Concept-to-Pixel: Prompt-Free Universal Medical Image Segmentation

Haoyun Chen[1,2,3†], Fenghe Tang[1,2,3†], Wenxin Ma[1,2,3], and Shaohua Kevin Zhou[1,2,3,4,5⋆]

[1] School of Biomedical Engineering, Division of Life Sciences and Medicine, University of Science and Technology of China (USTC), Hefei, Anhui, China 230026
[2] Medical Imaging, Robotics, Analytic Computing & Learning (MIRACLE) Lab, YRD-RIGHT, USTC Suzhou Institute for Advanced Research, Suzhou, Jiangsu, China
[3] Jiangsu Provincial Key Laboratory of Multimodal Digital Twin Technology, Suzhou, Jiangsu, China 215123
[4] Biomedical Basic Research Center (BBRC) of Jiangsu Province, Suzhou, Jiangsu, China 215123
[5] State Key Laboratory of Precision and Intelligent Chemistry, Hefei, Anhui, China
{chenhaoyun,fhtan9,wxma}@mail.ustc.edu.cn, {skevinzhou}@ustc.edu.cn

Abstract. Universal medical image segmentation seeks to use a single foundational model to handle diverse tasks across multiple imaging modalities. However, existing approaches often rely heavily on manual visual prompts or retrieved reference images, which limits their automation and robustness. In addition, naive joint training across modalities often fails to address large domain shifts. To address these limitations, we propose **Concept-to-Pixel (C2P)**, a novel prompt-free universal segmentation framework. C2P explicitly separates anatomical knowledge into two components: Geometric and Semantic representations. It leverages Multimodal Large Language Models (MLLMs) to distill abstract, high-level medical concepts into learnable Semantic Tokens and introduces explicitly supervised Geometric Tokens to enforce universal physical and structural constraints. These disentangled tokens interact deeply with image features to generate input-specific dynamic kernels for precise mask prediction. Furthermore, we introduce a Geometry-Aware Inference Consensus mechanism, which utilizes the model's predicted geometric constraints to assess prediction reliability and suppress outliers. Extensive experiments and analysis on a unified benchmark comprising eight diverse datasets across seven modalities demonstrate the significant superiority of our jointly trained approach, compared to universe- or single-model approaches. Remarkably, our unified model demonstrates strong generalization, achieving impressive results not only on zero-shot tasks involving unseen cases but also in cross-modal transfers across similar tasks. Code is available at: https://github.com/Yundi218/Concept-to-Pixel.

Keywords: Universal Segmentation · MLLM Distillation · Disentangled Representation · Dynamic Convolution · Medical Foundation Models

⋆ Corresponding author † Equal Contribution

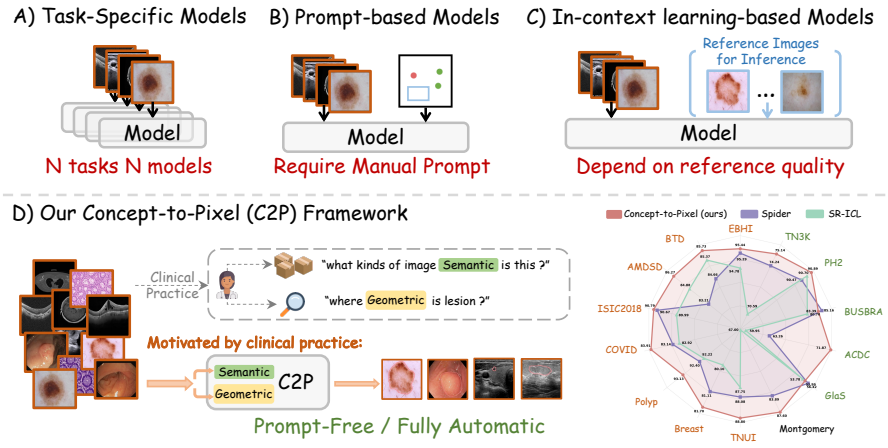

Fig. 1: Comparison of existing medical image segmentation paradigms and our proposed Concept-to-Pixel framework. Unlike existing paradigms that suffer from limited versatility **(A)**, limited manual prompts **(B)**, or reliance on reference quality **(C)**, our prompt-free model **(D)**, inspired by clinical diagnostic workflows, explicitly decouples modality semantics and physical geometry via `[SEM]` and `[GEO]`. The radar chart in **(D) Right** demonstrates that our zero-reference method (Concept-to-Pixel) significantly outperforms SOTA models (Spider and SR-ICL), exhibiting impressive in-domain performance and zero-shot generalization even on completely unseen modalities (*e.g.*, X-Ray).

1 Introduction

Medical image segmentation is an important task of computer-aided diagnosis [37, 46, 47, 49–51, 61], but the vast diversity of modalities and severe domain gaps in tissue contrast and noise distributions have historically confined the field to specialized, modality-specific solutions [23, 30, 39, 48, 59]. Beyond their computational cost and poor scalability, these "one-model-one-task" architectures fundamentally fail to exploit cross-modal anatomical knowledge for mutual performance gains. This has inspired a paradigm shift toward universal medical segmentation models [7, 9, 33, 38, 60], which aim to unify diverse segmentation tasks across modalities within a single framework.

However, building such a unified model that matches or exceeds specialized counterparts remains extremely difficult. As shown in Fig. 1 (B,C), existing approaches fall into three categories: (1) **Prompt-based methods** [13, 32] excel at interactive boundary delineation but rely heavily on manual visual prompts (points or boxes), lacking the semantic understanding necessary for fully automatic lesion and anatomy discovery. (2) **In-context learning (ICL) methods** [9, 60] perform segmentation by conditioning on reference image-mask pairs, but this inference-heavy paradigm suffers from high variance and depends critically on the quality and relevance of retrieved support examples, making it difficult to globally internalize medical knowledge. (3) **Naive joint training** on large-scale

heterogeneous datasets frequently induces negative transfer: conflicting low-level visual cues from different input sources drive gradients in opposing directions [53]. For example, a dark hypoechoic region signals malignancy in ultrasound yet represents healthy fluid in MRI, forcing the network to implicitly reconcile irreconcilable pixel-level patterns and ultimately causing feature collapse.

Previous work [10, 56] reveals that the root cause of this collapse is that the network is forced to reconcile conflicts at the pixel level, without any structured intermediate representation. Structured clinical diagnosis, by contrast, naturally avoids such confusion by evaluating *disentangled concepts*: a clinician assesses a lesion by implicitly decoupling its *Geometry*, including modality-agnostic physical properties such as shape, location, and boundaries, from its *Semantics*, which is modality-specific textural appearance and internal signal characteristics [18, 28, 29, 57]. We illustrate the procedure in Fig. 1 (D).

Motivated by this diagnostic principle, we propose **Concept-to-Pixel (C2P)**, a novel framework that fundamentally mitigates negative transfer by explicitly disentangling anatomical understanding into separate Geometric and Semantic representations. Specifically, to enforce universal shape priors, we introduce learnable Geometric Tokens, marked as `[GEO]`, that are explicitly supervised to regress physical attributes (*e.g.*, area, centroid), imposing a strict structural constraint largely ignored by conventional U-Nets. For semantic understanding, we leverage Multimodal Large Language Models (MLLMs) to generate and extract fine-grained concepts from clinical reports (*e.g.*, "irregular margin", "heterogeneous texture"). We distill these text concepts into Semantic Tokens marked as `[SEM]`. Subsequently, `[GEO]` and `[SEM]` serve as the sample-aware priors, interacting with visual features through a Token-Guided Dynamic Head (TGDH) to generate sample-aware dynamic segmentation head [26, 52]. Furthermore, we introduce an intrinsic Geometry-Aware Inference Consensus mechanism, utilizing the explicitly regressed geometric priors to self-evaluate prediction reliability across perturbed views and confidently suppress outliers. Together, these designs establish C2P as a principled and generalizable framework. Extensive experiments have proven that our method achieves state-of-the-art performance across eight datasets spanning seven modalities with a single universal model. Remarkably, our C2P demonstrates strong zero-shot generalization on five unseen datasets without relying on any references or prompts, and attains competitive performance even on entirely unseen modalities.

Our main contributions are summarized as follows:

- We propose Concept-to-Pixel (C2P), a universal segmentation framework that decouples anatomical understanding into `[GEO]` (supervised by physical attributes) and `[SEM]` (distilled from MLLM-extracted clinical concepts), enabling modality-agnostic representations that mitigate negative transfer in heterogeneous multi-modal training.
- While the learned tokens capture universal geometric and semantic representations, we design a Token-Guided Dynamic Head (TGDH) that conditions on per-sample tokens to generate instance-specific convolution kernels, enabling

the model to adapt its segmentation behavior to each individual sample rather than applying a shared static head across all targets.
– Extensive experiments on eight datasets across seven modalities show that C2P, as a *single* universal model, achieves 88.22% average Dice, surpassing both state-of-the-art universal methods and the task-specialized nnU-Net V2. Additionally, our C2P demonstrates strong generalization ability without requiring additional prompts or references, achieving competitive results on unseen datasets and modalities.

2 Method

2.1 Overview

We propose Concept-to-Pixel (C2P), a universal segmentation framework that explicitly disentangles anatomical understanding into Geometric and Semantic representations to mitigate negative transfer across modalities. Unlike conventional U-Nets that implicitly entangle these features, C2P leverages a deeply guided token-interaction mechanism. As illustrated in Fig. 2, our framework consists of four key components: (1) Token Formulation & Modality Injection: Initializes universal `[GEO]` and MLLM-distilled `[SEM]`, utilizing a Style-Content Fusion Module (SCFM) to resolve cross-modality ambiguities. (2) Bidirectional Token-Image Interaction: Enables deep interaction via Cross Attention, where tokens query image features while simultaneously guiding them with high-level priors. (3) Token-Guided Dynamic Head (TGDH): Synthesizes instance-specific convolutional kernels driven by the aggregated tokens and global context for precise mask prediction. (4) Geometry-Aware Inference Consensus: An intrinsic evaluation mechanism leveraging the physical self-consistency of `[GEO]` to filter unreliable predictions.

2.2 Disentangled Concepts to Pixel-Level Segmentation

The core philosophy of our framework is to translate abstract, modality-agnostic physical and clinical expectations into specific pixel-level decision boundaries. To achieve this, we design a coherent three-stage process: Disentangled Concept Extraction, Bidirectional Token-Image Interaction, and Token-Guided Dynamic Prediction.

Disentangled Token Formulation. To bridge the gap between low-level visual features and high-level medical concepts, we define two distinct sets of tokens: (i) Geometry Tokens `[GEO]` ($\mathcal{T}_{geo}$). To capture the universal physical attributes of lesions, we initialize a set of N_{geo} learnable tokens $\mathcal{T}_{geo} \in \mathbb{R}^{N_{geo} \times C}$. Crucially, instead of treating these tokens as latent black boxes, each token is explicitly and strictly supervised to regress one of the N_{geo} specific geometric properties. This explicit structural constraint forces the model to encode modality-agnostic "shape" and "location" physics, ensuring that a topologically identical lesion is recognized consistently regardless of its imaging modality. (ii) Semantic Tokens `[SEM]` ($\mathcal{T}_{sem}$).

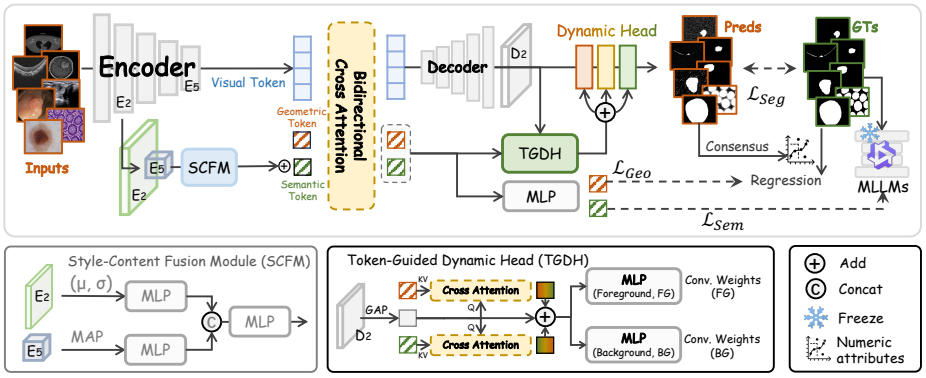

Fig. 2: Overview of the proposed Concept-to-Pixel (C2P) framework. (1) The model first extracts shallow (E2) and deep (E5) backbone features to infer modality embeddings via the Style-Content Fusion Module (SCFM), which are dynamically injected into the `[SEM]`. (2) The universal `[GEO]` and modality-aware `[SEM]` interact with visual tokens through bidirectional Cross Attention. (3) The aggregated concepts are then fed into the Token-Guided Dynamic Head (TGDH) to generate instance-specific parameters for the final Dynamic Head to predict precise masks. To guarantee explicit decoupling, the tokens are deeply supervised by physical properties ($\mathcal{L}_{Geo}$) and MLLM-distilled knowledge ($\mathcal{L}_{Sem}$).

To inject medical semantic knowledge, we leverage the reasoning capability of Large Multi-modal-Language Models, generating structured clinical descriptions for training images across N_{sem} explicit diagnostic dimensions. These texts are encoded into embeddings via PubMedBERT [21]. We initialize N_{sem} learnable `[SEM]` $\mathcal{T}_{sem} \in \mathbb{R}^{N_{sem} \times C}$ and align them with the MLLM-derived text embeddings using a contrastive loss. This explicit cross-modal alignment effectively distills the structured clinical semantics into the network's latent space, enabling the model to segment targets based on expert-level diagnostic concepts rather than mere pixel intensities.

Style-Content Fusion Module (SCFM). While geometry is essentially universal in segmentation, semantics are highly modality-dependent. To disentangle semantics, we introduce the Style-Content Fusion Module (SCFM). Specifically, we extract the channel-wise mean μ (*i.e.*, representing brightness/tone) and standard deviation σ (*i.e.*, representing texture/noise) from the shallow backbone features ($E2$) as the Style statistics. Concurrently, we extract the global context from the deep features ($E5$) via global average pooling (GAP) as the Content. These representations are compressed and fused via a gated MLP to infer a dense modality embedding $V_{modality}$. We then selectively inject this modality bias *only* into the `[SEM]`:

$$\widehat{\mathcal{T}}_{sem} = \mathcal{T}_{sem} + V_{modality}. \tag{1}$$

Bidirectional Token-Image Interaction. To establish a deep mapping between the abstract concepts and the concrete visual evidence, we propose a multi-layer

bidirectional interaction mechanism. First, the deepest visual features from the backbone (F_{E5}) are flattened and projected to the token dimension d. To preserve spatial topology, a learnable 2D positional encoding is added to F_{E5}. Concurrently, the geometric and modality-injected `[SEM]` are concatenated as $\mathcal{T} = [\mathcal{T}_{geo}, \widehat{\mathcal{T}}_{sem}] \in \mathbb{R}^{(N_{geo}+N_{sem})\times d}$. These representations then undergo an iterative, two-phase Cross Attention process:

Token-to-Image Attention (Abstraction). In this phase, the concept tokens act as queries (Q) to scan the global image features (K, V). This allows each token to adaptively aggregate relevant instance-specific visual evidence from the image (*e.g.*, the "margin" token searches for edge-like features). Structurally, this is implemented via Multi-head Cross-Attention (MCA) and a Feed-Forward Network (FFN) with residual connections:

$$\mathcal{T}' = \mathcal{T} + \mathrm{MCA}(\mathrm{LN}(\mathcal{T}), F_{E5}, F_{E5}), \tag{2}$$

$$\mathcal{T}_{out} = \mathcal{T}' + \mathrm{FFN}(\mathrm{LN}(\mathcal{T}')), \tag{3}$$

where $\mathrm{LN}(\cdot)$ denotes Layer Normalization. Through this process, the tokens transition from static global priors to dynamic, instance-aware concept representations.

Image-to-Token Attention (Guidance). Conversely, the image features must be refined by the high-level concepts to suppress irrelevant background noise. Here, the image features act as queries to retrieve information from the updated tokens dictionary:

$$F'_{E5} = F_{E5} + \mathrm{MCA}(\mathrm{LN}(F_{E5}), \mathcal{T}_{out}, \mathcal{T}_{out}), \tag{4}$$

$$F_{out} = F'_{E5} + \mathrm{FFN}(\mathrm{LN}(F'_{E5})). \tag{5}$$

By querying the concept tokens, the pixel-level features explicitly highlight regions that align with the predefined geometric and semantic expectations, yielding deeply coupled representations.

Token-Guided Dynamic Head (TGDH). To overcome the rigidity of static segmentation heads, which apply fixed weights across highly heterogeneous modalities, we propose the Token-Guided Dynamic Head (TGDH). This mechanism synthesizes image-specific convolutional parameters directly from the disentangled concept tokens. First, the refined image features undergo a decoder pathway to produce high-resolution representations D_2. We apply global average pooling to D_2 to obtain a global image query Q_{img}. This query is then fed into two parallel Cross-Attention modules to aggregate the most relevant insights from the final `[GEO]` and `[SEM]`:

$$\mathcal{F}_{geo} = \mathrm{CrossAttn}(Q_{img}, \mathcal{T}_{geo,out}), \quad \mathcal{F}_{sem} = \mathrm{CrossAttn}(Q_{img}, \mathcal{T}_{sem,out}). \tag{6}$$

The aggregated token features and the global image query are concatenated and processed by a Kernel Generator MLP (Ψ) to synthesize dynamic parameters for both foreground (fg) and background (bg) convolutions:

$$\mathcal{W}_{fg}, \mathcal{W}_{bg} = \Psi(\mathrm{Concat}(\mathcal{F}_{geo}, \mathcal{F}_{sem}, Q_{img})). \tag{7}$$

Finally, the segmentation masks are generated by the Dynamic Head, which applies these dynamically generated convolutional weights $\mathcal{W}$ directly onto the high-resolution feature map D_2. This ensures that the pixel-level decision boundary is dynamically governed by explicit physical and clinical reasoning.

2.3 Multi-Task Supervision Strategy

To ensure the tokens genuinely learn the disentangled representations, we optimize the network using a multi-task objective $\mathcal{L}_{total} = \lambda_{seg}\mathcal{L}_{seg} + \lambda_{geo}\mathcal{L}_{geo} + \lambda_{sem}\mathcal{L}_{sem}$.

- **Segmentation Loss ($\mathcal{L}_{seg}$):** We employ a combination of a Boundary-Aware Structure Loss ($\mathcal{L}_{struct}$) and a Dice Loss ($\mathcal{L}_{dice}$) for both foreground (fg) and background (bg) predictions:

$$\mathcal{L}_{seg} = \sum_{c\in\{fg,bg\}} \big(\mathcal{L}_{struct}(P_c, M_c) + \mathcal{L}_{dice}(P_c, M_c)\big), \tag{8}$$

  where P_c and M_c denote the predicted probabilities and ground-truth masks for class c. To compute $\mathcal{L}_{struct}$, we extract lesion boundaries by computing the absolute difference between the mask and its average-pooled version, assigning dynamic spatial weights W to hard-to-segment edge regions:

$$W = 1 + 5 \cdot |\text{AvgPool}_{31\times31}(M_c) - M_c|. \tag{9}$$

  This weight map is synergistically applied to emphasize boundary regions. To further address class imbalance and focus on hard-to-classify pixels, we combine this spatial weight with a Focal Loss ($\gamma = 2.0$) and an Intersection over Union (IoU) loss, formulating the structure loss as:

$$\mathcal{L}_{struct} = \mathcal{L}^W_{Focal}(P_c, M_c) + \mathcal{L}^W_{IoU}(P_c, M_c). \tag{10}$$

- **Geometric Regression Loss ($\mathcal{L}_{geo}$):** We apply the Mean Squared Error (MSE) loss across all N_{geo} geometric properties (including bounding box, area, perimeter, *etc.*). This strictly supervises the `[GEO]` to capture the physical ground truth:

$$\mathcal{L}_{geo} = \frac{1}{N_{geo}} \sum_{i=1}^{N_{geo}} \text{MSE}(\mathcal{T}^{(i)}_{geo}, g^{(i)}), \tag{11}$$

  where $g^{(i)}$ represents the normalized ground-truth scalar for the i-th geometric property.
- **Semantic Alignment Loss ($\mathcal{L}_{sem}$):** To inject abstract medical knowledge, we compute the Cosine Embedding Loss between the learned `[SEM]` and the pre-computed text embeddings derived from large multi-modal models:

$$\mathcal{L}_{sem} = \frac{1}{N_{sem}} \sum_{j=1}^{N_{sem}} \big(1 - \cos(\mathcal{T}^{(j)}_{sem}, E^{(j)}_{text})\big), \tag{12}$$

  where $E^{(j)}_{text}$ is the structured clinical embedding for the j-th diagnostic dimension. This forces the latent space to seamlessly align with expert medical semantics.

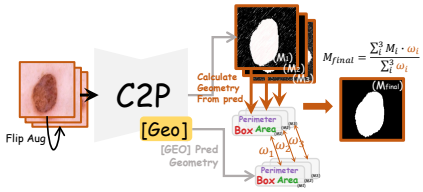

Fig. 3: Workflow of Geometry-Aware Inference Consensus

2.4 Geometry-Aware Inference Consensus

Standard inference paradigms often treat all predictions equally, making them vulnerable to ambiguous boundaries or severe imaging artifacts. To enhance the robustness of our framework, we propose a Geometry-Aware Inference Consensus mechanism that utilizes the explicitly learned `[GEO]` as intrinsic quality inspectors.

As shown in Fig. 3, our core insight relies on geometric self-consistency: a reliable prediction must satisfy strict physical constraints, where both the pixel-level area and spatial location of the generated mask tightly match the abstract, regression-level expectations predicted by the `[GEO]`. During inference, given multiple perturbed views x_i of an input image, the model outputs the segmentation mask M_i, along with a regressed area a_i^{reg} and a centroid c_i^{reg}. We then calculate the actual pixel-wise area a_i^{px} and centroid c_i^{px} directly from M_i. An inconsistency penalty is computed based on both scale and location shifts, and the confidence weight w_i for each view is derived as:

$$w_i = \exp\left(-\lambda \cdot (10 \cdot |a_i^{reg} - a_i^{px}| + \|c_i^{reg} - c_i^{px}\|_2)\right), \tag{13}$$

where λ controls the penalty sharpness. This weighting strategy heavily penalizes predictions where the network's abstract geometric reasoning contradicts its concrete visual output. Furthermore, an explicit false-positive suppression is applied to masks with contradicting near-zero predicted areas. The final prediction is dynamically aggregated as $M_{final} = (\sum w_i M_i)/(\sum w_i)$, effectively filtering out unreliable outliers based purely on multidimensional physical self-consistency.

3 Experiments

3.1 Datasets and Implementation Details

Datasets. Aiming to conduct fair comparisons with the previous works [9], we conduct our experiments on the same 8 public datasets, spanning 7 distinct imaging modalities. This comprehensive benchmark includes AMD-SD [22] (OCT), BTD [11, 12] (Brain MRI), EBHI-Seg [41] (Pathology), TNUI-2021 [62] and BUSI [1] (Ultrasound), Colon Polyp [4,25,42,45,54] (Endoscopy), COVID-19 [17] (CT), and ISIC 2018 [14] (Dermoscopy). Furthermore, to comprehensively evaluate the zero-shot generalization capability of our model, we extend our evaluation

Table 1: Comparison of different methods on multiple datasets (Dice %). The abbreviations Path., Endo., and Dermo. correspond to Pathology, Endoscopy, and Dermoscopy, respectively. The best results are highlighted in **bold**, and the second-best results are underlined.

Methods (#Volume)	Path.	Ultrasound		Endo.	CT	Dermo.	OCT	MRI	Avg
	EBHI	TNUI	Breast	Polyp	COVID	ISIC2018	AMDSD	BTD	
Specialized Models (One model for one task)									
UNet (MICCAI'25) [39]	94.61	85.11	73.00	87.66	78.62	87.76	85.57	80.92	84.16
AURA-Net (ISBI'21) [15]	95.08	87.55	79.34	90.53	79.28	89.92	85.91	84.75	86.54
UTANet (AAAI'25) [35]	94.82	86.79	78.78	89.75	79.14	89.32	85.72	83.57	85.99
Swin-UMamba (MICCAI'24) [30]	95.14	87.86	81.35	91.42	82.07	90.40	85.51	84.80	87.32
LGMSNet (ECAI'25) [16]	95.07	86.40	79.80	89.06	79.92	90.02	85.02	83.63	86.12
MEGANet (CVPR'25) [6]	95.06	87.96	80.30	91.29	81.70	90.45	85.89	84.73	87.17
RWKV-UNet (arXiv'25) [59]	95.26	88.21	81.40	91.41	81.41	90.55	85.65	85.15	87.38
nnUNetV2 (NM'21) [23]	95.26	88.22	78.20	90.47	79.44	87.54	**88.01**	**86.32**	86.68
Universal Models (One model for all tasks)									
UniverSeg (CVPR'23) [7]	91.28	71.75	66.04	63.18	38.69	82.57	68.30	63.37	68.15
Tyche (CVPR'24) [38]	94.20	85.00	73.39	83.47	74.18	88.24	82.44	79.34	82.53
Spider (ICML'24) [60]	95.29	88.08	81.11	92.40	83.14	90.67	83.21	84.66	87.32
SR-ICL (CVPR'25) [9]	94.78	87.75	80.10	92.22	82.92	89.99	84.88	85.37	87.25
C2P (ours)	**95.44**	**88.80**	**81.70**	**93.13**	**83.91**	**90.79**	86.27	85.73	**88.22**

to 8 additional datasets. Specifically, for unseen datasets within seen modalities, we utilize TN3K [20] and BUSBRA [19] (Ultrasound), PH2 [34] (Dermoscopy), ACDC [5] (MRI), and GlaS [43] (Pathology). For entirely unseen modalities, we test on Montgomery [24] and Covidquex [44] (X-Ray), as well as ISBI EM [8] (Electron Microscopy). Details of datasets are provided in the Appendix.

Implementation Details. To ensure fair and rigorous comparisons, we strictly conduct same experimental settings (*e.g.*, data splits, data augmentations, and evaluation criteria). We implemented our framework in PyTorch [36] with ConvNeXt-B [31, 58] as the backbone, setting the input resolution to 384×384. The model is trained for 50 epochs using the Adam [27] optimizer with a batch size of 8 and a cosine decayed initial learning rate of $1e-4$. To provide explicit geometric supervision, 9 physical attributes for `[GEO]` (*i.e.*, bounding box, area, perimeter, aspect ratio, compactness, centroid, eccentricity, orientation, and solidity) are pre-computed and extracted offline directly from the ground-truth (GT) masks. For semantic knowledge, we employ Qwen3-VL-Plus [2, 3, 55] to generate structured medical reports across 9 explicit clinical dimensions (*i.e.*, morphology, margin definition, internal texture, surrounding interaction, boundary distinctness, malignancy risk, pathological inference, differential reasoning, and predicted diagnosis). These reports are subsequently encoded into 768-dimensional token embeddings using PubMedBERT [21]. Detailed offline knowledge pre-processing and configurations are provided in the Appendix.

Comparison Baselines. We compare our C2P with baselines divided into two categories: (1) Specialized Models: UNet [39], nnUNetV2 [23], and a curated selection of top-tier models by U-Bench [46]: RWKV-UNet [59], AURA-Net [15], UTANet [35], MEGANet [6], Swin-UMamba [30], and LGMSNet [16]. (2) Universal Models: UniverSeg [7], Tyche [38], Spider [60], and SR-ICL [9].

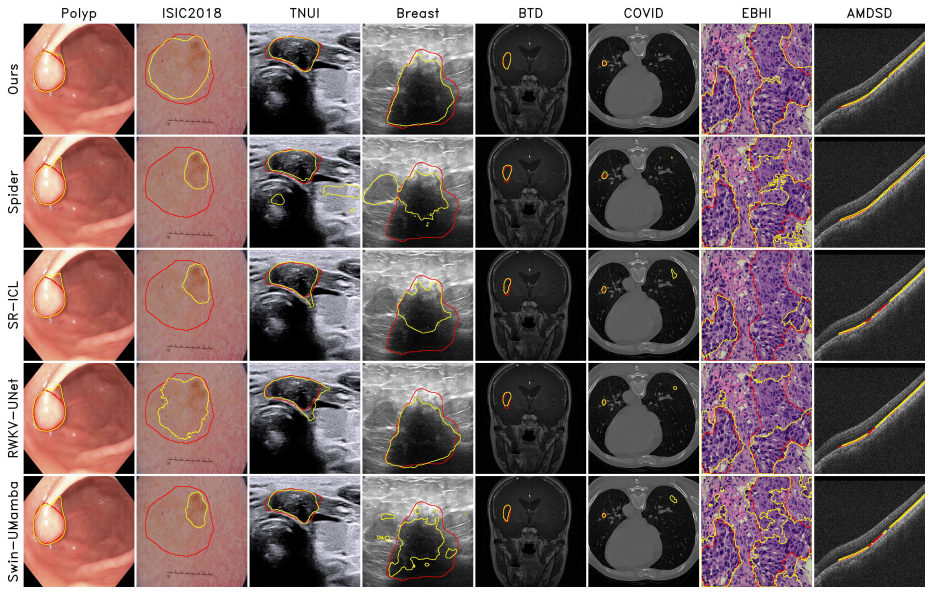

Fig. 4: Qualitative comparison on representative challenging cases. From left to right: Colon Polyp, ISIC2018, TNUI, Breast, BTD, COVID, EBHI, and AMDSD. The red contours indicate the Ground Truth masks, and the yellow contours denote the model predictions.

3.2 Comparison with State-of-the-Arts

Outperforming Specialized and Universal Baselines. As shown in Tab. 1, our method achieves an average Dice of 88.22%, surpassing strong specialized method nnU-NetV2 and RWKV-UNet. Compared to universal models, our C2P demonstrates a clear performance leap, outperforming the previous SOTA (Spider) by 0.90% in average Dice. Unlike existing methods (*e.g.*, UniverSeg and Tyche) that struggle with extreme domain shifts, our approach excels in challenging tasks such as Polyp (+0.73%) and COVID-19 (+0.77%) segmentation. This suggests that our design successfully improves the recognition of diverse textural and morphological variations compared to previous methods. Notably, in the modalities with fuzzy and blurred boundaries (e.g. Ultrasound, CT), as shown in Fig. 4, our method still predicts smooth, highly accurate, and geometrically plausible boundaries, setting new SOTAs across different datasets.

Strong Zero-Shot Generalization. We evaluate our method under a zero-shot inference setting, where no new `[GEO]`, `[SEM]`, or external MLLM text prompts are introduced or generated for unseen data. The evaluation is categorized into two scenarios: generalization to unseen datasets (intra-domain shift) and generalization to entirely unseen modalities. We compare our approach against the strongest specialized baseline nnUNetv2 and two competitive universal baselines:

Table 2: Zero-shot generalization results on unseen datasets and unseen modalities. C2P is evaluated without any support images, new tokens, or MLLM prompts for unseen data, and compared against specialized (nnUNetv2) and universal baselines (Spider with 64-shot references, SR-ICL with iterative refinement). The best results are highlighted in **bold**, and the second-best results are underlined. C2P achieves competitive performance on unseen datasets and outperforms all baselines on unseen X-ray and Ultrasound modalities, showing its generalization ability.

Dataset	Modality	nnUNetV2 [23]	Spider [60]	SR-ICL [9]	**C2P (ours)**
		(Zero-Ref)	(64-Ref)	(Self-Ref)	(Zero-Ref)
Generalization to Unseen Datasets (Seen Modalities)					
TN3K [20]	Ultrasound	52.61	74.24	70.59	**75.14**
PH2 [34]	Dermoscopy	56.39	90.47	90.70	**90.89**
ACDC [5]	MRI	29.17	63.26	59.95	**71.87**
BUSBRA [19]	Ultrasound	78.83	**85.16**	80.70	83.39
GlaS [43]	Pathology	**70.92**	56.42	53.78	53.99
Generalization to Unseen Modalities					
Montgomery [24]	X-Ray	-	83.89	67.00	**87.60**
Covidquex [44]	X-Ray	-	50.52	46.80	**53.31**
ISBI EM [8]	Electron Microscopy	-	64.63	**83.21**	26.79

Spider, which utilizes an extensive support set of 64 reference images, and SR-ICL with its self-refinement mechanism.

We first evaluate the models on five unseen datasets (TN3K, PH2, BUSBRA, ACDC, GlaS). As shown in Tab. 2, unlike Spider with 64-reference images or SR-ICL with two-stage iterative refinement, our model achieves competitive performance without requiring support images in most modalities (e.g. Dice of 75.15% in TN3K, 71.87% in ACDC). This striking success demonstrates that our `[SEM]` robustly adapt to the varying pixel intensity distributions, and `[GEO]` successfully capture essential structural priors.

Then, we evaluate on entirely unseen modalities (Montgomery X-ray [24], Covidquex and ISBI Electron Microscopy (ISBI EM) [8]). On the unseen X-ray modality dataset Montgomery and Covidquex, SR-ICL suffers a catastrophic failure, dropping to 67.00% and 46.80% Dice. Due to its self-refinement, when faced with a completely unseen modality, the massive domain gap renders the reference images ineffective. In contrast, our Concept-to-Pixel achieves a remarkable result, with Dice scores of 87.6% and 53.31%. We argue that `[GEO]` show their effectiveness in generalization: by learning compact shape priors across diverse lesion types during training, the model can recognize and segment unseen lesions that share similar geometries, even across previously unseen modalities. Conversely, our model struggles on the ISBI EM dataset (26.79% Dice). Since all 8 training datasets consist predominantly of macroscopic, compact structures (*e.g.*, polyps, nodules, tumors), the `[GEO]` have been trained to encode a strong compactness prior. This prior is incompatible with the filamentous, highly elongated topology

Table 3: Ablation studies of Concept-to-Pixel. (Static): Linear Segmentation Head. The best results are highlighted in **bold**.

ID	Geometry Tokens	Semantic Tokens	Dynamic Head	Inference Strategy	Partial Datasets (%)		Avg Dice (%)
					AMDSD	COVID	
1	-	-	- (Static)	None	82.19	76.27	82.38
2	-	✓	- (Static)	None	84.93	80.27	87.15
3	✓	-	- (Static)	None	84.98	82.66	87.31
4	-	-	✓	None	84.89	82.45	87.19
5	✓	-	✓	None	85.35	81.34	87.29
6	-	✓	✓	None	85.00	82.27	87.43
7	✓	✓	✓	None	86.07	83.49	87.79
8	✓	✓	✓	Standard TTA	86.20	83.60	88.16
9	✓	✓	✓	Geometry-Aware	**86.27**	**83.91**	**88.22**

of neuronal structures in EM images, causing geometric supervision to actively conflict with the correct prediction. This failure case is consistent with the design intent of C2P: the explicit geometric constraints are effective precisely because they are strict, and will generalize poorly when the target topology fundamentally violates the training distribution.

3.3 Ablation Studies

To investigate the contribution of each component in our Concept-to-Pixel framework, we conduct extensive ablation studies on the combined 8-dataset benchmark and report the average Dice scores. Our ablations start with a standard UNet with ConvNeXtV2-based backbone (ID 1).

Impact of Disentangled Concepts. The most significant performance leap comes from the introduction of explicit concept tokens. As shown in Tab. 3, the baseline model (ID 1) struggles with an Dice of 82.38%, suffering from negative transfer across diverse modalities. However, simply injecting `[SEM]` (ID 2) or `[GEO]` (ID 3) boosts the Dice to 87.15% and 87.31%, respectively. Remarkably, the static head guided by `[GEO]` (ID 3, 87.31%) even outperforms the blind Dynamic Head (ID 3 and 4, 87.19%). This strongly validates our core hypothesis: *explicit medical knowledge (what to look for and where) is more critical than complex architectural changes alone.*

Synergy of Dynamic Convolution and Dual Tokens. While tokens provide strong priors, the static segmentation head limits adaptability. By employing Dynamic Convolution conditioned on both token types (ID 7), we achieve a further improvement to 87.79%. Comparing ID 7 against single-token variants (ID 5 and 6), the combination of geometric constraints and semantic reasoning yields a synergistic effect (+0.50% Dice over ID 5), proving that "shape" and "texture" are complementary information streams that the dynamic head can effectively utilize to synthesize instance-specific weights.

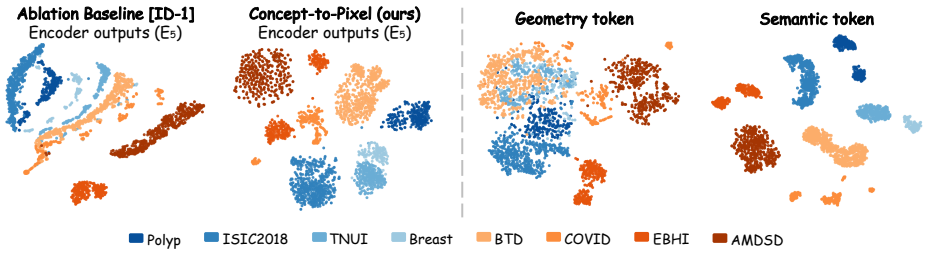

Fig. 5: In-domain representation analysis. (Left) t-SNE of encoder outputs (E5 features). Without C2P, features from different datasets show entanglements. Our C2P effectively separates the features. **(Right) t-SNE of `[GEO]` and `[SEM]`.** Some geometry tokens are clustered, since samples from different dataset may have similar geometry attributes. In contrast, semantic tokens are clearly separated, indicating that each token has successfully encoded discriminative semantic concepts tied to its corresponding domain.

Effectiveness of Geometry-Aware Consistency. Finally, we analyze the inference strategy. Standard Test-Time Augmentation [40] (ID 8), which averages predictions from augmented views, improves Dice to 88.16%. Our Geometry-Aware Consistency (ID 9) further pushes the performance to 88.22%. Unlike blind averaging, our method utilizes the regressed geometric attributes to penalize inconsistent predictions (e.g., when the predicted mask area contradicts the token's expectation). This demonstrates that the disentangled geometric knowledge serves not only as a training constraint but also as a reliable "self-check" mechanism during deployment.

3.4 Representation and Generalization Analysis

We further investigate the functional impact of our `[SEM]` and `[GEO]` by visualizing their feature spaces and dependency structures using t-SNE and attention maps, respectively. Our findings are summarized in this section.

In-Domain Representation Analysis. As shown in Fig. 5 (Left), we first visualize the deepest backbone representations (*i.e.*, E5) when the `[GEO]` and `[SEM]` supervision is removed (same as the configuration of the ablation ID 1). It shows that, lacking explicit concept guidance, the pure backbone features suffer from severe domain bias. In contrast, when introducing our C2P, the network intrinsically learns more cohesive modality representations. Furthermore, we provide t-SNE visualizations for both `[SEM]` and `[GEO]` in Fig. 5 (Right) to investigate their latent structures. The `[GEO]` exhibit an entangled and overlapping distribution across different domains, effectively verifying their capacity to extract modality-invariant features and universally shared concepts. In contrast, the `[SEM]` are distilled into highly compact and distinct clusters with minimal intra-class variance. This clear separation indicates that the `[SEM]` successfully capture modality-specific characteristics.

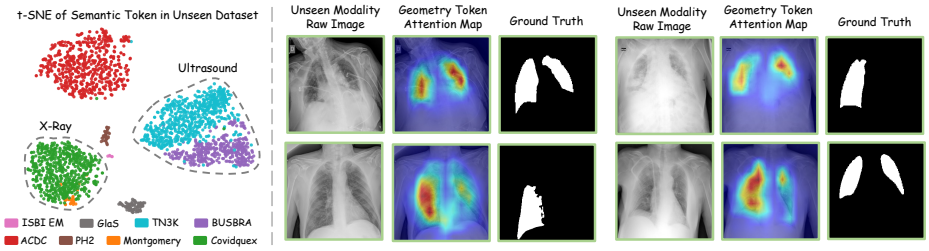

Fig. 6: Generalization Capability Analysis. (Left) t-SNE analysis of [SEM] in unseen dataset. Although our model has never seen the datasets, the result shows clear separation among modalities, while different datasets also belong to different clusters. **(Right) Visualization of the attention map between [GEO] and visual tokens.** Our [GEO] shows an attention pattern that is highly correlated with the ground truth mask.

Generalization Capability Analysis. We evaluate C2P's generalization by visualizing [SEM] and [GEO] on unseen datasets via t-SNE and attention maps, respectively. As shown in Fig. 6 (Left), [SEM] form cohesive clusters on unseen data. Notably, samples from different sources within the same modality (BUSBRA and TN3K, both Ultrasound) are mapped to the same latent region. Even for a completely unseen modality (X-Ray), [SEM] maintain clear clustering, suggesting that C2P develops generalizable modality-aware representations beyond its training distribution. As shown in Fig. 6 (Right), [GEO] accurately localize task-relevant regions on the Covidquex dataset (COVID-19 X-Ray, a completely unseen modality). The attention maps show that [GEO] simultaneously captures global lung structure and concentrates on infected regions, demonstrating that the learned geometric priors transfer effectively to unseen modalities.

4 Conclusion

In this paper, we presented Concept-to-Pixel (C2P), a novel prompt-free universal medical image segmentation framework. To overcome the long-standing challenge of negative transfer across heterogeneous imaging modalities, C2P explicitly disentangles anatomical reasoning into modality-agnostic [GEO] and MLLMs-distilled [SEM]. By inferring and injecting modality awareness solely into the semantic space and leveraging a bidirectional token-image transformer, our model generates instance-specific convolutional kernels driven entirely by explicit physical and clinical concepts. Furthermore, we introduced an intrinsic Geometry-Aware Inference Consensus mechanism to confidently filter unreliable predictions based on physical self-consistency during inference.

Extensive experiments across 8 diverse datasets demonstrate that C2P not only successfully suppresses negative transfer but remarkably outperforms state-of-the-art universal methods and strong task-specialized baselines. Notably, our framework exhibits exceptional zero-shot generalization to completely unseen

macroscopic modalities without requiring external reference images. While the current formulation naturally internalizes a topological inductive bias towards compact lesions, expanding the training mixture to encompass filamentous structures (*e.g.*, vessels, neurons) presents a straightforward and promising avenue for future work. Ultimately, Concept-to-Pixel paves a highly interpretable, robust, and generalizable path toward true foundation models in medical image analysis.

Acknowledgements

Supported by Natural Science Foundation of China under Grant 62271465, National Key R&D Program of China under Grant 2025YFC3408300, Natural Science Foundation of Jiangsu Province (No. BK20255001), and Suzhou Basic Research Program under Grant SYG202338.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020)
2. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization. Text Reading, and Beyond **2**(1), 1 (2023)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized medical imaging and graphics **43**, 99–111 (2015)
5. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE transactions on medical imaging **37**(11), 2514–2525 (2018)
6. Bui, N.T., Hoang, D.H., Nguyen, Q.T., Tran, M.T., Le, N.: Meganet: Multi-scale edge-guided attention network for weak boundary polyp segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 7985–7994 (2024)
7. Butoi, V.I., Ortiz, J.J.G., Ma, T., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Universeg: Universal medical image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21438–21451 (2023)
8. Cardona, A., Saalfeld, S., Preibisch, S., Schmid, B., Cheng, A., Pulokas, J., Tomancak, P., Hartenstein, V.: An integrated micro-and macroarchitectural analysis of the drosophila brain by computer-assisted serial section electron microscopy. PLoS biology **8**(10), e1000502 (2010)
9. Chang, S., Zhao, X., Zhang, L., Wang, T.: Unified medical lesion segmentation via self-referring indicator. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10414–10424 (2025)

10. Chartsias, A., Joyce, T., Papanastasiou, G., Semple, S., Williams, M., Newby, D.E., Dharmakumar, R., Tsaftaris, S.A.: Disentangled representation learning in cardiac image analysis. Medical image analysis **58**, 101535 (2019)
11. Cheng, J., Huang, W., Cao, S., Yang, R., Yang, W., Yun, Z., Wang, Z., Feng, Q.: Enhanced performance of brain tumor classification via tumor region augmentation and partition. PloS one **10**(10), e0140381 (2015)
12. Cheng, J., Yang, W., Huang, M., Huang, W., Jiang, J., Zhou, Y., Yang, R., Zhao, J., Feng, Y., Feng, Q., et al.: Retrieval of brain tumors by adaptive spatial pooling and fisher vector representation. PloS one **11**(6), e0157112 (2016)
13. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sam-med2d. arXiv preprint arXiv:2308.16184 (2023)
14. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)
15. Cohen, E., Uhlmann, V.: aura-net: Robust segmentation of phase-contrast microscopy images with few annotations. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 640–644. IEEE (2021)
16. Dong, C., Tang, F., Mao, R., Gao, X., Zhou, S.K.: Lgmsnet: Thinning a medical image segmentation model via dual-level multiscale fusion. arXiv preprint arXiv:2508.15476 (2025)
17. Fan, D.P., Zhou, T., Ji, G.P., Zhou, Y., Chen, G., Fu, H., Shen, J., Shao, L.: Inf-net: Automatic covid-19 lung infection segmentation from ct images. IEEE transactions on medical imaging **39**(8), 2626–2637 (2020)
18. Gillies, R.J., Kinahan, P.E., Hricak, H.: Radiomics: images are more than pictures, they are data. Radiology **278**(2), 563–577 (2016)
19. Gómez-Flores, W., Gregorio-Calas, M.J., Coelho de Albuquerque Pereira, W.: Busbra: A breast ultrasound dataset for assessing computer-aided diagnosis systems. Medical physics **51**(4), 3110–3123 (2024)
20. Gong, H., Chen, G., Wang, R., Xie, X., Mao, M., Yu, Y., Chen, F., Li, G.: Multi-task learning for thyroid nodule segmentation with thyroid region prior. In: 2021 IEEE 18th international symposium on biomedical imaging (ISBI). pp. 257–261. IEEE (2021)
21. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH) **3**(1), 1–23 (2021)
22. Hu, Y., Gao, Y., Gao, W., Luo, W., Yang, Z., Xiong, F., Chen, Z., Lin, Y., Xia, X., Yin, X., et al.: Amd-sd: an optical coherence tomography image dataset for wet amd lesions segmentation. Scientific Data **11**(1), 1014 (2024)
23. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
24. Jaeger, S., Candemir, S., Antani, S., Wáng, Y.X.J., Lu, P.X., Thoma, G.: Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. Quantitative imaging in medicine and surgery **4**(6), 475 (2014)
25. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., De Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: International conference on multimedia modeling. pp. 451–462. Springer (2019)
26. Jia, X., De Brabandere, B., Tuytelaars, T., Gool, L.V.: Dynamic filter networks. Advances in neural information processing systems **29** (2016)

27. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
28. Kundel, H.L., Nodine, C.F., Conant, E.F., Weinstein, S.P.: Holistic component of image perception in mammogram interpretation: gaze-tracking study. Radiology **242**(2), 396–402 (2007)
29. Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., Van Stiphout, R.G., Granton, P., Zegers, C.M., Gillies, R., Boellard, R., Dekker, A., et al.: Radiomics: extracting more information from medical images using advanced feature analysis. European journal of cancer **48**(4), 441–446 (2012)
30. Liu, J., Yang, H., Zhou, H.Y., Xi, Y., Yu, L., Li, C., Liang, Y., Shi, G., Yu, Y., Zhang, S., et al.: Swin-umamba: Mamba-based unet with imagenet-based pretraining. In: International conference on medical image computing and computer-assisted intervention. pp. 615–625. Springer (2024)
31. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
32. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature communications **15**(1), 654 (2024)
33. Ma, W., Yao, Q., Zhang, X., Huang, Z., Jiang, Z., Zhou, S.K.: Towards accurate unified anomaly segmentation. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1342–1352. IEEE (2025)
34. Mendonça, T., Celebi, M., Mendonca, T., Marques, J.: Ph2: A public database for the analysis of dermoscopic images. Dermoscopy image analysis **2** (2015)
35. Nguyen, V.T., Pham, V.T., Tran, T.T.: Ac-mambaseg: An adaptive convolution and mamba-based architecture for enhanced skin lesion segmentation. In: International Conference on Green Technology and Sustainable Development. pp. 13–26. Springer (2024)
36. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
37. Quan, Q., Tang, F., Xu, Z., Zhu, H., Zhou, S.K.: Slide-sam: Medical sam meets sliding window. arXiv preprint arXiv:2311.10121 (2023)
38. Rakic, M., Wong, H.E., Ortiz, J.J.G., Cimini, B.A., Guttag, J.V., Dalca, A.V.: Tyche: Stochastic in-context learning for medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11159–11173 (2024)
39. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
40. Shanmugam, D., Blalock, D., Balakrishnan, G., Guttag, J.: Better aggregation in test-time augmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1214–1223 (2021)
41. Shi, L., Li, X., Hu, W., Chen, H., Chen, J., Fan, Z., Gao, M., Jing, Y., Lu, G., Ma, D., et al.: Ebhi-seg: a novel enteroscope biopsy histopathological hematoxylin and eosin image dataset for image segmentation tasks. Frontiers in Medicine **10**, 1114673 (2023)
42. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. International journal of computer assisted radiology and surgery **9**(2), 283–293 (2014)

43. Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., et al.: Gland segmentation in colon histology images: The glas challenge contest. Medical image analysis **35**, 489–502 (2017)
44. Tahir, A.M., Chowdhury, M.E., Khandakar, A., Rahman, T., Qiblawey, Y., Khurshid, U., Kiranyaz, S., Ibtehaz, N., Rahman, M.S., Al-Maadeed, S., et al.: Covid-19 infection localization and severity grading from chest x-ray images. Computers in biology and medicine **139**, 105002 (2021)
45. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE transactions on medical imaging **35**(2), 630–644 (2015)
46. Tang, F., Dong, C., Ma, W., Xu, Z., Zhu, H., Jiang, Z., Wang, R., Wang, Y., Wu, C., Zhou, S.K.: U-bench: A comprehensive understanding of u-net through 100-variant benchmarking. arXiv preprint arXiv:2510.07041 (2025)
47. Tang, F., Ma, W., He, Z., Tao, X., Jiang, Z., Zhou, S.K.: Pre-trained llm is a semantic-aware and generalizable segmentation booster. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 402–412. Springer (2025)
48. Tang, F., Nian, B., Ding, J., Ma, W., Quan, Q., Dong, C., Yang, J., Liu, W., Zhou, S.K.: Mobile u-vit: Revisiting large kernel and u-shaped vit for efficient medical image segmentation. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 3408–3417 (2025)
49. Tang, F., Nian, B., Li, Y., Jiang, Z., Yang, J., Liu, W., Zhou, S.K.: Mambamim: Pre-training mamba with state space token interpolation and its application to medical image segmentation. Medical Image Analysis **103**, 103606 (2025)
50. Tang, F., Xu, R., Yao, Q., Fu, X., Quan, Q., Zhu, H., Liu, Z., Zhou, S.K.: Hyspark: Hybrid sparse masking for large scale medical image pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 330–340. Springer (2024)
51. Tang, F., Yao, Q., Ma, W., Wu, C., Jiang, Z., Zhou, S.K.: Hi-end-mae: Hierarchical encoder-driven masked autoencoders are stronger vision learners for medical image segmentation. Medical Image Analysis p. 103770 (2025)
52. Tian, Z., Shen, C., Chen, H.: Conditional convolutions for instance segmentation. In: European conference on computer vision. pp. 282–298. Springer (2020)
53. Vandenhende, S., Georgoulis, S., Van Gansbeke, W., Proesmans, M., Dai, D., Van Gool, L.: Multi-task learning for dense prediction tasks: A survey. IEEE transactions on pattern analysis and machine intelligence **44**(7), 3614–3633 (2021)
54. Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdzal, M., Courville, A.: A benchmark for endoluminal scene segmentation of colonoscopy images. Journal of healthcare engineering **2017**(1), 4037190 (2017)
55. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
56. Wang, X., Guo, Y., Xia, J., Zhang, K., Balu, N., Mossa-Basha, M., Shapiro, L., Yuan, C.: Unified and semantically grounded domain adaptation for medical image segmentation. arXiv preprint arXiv:2508.08660 (2025)
57. Weinreb, J.C., Barentsz, J.O., Choyke, P.L., Cornud, F., Haider, M.A., Macura, K.J., Margolis, D., Schnall, M.D., Shtern, F., Tempany, C.M., et al.: Pi-rads prostate imaging–reporting and data system: 2015, version 2. European urology **69**(1), 16–40 (2016)

58. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16133–16142 (2023)
59. Ye, H., Tang, F., Zhao, P., Huang, Z., Zhao, D., Bian, M., Zhou, S.K.: U-rwkv: Lightweight medical image segmentation with direction-adaptive rwkv. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 613–623. Springer (2025)
60. Zhao, X., Pang, Y., Ji, W., Sheng, B., Zuo, J., Zhang, L., Lu, H.: Spider: A unified framework for context-dependent concept segmentation. arXiv preprint arXiv:2405.01002 (2024)
61. Zhou, K., Greenspan, H., Shen, D.: Deep learning for medical image analysis. Academic Press (2017)
62. Zhou, X., Nie, X., Li, Z., Lin, X., Xue, E., Wang, L., Lan, J., Chen, G., Du, M., Tong, T.: H-net: A dual-decoder enhanced fcnn for automated biomedical image diagnosis. Information Sciences **613**, 575–590 (2022)