

Scaling In-Context Segmentation with Hierarchical Supervision

Tidiane Camaret Ndir Marco Reisert Robin T. Schirrmeister
Division of Medical Physics, Department of Radiology,
Medical Center – University of Freiburg
Freiburg, Germany

{tidiane.camaret.ndir, marco.reisert, robin.schirrmeister}@uniklinik-freiburg.de

Abstract

In-context learning (ICL) enables medical image segmentation models to adapt to new anatomical structures from limited examples, reducing the clinical annotation burden. However, standard ICL methods typically rely on dense, global cross-attention, which scales poorly with image resolution. While recent approaches have introduced localized attention mechanisms, they often lack explicit supervision on the selection process, leading to redundant computation in non-informative regions. We propose PatchICL, a hierarchical framework that combines selective image patching with multi-level supervision. Our approach learns to actively identify and attend only to the most informative anatomical regions. Compared to UniverSeg, a strong global-attention baseline, PatchICL achieves competitive in-domain CT segmentation accuracy while reducing compute by 44% at 512×512 resolution. On 35 out-of-domain datasets spanning diverse imaging modalities, PatchICL outperforms the baseline on 6 of 13 modality categories, with particular strength on modalities dominated by localized pathology such as OCT and dermoscopy. Training and evaluation code is available at https://github.com/tidiane-camaret/ic_segmentation.

1. Introduction

In-context learning (ICL) has emerged as a promising paradigm for medical image segmentation, allowing models to generalize to new anatomical structures from a handful of annotated examples without retraining. This capability is especially valuable in clinical settings, where acquiring dense annotations is expensive and time-consuming. However, deploying ICL methods on high-resolution medical volumes remains challenging due to the computational cost of dense cross-attention between target and context images.

Existing ICL approaches for segmentation fall into three broad categories, each with distinct limitations. Global methods such as UniverSeg [1] and Iris [3] employ dense

cross-attention that scales quadratically with spatial resolution. Selective attention methods like Tyche [7] and EICSeg [10] improve focus but learn selection implicitly through segmentation loss alone. Patch-based approaches such as Medverse [4] handle high resolutions via sliding windows but sever global context.

We propose PatchICL, a hierarchical framework that addresses these limitations through explicit patch selection with multi-level supervision. Our method processes images in a coarse-to-fine cascade, learning to identify and attend only to the most informative anatomical regions at each level. Unlike implicit attention methods, our patch selection is directly supervised, allowing early discarding of uninformative background before heavy processing.

Our contributions are as follows. First, we introduce a coarse-to-fine cascade with entropy-guided Gumbel-top- K patch sampling that focuses computation on uncertain regions. Second, we apply multi-level supervision at each resolution stage, providing an explicit training signal for the selection process. Third, we demonstrate that PatchICL achieves competitive in-domain CT accuracy while reducing compute by 44% at 512×512 resolution, and shows favorable scaling on out-of-domain high-resolution datasets.

2. Related Work

Global in-context learning. Early adaptations of ICL for segmentation, such as UniverSeg [1], utilize encoder-decoder architectures with dense cross-attention (Cross-Block) to fuse target and context features. Similarly, Iris [3] encodes context pairs into global task embeddings. These methods require processing the full resolution of both target and context images, resulting in quadratic complexity $\mathcal{O}((HW)^2)$ that prohibits scaling to large inputs.

Selective attention and retrieval. To address scalability, recent works have moved toward selective processing. Tyche [7] introduces stochasticity to explore multiple potential masks, effectively attending to different modes of the output

distribution. EICSeg [10] employs retrieval mechanisms to align features before processing. While these methods focus attention on specific zones or features, selection is often implicit—learned solely through the final segmentation loss without direct supervision on what to select. Consequently, they often still require encoding the entire image volume before discarding irrelevant regions.

Patch-based and efficient approaches. Direct patch-based methods like Medverse [4] use sliding window inference to handle high resolutions, while Patchwork [8] employs a hierarchical stacking strategy to iteratively sample uncertain regions. Pure sliding windows treat patches independently, severing global context.

Positioning. Unlike implicit attention methods, we employ explicit patch selection trained via multi-level supervision. This allows our model to discard uninformative background regions before heavy processing, retaining global context efficiency without the computational cost of full-volume attention.

3. Method

Input: Target image $I^t \in \mathbb{R}^{H \times W}$ and context set $\{(I_i^c, L_i^c)\}_{i=1}^{N_c}$ of image–label mask pairs.

Output: Predicted segmentation mask \hat{L}^t at original resolution.

3.1. Overview

Our method processes images through a coarse-to-fine cascade of M resolution levels $\{r_1, \dots, r_M\}$ (e.g., $24 \rightarrow 96 \rightarrow 128$). At each level: (1) *Sampling*: select K patches via Gumbel-top- K sampling, using uncertainty from the previous level as sampling weights; (2) *Attention*: process target and context patches jointly through transformer layers; (3) *Aggregation*: combine potentially overlapping patch predictions via averaging; and (4) *Refinement*: additively fuse with upsampled previous-level predictions.

3.2. Patch Sampling

At level $\ell > 1$, target patch sampling is guided by the entropy of the previous level’s prediction:

$$w_\ell(x, y) = H(p_{\ell-1}(x, y)),$$

$$H(p) = -p \log p - (1 - p) \log(1 - p), \quad (1)$$

where $p_{\ell-1} = \sigma(\text{logits}_{\ell-1})$, focusing computation on uncertain regions. Patch-level sampling weights are obtained by averaging entropy over each candidate patch region. These weights are passed through a Gumbel-top- K sampler [5]:

$$S_\ell = \text{top-K}(\log \bar{w}_\ell(k) + g_k, K), \quad g_k \sim \text{Gumbel}(0, 1), \quad (2)$$

where $\bar{w}_\ell(k)$ denotes the mean entropy weight for candidate patch k and g_k is an independent Gumbel noise sample. The temperature of the Gumbel distribution is fixed at $\tau = 1$ throughout training.

For context patches, sampling weights are derived from the boundary proximity of the ground-truth mask: we compute a distance transform of the label and assign higher weight to patches whose mean distance to the foreground boundary is small. This encourages context patches to cover the most structurally informative regions.

3.3. Feature Extraction and Encoding

Input images are encoded using the frozen encoder from UniverSeg [1]. A lightweight CNN encoder projects these features to embeddings $\mathbf{z}_k \in \mathbb{R}^d$, producing skip connections for the decoder. To enable a single backbone across all resolution levels, we condition on resolution via sinusoidal encoding, allowing generalization to unseen resolutions.

3.4. Patch-level Attention

Target and context patch embeddings are processed jointly through L transformer layers with: (1) 2D rotary position embeddings (RoPE) encoding spatial patch coordinates (y, x) , enabling reasoning about relative patch positions; (2) per-layer learnable type embeddings distinguishing target from context patches; and (3) bidirectional attention, where all patches attend to all others. A CNN decoder reconstructs patch-level predictions from the attended features.

3.5. Aggregation and Level Combination

Patch predictions are aggregated using simple averaging over overlapping regions:

$$\hat{L}_\ell^t(x, y) = \frac{\sum_k \mathbf{1}_{(x,y) \in P_k} \cdot p_k(x, y)}{\sum_k \mathbf{1}_{(x,y) \in P_k}}, \quad (3)$$

where $\mathbf{1}_{(x,y) \in P_k}$ indicates pixel (x, y) is covered by patch k .

Level predictions are combined via additive fusion within the coverage mask:

$$\hat{L}_{\text{comb}}^t = \hat{L}_{\ell-1, \uparrow}^t + M_\ell \odot \hat{L}_\ell^t, \quad (4)$$

where $\hat{L}_{\ell-1, \uparrow}^t$ is the upsampled previous prediction and M_ℓ is the binary coverage mask indicating which pixels were sampled at level ℓ .

3.6. Training

The model is trained with supervision at each level. Each level loss combines binary cross-entropy and soft Dice:

$$\mathcal{L}_{\text{combined}}^\ell = \mathcal{L}_{\text{BCE}}^\ell + \mathcal{L}_{\text{Dice}}^\ell, \quad (5)$$

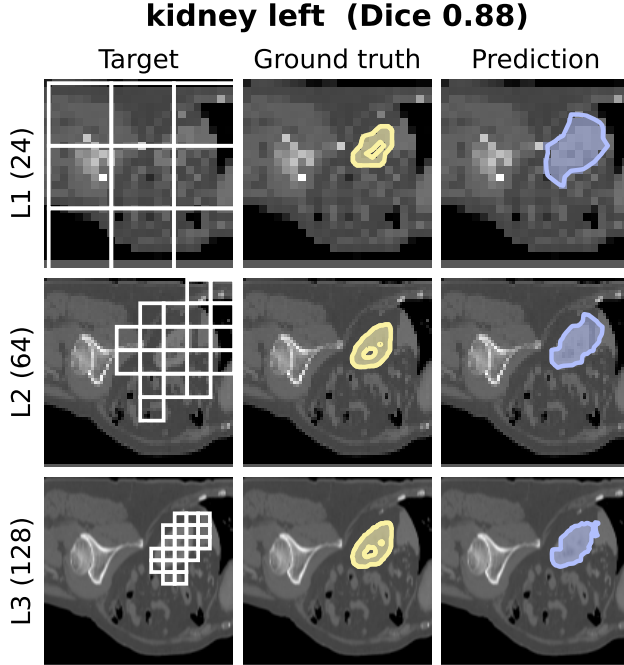


Figure 1. Coarse-to-fine patch selection across resolution levels on subject 0013 of the TotalSegmentator CT dataset. Columns show the target image with patch boxes, ground-truth overlay, and prediction overlay. Rows correspond to cascade levels at increasing resolution.

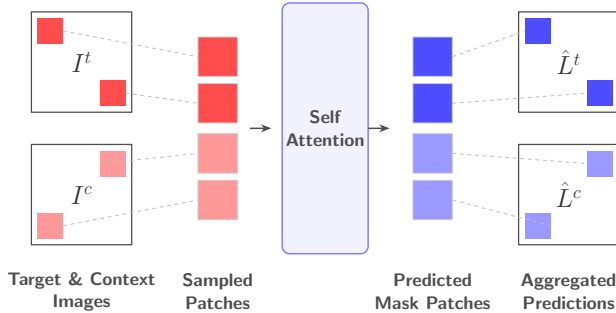


Figure 2. Overview of the PatchICL architecture. Image patches are selectively sampled on the target and context images, and mask predictions are aggregated.

applied to the combined prediction $\hat{L}_{\text{comb},\ell}^t$ against the ground-truth mask at the corresponding resolution. The total loss sums over all levels with uniform weights:

$$\mathcal{L} = \sum_{\ell=1}^M \mathcal{L}_{\text{combined}}^{\ell}. \quad (6)$$

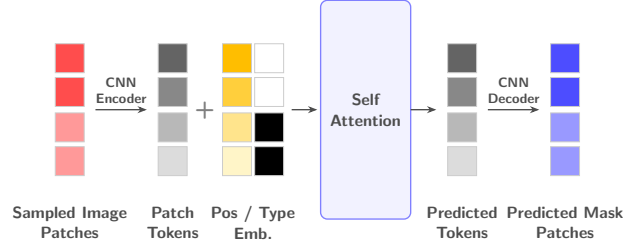


Figure 3. Patch-level attention mechanism. Target and context patches are processed jointly through transformer layers with 2D RoPE, type embeddings, and bidirectional attention.

4. Experiments

4.1. Dataset and Preprocessing

We train PatchICL on TotalSegmentator [9], a large-scale CT dataset with various annotated anatomical structures. To test in-context generalization, we split anatomical classes into 47 structures for training and 45 held-out structures for validation. This ensures the model must generalize to entirely unseen anatomical categories at test time, not just unseen patients. Tab. 1 lists all structures grouped by reporting category.

2D slice extraction. We extract 2D slices from 3D volumes along all three anatomical axes (axial, coronal, sagittal), increasing data diversity and providing multi-view perspectives of each structure. Slices are filtered by coverage: only slices with ≥ 30 foreground pixels are retained, preventing training on uninformative near-empty slices.

Context selection. Context slices are preferentially drawn from the same case as the target, providing anatomically consistent support. When insufficient same-case examples exist, we fall back to other patients with the same structure. We use $N_c = 3$ context pairs during training and validation.

Evaluation protocol. We assess our method across three distinct generalization settings. For *in-domain* performance, we evaluate on TotalSegmentator CT using the 45 held-out anatomical classes on validation cases. To test *cross-modality* transfer, we utilize the TotalSegmentator MRI dataset [2] (all 46 classes, validation split). Finally, we measure *out-of-domain* generalization on MedSegBench [6], challenging the model with 35 diverse datasets spanning modalities such as ultrasound, X-ray, and microscopy.

4.2. Scaling with Input Resolution

A key motivation for PatchICL is efficient scaling to high-resolution inputs. We therefore begin by examining how

Table 1. Anatomical structure categories for TotalSegmentator evaluation. Left/right pairs abbreviated (L/R). Unmarked structures appear in both CT and MRI; ^c = CT only, ^m = MRI only.

Category	Structures
Organs (Abd./Pelvis)	adrenal gland (L/R), colon, duodenum, esophagus, gallbladder, kidney (L/R), kidney cyst ^c (L/R), liver, pancreas, prostate, small bowel, spleen, stomach, urinary bladder
Organs (Thrx./Head/Spine)	brain, heart, spinal cord, atrial appendage ^c (L), lung lobes ^c (5), lung ^m (L/R), thyroid gland ^c , trachea ^c
Bones (Ribs/Sternum)	costal cartilages ^c , ribs ^c L 1–12, ribs ^c R 1–12, sternum ^c
Bones (Limbs/Shldr./Pelvis)	femur (L/R), hip (L/R), humerus (L/R), clavicle ^c (L/R), scapula ^c (L/R), skull ^c , fibula ^m , tibia ^m
Muscles (Trunk)	autochthon (L/R), gluteus max./med./min. (L/R), iliopsoas (L/R)
Muscles (Thigh)	quadriceps femoris ^m (L/R), sartorius ^m (L/R), thigh med./post. comp. ^m (L/R)
Vessels	aorta, iliac artery (L/R), iliac vein (L/R), inferior vena cava, portal/splenic vein, brachiocephalic trunk ^c , brachiocephalic vein ^c (L/R), common carotid art. ^c (L/R), pulmonary vein ^c , subclavian art. ^c (L/R), superior vena cava ^c

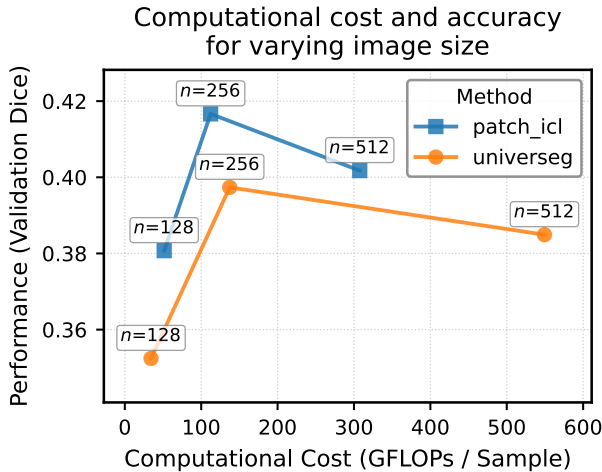


Figure 4. Dice score and computational cost (GFLOPs) as a function of input image resolution on TotalSegmentator CT. PatchICL achieves competitive accuracy with significantly reduced compute at higher resolutions.

accuracy and computational cost evolve with input resolution on the TotalSegmentator CT validation set.

Table 2. Dice scores on TotalSegmentator CT validation cases and held-out classes (mean \pm std). N denotes the number of evaluated slices. Best per category in **bold**.

Category	N	PatchICL	UniverSeg
Bones (Limbs/Shldr./Pelvis)	1,992	0.572 \pm 0.296	0.523 \pm 0.331
Bones (Spine)	780	0.464 \pm 0.271	0.552 \pm 0.281
Organs (Thorax/Head/Spine)	1,077	0.568 \pm 0.328	0.584 \pm 0.316
Organs (Abd./Pelvis)	8,498	0.461 \pm 0.314	0.470 \pm 0.335
Muscles	4,494	0.444 \pm 0.328	0.454 \pm 0.332
Bones (Ribs/Sternum)	13,620	0.390 \pm 0.308	0.321 \pm 0.334
Vessels	8,298	0.340 \pm 0.326	0.348 \pm 0.339
Overall	38,759	0.417 \pm 0.321	0.397 \pm 0.343

Fig. 4 shows that PatchICL achieves higher Dice scores at all tested resolutions while requiring significantly less compute at high resolutions. At 512×512 , PatchICL uses 308 GFLOPs compared to UniverSeg’s 549 GFLOPs—a 44% reduction—while maintaining a Dice advantage (see figure for exact values). Both methods peak at 256×256 resolution, suggesting diminishing returns from higher resolutions on this benchmark. The favorable scaling of PatchICL stems from its patch-based attention, which processes only selected regions rather than the full image. This efficiency advantage motivates the detailed accuracy analysis that follows.

4.3. In-Domain CT Segmentation

Having established PatchICL’s computational advantage, we now examine per-category accuracy on the in-domain TotalSegmentator CT benchmark.

Tab. 2 shows that PatchICL outperforms UniverSeg overall (0.417 vs. 0.397), winning 2 of 7 categories and achieving a higher score on 27 of 45 individual classes (60%). UniverSeg leads on the remaining 5 categories, with its largest advantage on Bones (Spine).

Fig. 6 provides additional examples of the coarse-to-fine patch selection cascade, complementing Fig. 1.

4.4. Cross-Modality Generalization (MRI)

We next evaluate whether the accuracy–efficiency trade-off observed on CT transfers to an unseen imaging modality. We test on TotalSegmentator MRI [2], which contains 46 anatomical structures across abdominal and musculoskeletal regions.

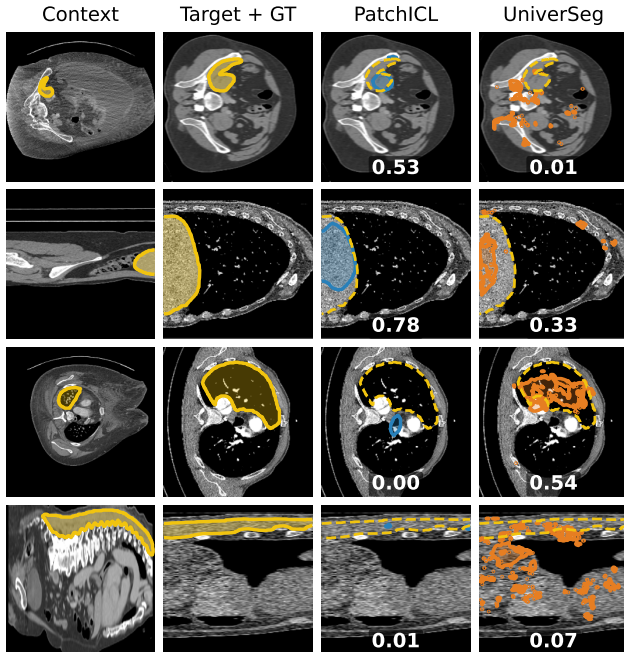


Figure 5. Qualitative comparison on TotalSegmentator CT held-out classes. Columns: context pair, target with **yellow** ground truth, **blue** PatchICL prediction, **orange** UniverSeg prediction. Dashed yellow contours show GT on prediction columns. Dice scores inset. Top rows: PatchICL wins (iliopsoas, liver); bottom rows: UniverSeg wins (lung lobe, autochthon).

Table 3. Dice scores on TotalSegmentator MRI (mean \pm std). N denotes the number of evaluated slices. The overall row averages across all individual classes; category rows are shown for interpretability but contain varying numbers of classes. Best per category in **bold**.

Category	N	PatchICL	UniverSeg
Organs (Thorax/Head/Spine)	936	0.550 \pm 0.291	0.547 \pm 0.294
Muscles (Trunk)	1 396	0.447 \pm 0.304	0.488 \pm 0.305
Organs (Abd./Pelvis)	2 072	0.436 \pm 0.298	0.459 \pm 0.307
Bones (Spine)	441	0.405 \pm 0.234	0.500 \pm 0.236
Bones (Limbs/Pelvis)	520	0.417 \pm 0.340	0.423 \pm 0.350
Vessels	831	0.381 \pm 0.312	0.398 \pm 0.326
Overall	6 196	0.444 \pm 0.304	0.471 \pm 0.310

As shown in Tab. 3, UniverSeg outperforms PatchICL on overall Dice by +2.9%. The largest category-level gaps appear on Bones (Spine) (−8.2%) and Muscles (Trunk)

Table 4. Dice scores on MedSegBench by imaging modality (mean \pm std), using 3 context pairs. N denotes the number of evaluated samples. Best per modality in **bold**.

Modality	N	PatchICL	UniverSeg
X-Ray	11	0.727 \pm 0.045	0.896 \pm 0.025
Dermoscopy	209	0.643 \pm 0.204	0.578 \pm 0.190
Chest X-Ray	2 581	0.610 \pm 0.174	0.684 \pm 0.216
Microscopy	263	0.539 \pm 0.252	0.495 \pm 0.285
Ultrasound	1 209	0.484 \pm 0.241	0.553 \pm 0.261
MRI	147	0.433 \pm 0.270	0.565 \pm 0.211
Nuclei	67	0.344 \pm 0.150	0.356 \pm 0.181
Fundus	26	0.279 \pm 0.212	0.223 \pm 0.185
OCT	101	0.272 \pm 0.223	0.140 \pm 0.188
Nuclear Cell	1 757	0.254 \pm 0.120	0.418 \pm 0.179
Endoscopy	697	0.174 \pm 0.231	0.147 \pm 0.220
CT	272	0.121 \pm 0.178	0.098 \pm 0.165
Pathology	40	0.118 \pm 0.146	0.129 \pm 0.146
Overall	7 380	0.430 \pm 0.260	0.500 \pm 0.286

(−4.1%), consistent with CT findings that UniverSeg excels on complex articulated anatomy. PatchICL maintains near-parity on Organs (Thorax/Head/Spine) and Vessels.

4.5. Out-of-Domain Generalization (MedSeg-Bench)

Finally, we evaluate generalization beyond CT imaging on MedSegBench [6], a comprehensive benchmark spanning 35 datasets across diverse modalities including ultrasound, X-ray, endoscopy, dermoscopy, fundus imaging, and microscopy. Both methods were trained exclusively on TotalSegmentator CT.

Tab. 4 stratifies MedSegBench results by modality category. PatchICL outperforms UniverSeg on 6 of 13 modalities, with the largest gains on OCT (+13.2%) and dermoscopy (+6.6%), where localized features and fine-grained textures dominate. UniverSeg leads on the remaining 7 modalities, with its strongest advantages on X-Ray (−16.8%) and Nuclear Cell (−16.4%), both characterized by high-contrast, globally structured targets. Overall, PatchICL tends to excel on modalities with subtle, spatially localized pathology, while UniverSeg benefits from modalities with well-defined, high-contrast structures.

5. Conclusion

We presented PatchICL, a scalable hierarchical framework for in-context medical image segmentation. By explicitly operating at multiple resolution levels with entropy-guided patch selection, our approach reduces computational overhead while preserving global context. Our evaluations demonstrate that PatchICL achieves competitive performance on in-domain CT and cross-modality MRI tasks,

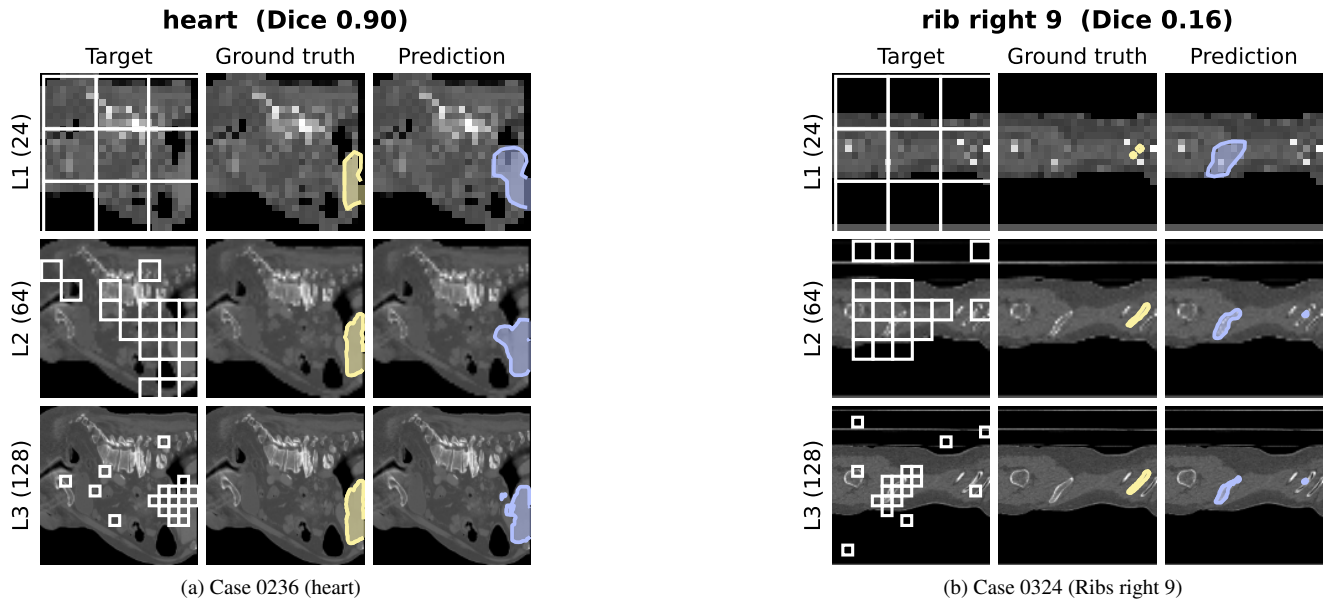


Figure 6. Additional coarse-to-fine patch selection examples on TotalSegmentator CT. Columns show the target image with patch boxes, ground-truth overlay, and prediction overlay. Rows correspond to cascade levels at increasing resolution.

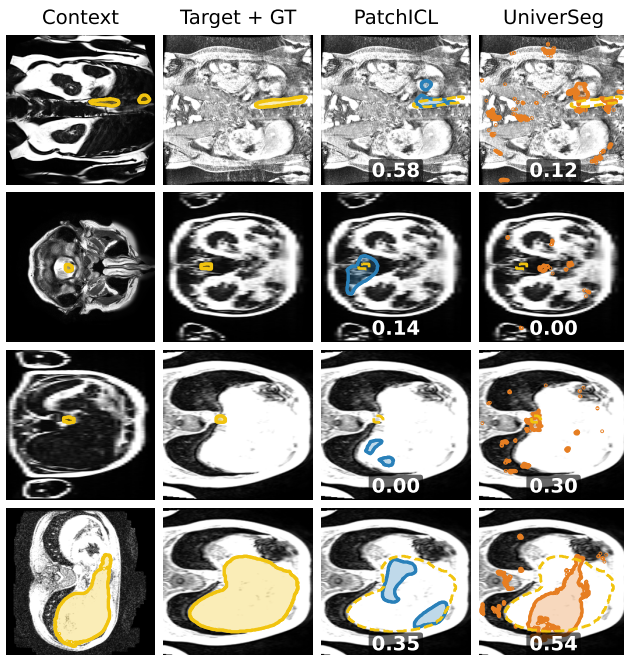


Figure 7. Cross-modality comparison on TotalSegmentator MRI (both methods trained on CT only). Columns: context pair, target with yellow ground truth, blue PatchICL prediction, orange UniverSeg prediction. Dice scores inset.

excelling particularly on repetitive skeletal structures and localized out-of-domain pathologies. While the UniverSeg baseline maintains a slight advantage on complex artic-

ulated anatomies, PatchICL offers a favorable accuracy–compute trade-off, reducing compute by 44% at 512×512 resolution without sacrificing performance. Future work will explore learned patch sampling strategies to better exploit specific medical structures and extend the cascade with adaptive level selection.

References

- [1] Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. Universeg: Universal medical image segmentation. *arXiv preprint arXiv:2304.06131*, 2023. 1, 2
- [2] Tommaso A. D’Antonoli et al. TotalSegmentator MRI: Robust segmentation of anatomic structures in MRI. *arXiv preprint*, 2025. 3, 4
- [3] Yunhe Gao et al. Show and segment: Universal medical image segmentation via in-context learning. *arXiv preprint*, 2025. 1
- [4] others Hu. Medverse: A universal model for medical image segmentation. *arXiv preprint*, 2025. 1, 2
- [5] Wouter Kool, Herke van Hoof, and Max Welling. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement, 2019. 2
- [6] Zeki Kus et al. MedSegBench: A comprehensive benchmark for medical image segmentation. *arXiv preprint*, 2024. 3, 5
- [7] Marianne Rakic et al. Tyche: Stochastic in-context learning for medical image segmentation. *arXiv preprint*, 2025. 1
- [8] Marco Reisert et al. Deep neural patchworks: Coping with large segmentation tasks. In *Medical Image Computing and Computer Assisted Intervention*, 2022. 2
- [9] Jakob Wasserthal et al. TotalSegmentator: Robust segmen-

tation of 104 anatomic structures in CT images. *Radiology: Artificial Intelligence*, 2023. [3](#)

- [10] others Xie. EICSeg: Universal medical image segmentation via efficient in-context learning. *arXiv preprint*, 2026. [1](#), [2](#)