAI*, 560–570 (Springer, 2024).
16. Lin, F. et al. Gs-ema: Integrating gradient surgery exponential moving average with boundary-aware contrastive learning for enhanced domain generalization in aneurysm segmentation. *arXiv* https://doi.org/10.48550/arXiv.2402.15239 (2024).
17. Ceballos-Arroyo, A. M. et al. Vessel-aware aneurysm detection using multi-scale deformable 3d attention. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 754–765 (Springer, 2024).
18. Li, Y. et al. Deep learning-based platform performs high detection sensitivity of intracranial aneurysms in 3d brain tof-mra: an external clinical validation study. *Int. J. Med. Inform.* **188**, 105487 (2024).
19. Kiewitz, J. et al. Deep learning-based multiclass segmentation in aneurysmal subarachnoid hemorrhage. *Front. Neurol.* **15**, 1490216 (2024).
20. Nishi, H. et al. Deep learning-based cerebral aneurysm segmentation and morphological analysis with three-dimensional rotational angiography. *J. NeuroInt. Surg.* **16**, 197–203 (2024).

21. Mazurowski, M. A. et al. Segment anything model for medical image analysis: an experimental study. *Radiol. Artif. Intell.* **5**, e230217 (2023).
22. Ma, J. et al. Segment anything in medical images. *Nat. Commun.* **15**, 6543 (2024).
23. Ma, J. et al. Segment anything in medical images. *Nat. Commun*. **15**, 654 (2023).
24. Dong, G. et al. An efficient segment anything model for the medical image segmentation (emedsam). *Sci. Rep.* **14**, 9425 (2024).
25. Wang, H. et al. Sam-med3d: Towards general-purpose segmentation models for volumetric medical images. *Sci. Rep*. **14**, 19425 (2023).
26. Wang, G. et al. Sam-med3d-moe: Towards a non-forgetting segment anything model via mixture of experts for 3d medical image segmentation. *arXiv* https://doi.org/10.48550/arXiv.2407.04938 (2024).
27. Ma, J. et al. Medsam2: Segment anything in 3d medical images and videos. *arXiv* https://doi.org/10.48550/arXiv.2504.03600 (2025).
28. Hartsock, J., Fang, E. & Xu, M. Vision-language models in radiology: a systematic review. *Insights Imag.* **15**, 64 (2024).
29. Wei, Y. et al. More-brain: Routed mixture of experts for interpretable and generalizable cross-subject fmri visual decoding. *arXiv* https://arxiv.org/abs/2505.15946 (2025).
30. Xiao, X. et al. Describe anything in medical images. *arXiv* https://arxiv.org/abs/2505.05804 (2025).
31. Xiao, X. et al. Hgtdp-dta: Hybrid graph-transformer with dynamic prompt for drug-target binding affinity prediction. In *International Conference on Neural Information Processing*, 340–354 (Springer, 2024).
32. Wei, Y. et al. 4d multimodal co-attention fusion network with latent contrastive alignment for alzheimer's diagnosis. *arXiv* https://arxiv.org/abs/2504.16798 (2025).
33. Tang, Y. et al. Swin unetr: swin transformers for semantic segmentation of brain tumors in mri images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2021. Lecture Notes in Computer Science*, (eds. Crimi, A., Bakas, S.) 12962 (Springer (Springer, 2022).
34. Xie, B., Zhang, K. & Liu, Z. Knowledge-guided masked image modeling for radiology representation learning. *Med. Image Anal.* **89**, 102907 (2024).
35. Oh, H., Park, J. & Lee, D. Meta-learning based domain generalization for medical image classification. *Med. Image Anal.* **87**, 102830 (2023).
36. Wang, D., Shelhamer, E., Liu, S., Olshausen, B. & Darrell, T. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)* (ICLR, 2021).
37. Gupta, A., Osman, I., Shehata, M. S., Braun, W. J. & Feldman, R. E. Medmae: A self-supervised backbone for medical imaging tasks. *Computation* **13**, 88 (2025).
38. Li, D., Yang, Y., Song, Y.-Z. & Hospedales, T. Learning to generalize: meta-learning for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3490−3497 (AAAI, 2018).
39. Jahanifar, M. et al. Domain generalization in computational pathology: survey and guidelines. *ACM Comput. Surv.* **57**, 1–37 (2025).
40. Timmins, K. et al. *The Aneurysm Detection And Segmentation (Adam) Challenge Dataset*. https://adam.isi.uu.nl/ (2020).
41. Qin, C. et al. Intra: An intracranial aneurysm dataset for clinical evaluation of detection and segmentation algorithms. *Sci. Data* **8**, 47 (2021).
42. Chilamkurthy, S. et al. Development and validation of deep learning algorithms for detection of critical findings in head ct scans. *Lancet* **392**, 2388–2396 (2018).
43. Yang, D., Xu, H. & Zhao, Y. *Open-Source Intracranial Aneurysm Annotations on the cq500 Dataset*. https://github.com/ycchen218/CQ500-IA (2022).

Acknowledgements

We would like to thank the contributors of the ADAM, IntrA, and CQ500 datasets for making their data publicly available, which enabled this research.This work was supported by the Beijing-Tianjin-Hebei Basic Research Cooperation Project (Grant No. H2024102009), the Natural Science Foundation of Beijing Municipality (Grant No. L242045), the Continuing Education Center of the National Health (Grant No. GWJJ2023100103), the Key-Area Research and Development Program of Guangdong Province (Grant No. 2023B0303030002), the STI 2030-Major Projects (Grant No. 2022ZD0208500), and the National Natural Science Foundation of China (Grant No. 62336002, 62406025).

Author contributions

M.H., T.L., J.Z. and X.S. contributed equally to this work. Y.D., T.Y., L.Y. and H.R.C. contributed equally to this work and are all corresponding authors. M.H., X.S., K.Z. and J.G. conceptualized the study, designed the methodology, and participated in securing research funding (Conceptualization, Methodology, Funding acquisition). T.L., X.C. and H.L.C. carried out data acquisition, curation, and investigation (Investigation, Data curation) and provided key resources, instruments, and technical support (Resources, Software). J.Z., M.L. and Y.S. drafted the initial manuscript and generated visualizations (Writing – Original Draft, Visualization). Y.D., T.Y., L.Y. and H.R.C. supervised the project, coordinated collaborations, and ensured administrative support (Supervision, Project administration). All authors contributed to reviewing and revising the manuscript critically for important intellectual content (Writing – Review & Editing) and approved the final version for submission.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Hengri Cong, Long Yan, Tianyi Yan or Yiming Deng.