VISION TRANSFORMER-CONDITIONED UNET FOR DOMAIN-ADAPTIVE SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite significant advances in natural domain vision, biomedical semantic segmentation remains challenged by severe label scarcity, low signal-to-noise ratios, and fine anatomical structures. We introduce ViTC-UNet, a framework that bridges this gap by conditioning a UNet pixel decoder on frozen Vision Transformer (ViT) representations via a multi-stage, two-way attention decoder. Semantic identity is decoupled from output geometry using target-specific structure tokens, implemented either as learnable class vectors or embeddings from a language model. This design unifies the global representational capacity of large-scale ViTs with the spatial inductive biases and high-resolution decoding precision of UNets, bypassing computationally intensive and data-hungry Transformer fine-tuning. ViTC-UNet consistently outperforms baselines across CT and MRI modalities in both closed- and open-vocabulary settings, as well as out of distribution, demonstrating that structure-conditioned UNet decoding provides a sample-efficient, highly adaptable pathway for biomedical foundation models. Code will be released upon publication.

1 INTRODUCTION

Semantic segmentation in natural images has largely transitioned from convolutional neural networks (CNNs) (Long et al., 2015; Badrinarayanan et al., 2017) to ViTs (Dosovitskiy et al., 2021). Paired with task-specific decoders ranging from linear heads (Oquab et al., 2024) and lightweight mask decoders (Ranftl et al., 2021; Cheng et al., 2022) to promptable (Kirillov et al., 2023; Ravi et al., 2025) and open-vocabulary interfaces (Liu et al., 2025; Carion et al., 2026), ViTs now form the backbone of state-of-the-art vision models.

Translating this paradigm to biomedical imaging, however, remains exceptionally difficult. Biomedical applications face severe data scarcity (Chi et al., 2020; Payer et al., 2023) and are dominated by targets with complex, sparse, thin, or low-contrast morphology (Wasserthal, 2023; Last et al., 2025). While fine-tuning large ViTs can improve domain adaptation (Matsoukas et al., 2022; 2023; Koch et al., 2024), it is computationally prohibitive and prone to overfitting on limited data. Crucially, the lightweight decoders standard in ViTs lack the localized inductive biases necessary to reconstruct fine anatomical boundaries. Consequently, UNet variants (Ronneberger et al., 2015), most notably nnU-Net (Isensee et al., 2021), remain a central option across biomedical modalities (Lamm et al., 2024; Mehmood et al., 2025). Existing hybrid models attempt to bridge this gap by placing skip connections between transformer and convolutional blocks (Chen et al., 2021; Tang et al., 2022) or directly concatenating latent outputs (Valdivia Ortega et al., 2026), but they fail to fully decouple global feature extraction from high-resolution spatial decoding.

To exploit these complementary strengths, we propose ViTC-UNet (Figures 1, 2), a novel framework that uses a frozen ViT pretrained in natural images as a general-purpose visual prior generator to condition a UNet pixel decoder optimized for complex target topologies. By freezing the ViT backbone, ViTC-UNet eliminates heavy transformer optimization while leveraging the UNet's native sample efficiency and spatial inductive bias, establishing a new state of the art across in-domain biomedical modalities (Figure 3). Furthermore, because our conditioning mechanism explicitly decouples semantic identity from the output layer, the architecture naturally extends to open-vocabulary settings, outperforming existing models in out-of-domain (OOD) testing (Table 1), while proving to be able to add new classes in a close-vocabulary regime (Table 2).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

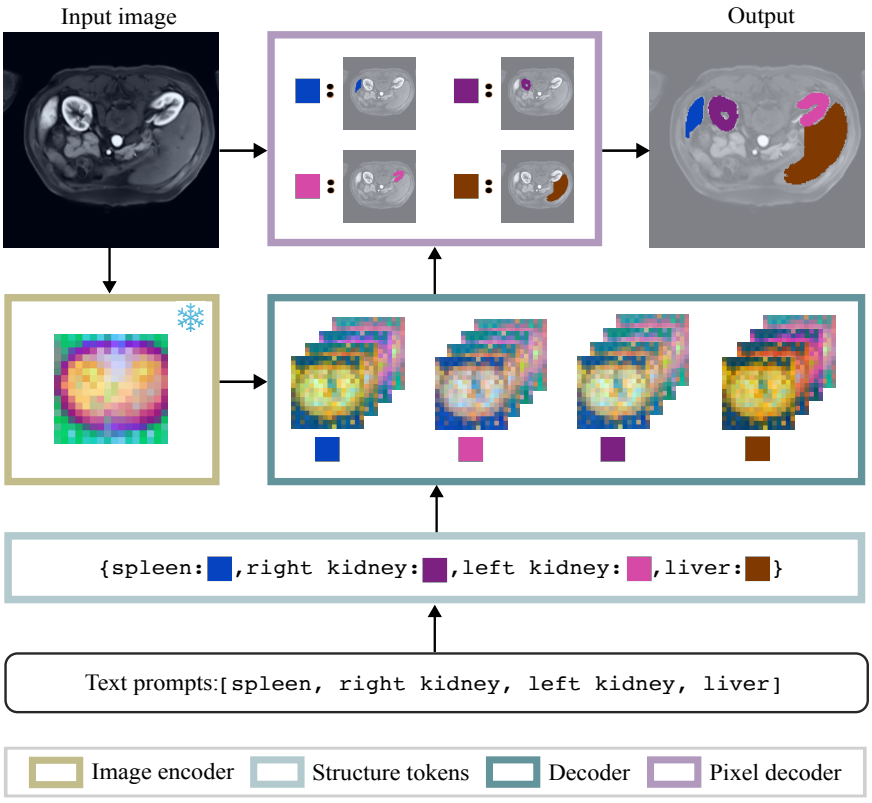


Figure 1: Information flow of ViTC-UNet: a multi-modal framework for semantic segmentation. The input image is processed through a frozen ViT to extract high-level visual priors. These priors are modulated by structure tokens in a two-way attention decoder via latent conditioning. The resulting feature maps, along with the original image, are fed into a UNet acting as a pixel-level decoder. By fusing the global context from the conditioned ViT priors with the local inductive biases of CNNs, the UNet generates precise, structure-specific masks. Notably, the architecture produces individual masks for each input structure through a single-channel output, efficiently leveraging the learned latent conditioning via prompting.

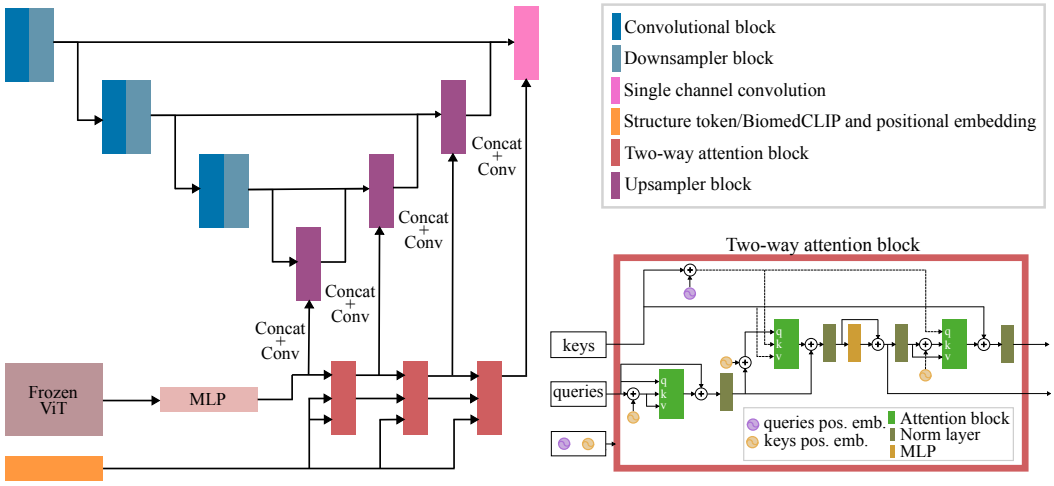


Figure 2: Overview of ViTC-UNet. A ViT generates image embeddings which are integrated with structure tokens through a two-way attention-based decoder modified from SAM (Kirillov et al., 2023). These conditioned features then serve to prompt the nnU-Net architecture as pixel decoder. Structure tokens can be embeddings from BiomedCLIP (Zhang et al., 2025) or learned parameters.

2 Related Work

2.1 Hybrid Convolutional-Attention Architectures

While UNet variants remain the standard for biomedical segmentation in data scarcity modalities (Lamm et al., 2024; Mehmood et al., 2025), recent architectures integrate ViTs to model global context (Chen et al., 2021; Tang et al., 2022). Early hybrids replace the convolutional bottleneck or augment skip connections with attention blocks (Li et al., 2024; Afridi et al., 2026). However, these models demand computationally intensive end-to-end fine-tuning and frequently fail to outperform well-tuned pure-convolutional baselines under rigorous validation (Isensee et al., 2024).

To reduce computational overhead, recent strategies leverage frozen ViT backbones. Li et al. (2025) fuse features from a static Masked Autoencoder into a convolutional decoder, but their approach depends on domain-specific pre-training. Similarly, Wei et al. (2024) deploy a frozen DINOv2 backbone for surgical instrument segmentation, though their evaluation focuses on minimal domain shifts rather than complex modalities like CT or MRI. Most closely related to our work, Valdivia Ortega et al. (2026) concatenate ViT embeddings directly into a UNet bottleneck. Yet, their framework still relies on domain-adapted backbones and enforces rigid architectural constraints linking output dimensionality to label sets. In contrast, ViTC-UNet incorporates frozen natural image priors directly, bypassing domain-specific fine-tuning and strict output limitations.

2.2 Promptable Segmentation and Pixel Decoder Conditioning

ViT architectures like SAM (Kirillov et al., 2023; Ravi et al., 2025; Carion et al., 2026) have advanced interactive segmentation via multi-modal prompting. However, their reliance on lightweight upsampling decoders often necessitates extensive domain-specific fine-tuning to resolve complex biomedical structures (Ma et al., 2024; Archit et al., 2025). Text-conditioned decoders designed for biomedical applications (Zhao et al., 2024; 2025; Hoopes et al., 2025) address this through joint vision-language supervision, but adaptation remains costly and performs poorly on underrepresented target distributions. The continued competitiveness of sample-efficient architectures like nnU-Net (Isensee et al., 2021) underscores that strong spatial inductive biases are essential for reconstructing fine anatomical boundaries.

Despite their spatial precision, standard UNet models (Pal et al., 2025; Xu et al., 2026) rely on fixed output channels mapped directly to specific target classes. This design inherently ties semantic identity to spatial output dimensions, precluding flexible latent conditioning. We resolve this limitation by introducing a conditioned UNet pixel decoder (Figure 2), unifying the open-vocabulary flexibility of promptable attention-based models with the spatial fidelity of convolutional decoders.

3 Model Architecture

The ViTC-UNet architecture bridges global visual representations and localized spatial precision through a modular design comprising a frozen backbone encoder, a learnable conditioning decoder, and a UNet pixel decoder. By leaving the backbone weights intact, ViTC-UNet avoids prohibitive transformer fine-tuning while leveraging strong natural image priors. Instead, target-specific guidance is injected directly into the multi-scale reconstruction path of the UNet, enabling high-capacity, structure-conditioned decoding for complex biomedical targets.

3.1 Backbone Encoder

The encoder utilizes a frozen ViT backbone pre-trained on large-scale natural images to provide strong visual priors. To assess generalization across distinct pre-training paradigms, we evaluate our framework using both self-supervised representations from DINOv2 (Oquab et al., 2024) and prompt-supervised representations from SAM 2 (Ravi et al., 2025). Freezing the backbone eliminates the massive data and compute requirements of end-to-end transformer fine-tuning, preserving rich general-purpose features while preventing overfitting in data-scarce biomedical settings (Chi et al., 2020; Payer et al., 2023).

3.2 CONDITIONING DECODER

To translate natural image embeddings into target-specific biomedical representations, we introduce a conditioning decoder. This module projects target structure tokens into a continuous segmentation space via a linear MLP and processes them alongside image embeddings through a series of modified two-way attention blocks (Kirillov et al., 2023). These blocks enable bidirectional feature modulation by allowing image representations to adapt to the target structure, while the structure token aggregates contextual visual cues. Ablation experiments confirm that two-way cross-attention yields superior conditioning compared to simpler mechanisms like vector addition or one-way attention (Figure 3b).

Unlike standard architectures that bind classes to fixed output channels, ViTC-UNet decouples semantic identity from spatial output geometry using these structure tokens. Depending on the availability of vision-language models for a given modality, structure tokens can be implemented either as frozen text embeddings from BiomedCLIP (Zhang et al., 2025) or as learnable class-specific vectors initialized at training. Spatially replicated structure tokens interact with image embeddings across the attention stack, providing dynamic, queryable guidance for the downstream pixel decoder.

3.3 STRUCTURE-CONDITIONED UNET PIXEL DECODER

To resolve fine anatomical boundaries, we employ nnU-Net (Isensee et al., 2021) as pixel decoder. The output of the n^{th} attention block in the conditioning decoder is concatenated with the skip connection of the n^{th} shallowest convolutional block (Figure 2). This fusion combines localized convolutional inductive biases with global, target-conditioned latent features (Figure 4).

Crucially, the pixel decoder predicts a single-channel binary mask regardless of the number of target classes. Because semantic identity is handled entirely by the input structure token, ViTC-UNet supports open-vocabulary segmentation when using BiomedCLIP embeddings (Table 1), and allows new target structures to be integrated without modifying the pixel decoder when learnable tokens are used to represent structures (Table 2). At inference, pairing an image embedding with individual structure tokens generates target-specific masks, with multi-class predictions obtained by iterating over the set of desired prompts.

3.4 MULTI-STAGE AND MULTI-SCALE LATENT CONDITIONING

Dense prediction tasks benefit significantly from multi-level features rather than relying solely on a final latent vector (Ranftl et al., 2021; Oquab et al., 2024). Instead of extracting intermediate features from the frozen backbone, ViTC-UNet generates a sequence of refined decoder states by passing the final ViT embedding through successive two-way attention blocks. This design is grounded in the topological properties of ViTs. Valdivia Ortega et al. (2026) demonstrate that, when injectivity constraints are satisfied and given a compact domain, ViT encoders act as homeomorphisms, i.e. continuous, bijective mappings with continuous inverses. In DINOv2, injectivity is explicitly enforced via the KoLeo regularizer (Sablayrolles et al., 2019)

$$\mathcal{L}_{\text{KoLeo}}\left(\{x_1, \ldots, x_n\}\right) = -\frac{1}{n}\sum_{i=1}^{n} \log\left(\min_{j\neq i} \|x_i - x_j\|_2\right).$$

This property can be extended to the image encoder of SAM 2 (Corollary 17, Table 3) since, given fixed-resolution natural image domains and binary mask codomains, an optimal segmentation mapping (Definition 12) requires for the model to be capable of grouping prompts using the image's content (Lemma 15), turning the image encoder into an injective function (Theorem 16). The resulting intermediate decoder outputs form a continuous curve through the latent space, preserving topological structures even across domain shifts (Remark 19). Architecturally, these states provide multi-stage conditioning by iteratively refining visual-target interactions, and multi-scale conditioning by feeding into corresponding levels of the UNet decoding path. Deep decoder stages capture target-specific global context, while shallow stages preserve fine spatial details required for narrow or sparse structures. Detailed mathematical proofs can be found in Appendix A.1.

4 Implementation and Training Details

Encoder Configurations. To evaluate architectural flexibility, we benchmark ViTC-UNet using two distinct ViT backbones: self-supervised DINOv2 (Oquab et al., 2024) and prompt-supervised SAM 2 (Ravi et al., 2025). To show that our performance gains do not depend on massive model scales, we utilize the small variant of both encoders. These choices test our framework's adaptability across disparate latent space geometries and pre-training objectives. Across all experiments, the encoder weights remain strictly frozen to preserve pre-trained visual priors and minimize compute requirements.

Decoder Specifications. The two-way attention blocks maintain a fixed structure across trials, with latent dimensions matched to the corresponding backbone. In open-vocabulary settings, structure tokens are formed by concatenating BiomedCLIP (Zhang et al., 2025) image and text embeddings. In closed-vocabulary settings simulating modalities without language models, we represent classes using a dictionary of learnable 1D tokens initialized to the ViT's latent dimension, excluding the background. Additionally, the decoder incorporates learnable spatial positional encodings aligned with DINOv2's embedding scheme.

Ablation Studies. We conduct targeted ablations using learnable structure tokens to isolate the source of performance gains (Figure 3b). First, an *Unconditioned* variant removes structure tokens entirely to ensure gains stem from target-specific conditioning rather than increased decoder capacity. Second, to validate our multi-stage and multi-scale design, we evaluate passing only the final decoder state to either the deepest (*Last deep*) or shallowest (*Last shallow*) UNet layers. Finally, to evaluate the two-way cross-attention mechanism, we compare it against simpler alternatives, namely adding tokens directly to image embeddings applying self-attention only (*Addition*), and testing one-way attention directed either from tokens to image embeddings (*Oneway att. t2e*) or vice versa (*Oneway att. e2t*). Diagrams of the variants can be found in Figure S1.

Pixel Decoder Baselines. We compare ViTC-UNet against two pixel decoder configurations: a linear classification head and the simple hybrid ViT-UNet baseline used by Valdivia Ortega et al. (2026). For both the hybrid baseline and ViTC-UNet, the convolutional backbone adopts a 2D DynUNet architecture from nnU-Net. Unlike the baselines, which allocate a dedicated output channel to the background, ViTC-UNet omits a background channel entirely, relying on token conditioning for foreground selection.

Training and Optimization. All models are trained on a single NVIDIA A100 GPU (64GB) for under 12 hours per job. We optimize using AdamW (Loshchilov & Hutter, 2019) with a batch size of 32 for up to 100 epochs (300 iterations per epoch). Training stops early if validation metrics fail to improve by 1% relative over a 10-epoch plateau. The loss function is a linear combination of Focal loss ($\gamma = 2$) (Lin et al., 2017) and Dice loss (Milletari et al., 2016). Following nnU-Net's protocol, every backbone and decoder combination is evaluated across five independent trials with distinct random initializations using an 80-20 random train/validation split.

Datasets and Benchmark Evaluation. Experiments utilize the curated BiomedParse data compilation (Zhao et al., 2024). For open-vocabulary evaluations, models are trained on a combined dataset of TotalSeg (Wasserthal, 2023) (CT) and AMOS (Ji et al., 2022) (MRI) to maximize class variety, and evaluated OOD across the remaining benchmark datasets (Table 1). For closed-vocabulary evaluations, models are trained and tested individually on each dataset to simulate data-constrained regimes (Figure 3a). Generalization is measured on the independent holdout set from the Interactive 3D Biomedical Image Segmentation benchmark, with training and validation folds randomized independently across all five runs. Datasets sources and nnU-Net's performance on them can be found in Table S1.

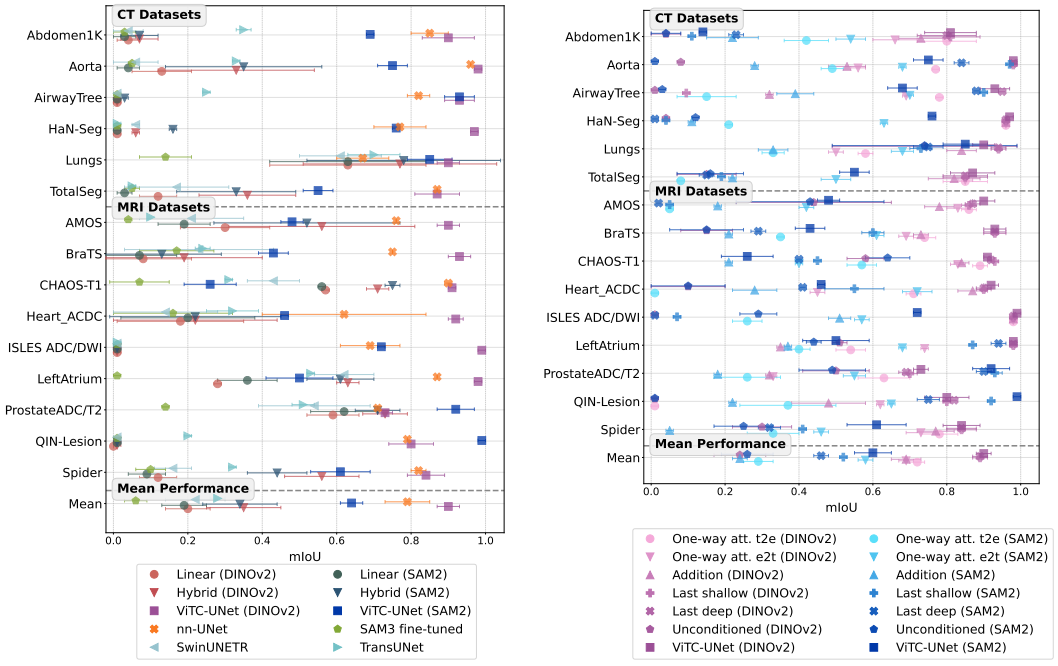
Class Sampling Strategy and Label Space Expansion. In closed-vocabulary training, optimal performance requires optimizing all target structures simultaneously within each step. Each input volume is processed alongside the full set of dataset class tokens, generating masks for all target structures regardless of their presence in the image. This exhaustive sampling mirrors standard UNet optimization, exposing the conditioning mechanism to balanced positive and negative signals across

target structures. Further, to test incremental learning capability, we execute a two-stage training experiment where models are initially trained on half of the available classes before expanding to the full label set. As shown in Table 2, ViTC-UNet integrates novel classes without degrading performance on previously learned structures, confirming that token conditioning supports seamless label-space expansion.

5 Evaluation and Discussion

We evaluate segmentation performance using mean Intersection over Union (mIoU), excluding the background class to focus on discriminative accuracy. Volumetric scans are processed slice-wise following the backbones' geometry, with 2D predictions reassembled into 3D volumes for final metric computation.

Pixel Decoder Baselines. The performance gain of the ViT-UNet hybrid over the linear baseline confirms that dense biomedical segmentation benefits significantly from convolutional inductive biases. However, Table S2 demonstrates that bottleneck-level fusion into a fixed-channel UNet is insufficient for ViT domain adaptation; despite leveraging skip connections, the lack of multi-stage conditioning and cross-domain adaptation causes the hybrid baseline to misclassify target structures. In contrast, ViTC-UNet using a DINOv2 backbone correctly resolves fine anatomical morphologies (Figure 4). This confirms that frozen ViT decoding requires target-specific guidance injected continuously across the multi-scale UNet path, mitigating the loss of spatial inductive bias common to pure transformer architectures in data-scarce settings (Soh & Rajapakse, 2023; He et al., 2024; Sang et al., 2026). Furthermore, ViTC-UNet outperforming nnU-Net shows that convolutional decoders are substantially boosted when paired with general-purpose ViT representations.



(a) Comparison of models fine-tuned to the dataset. Values can be found in Tables S2 and S3.

(b) Ablation study on decoder configurations (Figure S1). Values can be found in Tables S4 and S5.

Figure 3: Performance comparison of ViTC-UNet using frozen DINOv2 and SAM2 as image encoders and learnable parameters as structure tokens. We report mIoU as mean±std of 5-fold independent runs for all of the presented results except for nnU-Net, whose performance was extracted from literature (Table S1). Quantitative results can be found in Appendix A.2.

Comparison with Fine-Tuned and Closed-Vocabulary Baselines. As shown in Table S3, regardless of the used backbone, ViTC-UNet with learnable structure tokens outperforms fully fine-tuned

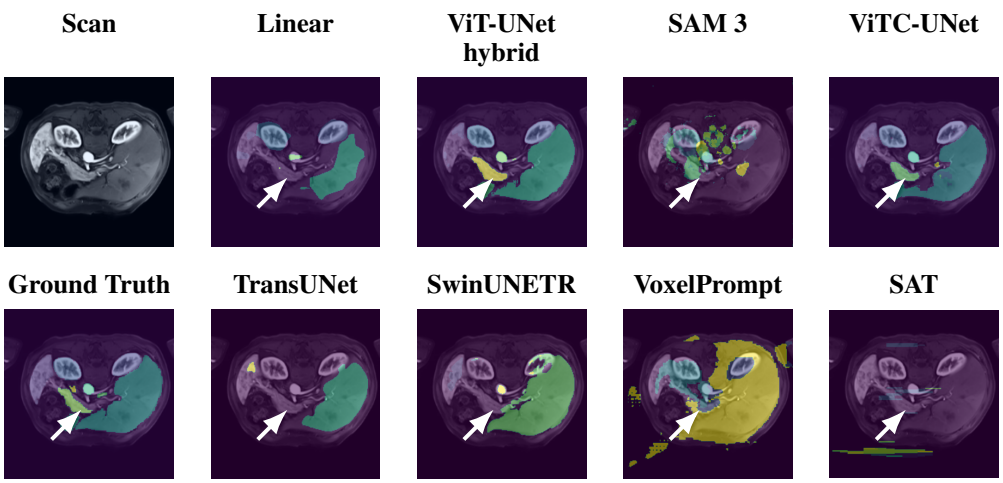


Figure 4: Qualitative comparison of segmentation on the AMOS (Ji et al., 2022) MRI dataset. ViTC-UNet using DINOv2 as encoder demonstrates superior boundary resolution and structure classification compared to promptable, agentic and hybrid baselines as well as alternative heads. Further examples can be found in the Appendix A.3.

hybrid baselines, including TransUNet (Chen et al., 2021) and Swin UNETR (Tang et al., 2022). It also surpasses SAM 3 under equal conditions with frozen image and text encoders. In closed-vocabulary benchmarks (Figure 3a), DINOv2-based ViTC-UNet achieves the top overall performance, reaching an average foreground mIoU of 0.90 across all datasets under equal weighting, outperforming pure UNets, hybrid models, and promptable baselines alike.

Label Space Flexibility. ViTC-UNet supports incremental learning. In a two-stage experiment representing structures with learnable tokens where models were initially trained on half the class labels before fine-tuning on the whole class set, ViTC-UNet integrated new target classes without degrading performance on previously learned structures, while reducing fine-tuning time by 36% (Table 2). This flexibility is a direct benefit of our single-channel, token-conditioned formulation.

Open-Vocabulary Out-of-Domain Generalization. When paired with BiomedCLIP (Zhang et al., 2025) embeddings for open-vocabulary segmentation, ViTC-UNet demonstrates strong cross-domain robustness (Table 1). Compared against zero-shot SAT (Pro version) (Zhao et al., 2025) and the agentic baseline VoxelPrompt (Hoopes et al., 2025) (trained on TotalSeg (Wasserthal, 2023) and AMOS (Ji et al., 2022) as our proposed model), ViTC-UNet consistently achieves superior mIoU on unseen OOD datasets, confirming the strong generalization enabled by decoupled semantic conditioning and preservation of local inductive bias.

Ablation of Decoder Mechanisms. Ablation experiments confirm the necessity of each architectural component (Figure 3b, Tables S4, S5). Removing structure tokens (*Unconditioned*) leads to severe overfitting, confirming that performance gains stem from dynamic target conditioning rather than added decoder parameters. Restricting latent injection to only the deepest (*Last deep*) or shallowest (*Last shallow*) layers degraded segmentation metrics, validating our multi-stage, multi-scale design (Section 4). Moreover, simple feature addition and self-attention (*Addition*) failed to achieve consistent precision across modalities. Finally, while one-way attention variants (*Oneway att. t2e*, *Oneway att. e2t*) performed competitively on DINOv2, their accuracy dropped sharply on SAM 2, proving that two-way cross-attention is essential for cross-backbone stability.

Latent Space Geometry and Orthogonal Separability. To explain why DINOv2 consistently outperforms SAM 2 within our framework (Figure 3a), we analyzed their respective latent space

Table 1: Zero-shot OOD segmentation performance (mIoU, mean ± std over 5 runs). SAT (Zhao et al., 2025) is evaluated out-of-the-box, whereas VoxelPrompt (Hoopes et al., 2025) and ViTC-UNet using frozen BiomedCLIP (Zhang et al., 2025) concatenated image and text embeddings as structure tokens with a frozen DINOv2 backbone are trained on TotalSeg (Wasserthal, 2023) and AMOS (Ji et al., 2022) before zero-shot evaluation on unseen target datasets. Bold values indicate best performance while † shows statistically significant improvements over all other models ($p<0.05$, Mann-Whitney U test (Virtanen et al., 2020)).

(a) Zero-shot OOD mIoU performance on CT datasets.

CT dataset	SAT	VoxelPrompt	ViTC-UNet
Abdomen1K	0.28±0.04	0.10±0.02	**0.88±0.03**†
Aorta	0.50±0.01	0.14±0.17	**0.94±0.01**†
AirwayTree	0.60±0.20	0.01±0.01	**0.94±0.01**†
HaN-Seg	0.02±0.01	0.01±0.01	**0.94±0.01**†
Lungs	0.51±0.01	0.07±0.06	**0.90±0.01**†

(b) Zero-shot OOD mIoU performance on MRI datasets.

MRI dataset	SAT	VoxelPrompt	ViTC-UNet
BraTS	0.32±0.05	0.04±0.03	**0.92±0.01**†
CHAOS-T1	0.20±0.01	0.15±0.05	**0.92±0.01**†
Heart_ACDC	0.44±0.09	0.03±0.03	**0.93±0.01**†
ISLES ADC/DWI	0.25±0.09	0.17±0.10	**0.93±0.01**†
LeftAtrium	0.17±0.01	0.06±0.06	**0.94±0.01**†
ProstateADC/T2	0.06±0.02	0.01±0.01	**0.93±0.01**†
QIN-Lesion	0.06±0.02	0.01±0.01	**0.94±0.01**†
Spider	0.05±0.08	0.01±0.01	**0.92±0.02**†

Table 2: Impact of incremental class acquisition on ViTC-UNet performance. We compare a standard joint training baseline (all classes from initialization, top) against a two-phase training where the model is first trained on half subset of classes before incorporating the remainder (bottom). Results are reported as mean mIoU ± standard deviation over five independent runs.

	CT		MRI				
Two-stage Training	HaN-Seg	TotalSeg	AMOS	CHAOS-T1	ISLES ADC/DWI	QIN-Lesion	Spider
No	0.97±0.01	0.87±0.06	0.90±0.03	0.91±0.01	0.99±0.01	0.80±0.06	0.84±0.05
Yes	0.94±0.01	0.93±0.05	0.87±0.02	0.86±0.03	0.99±0.01	0.79±0.03	0.84±0.05

Table 3: Quantitative feature space analysis for ViTC-UNet comparing DINOv2 and SAM 2 backbones. Cosine Similarity measures the maximum similarity between non-empty slice embeddings across the dataset. Encoder and Decoder columns report the mean ± std angle between structure anchors (patch tokens belonging to distinct target organs within a slice; Figure S2). mIoU reflects overall segmentation performance of ViTC-UNet using learnable structure tokens. Pearson correlation (Virtanen et al., 2020) is calculated between inter-organ anchor angles and mIoU metrics.

	DINOv2				SAM 2			
Dataset	Cosine similarity	Encoder output	Decoder output	mIoU	Cosine similarity	Encoder output	Decoder output	mIoU
	CT							
Abdomen1K	0.57±0.42	52±18	48±5	0.90±0.07	0.05±0.05	36±11	1±1	0.69±0.01
Aorta	0.16±0.07	47±9	163±1	0.98±0.01	0.01±0.01	1±1	1±1	0.75±0.04
Lungs	0.36±0.19	50±11	98±1	0.90±0.03	0.11±0.09	32±10	1114±1	0.85±0.06
TotalSeg	0.01±0.01	45±8	60±6	0.87±0.06	0.1±0.01	59±6	23±7	0.55±0.04
	MRI							
AMOS	0.01±0.22	58±6	74±4	0.90±0.03	0.17±0.16	64±7	36±17	0.48±0.03
BraTS	0.02±0.01	45±10	71±4	0.93±0.03	0.10±0.07	44±13	27±1	0.43±0.04
ISLES ADC/DWI	0.08±0.07	19±7	144±28	0.99±0.01	0.10±0.04	51±9	1±1	0.72±0.01
LeftAtrium	0.01±0.01	4±2	147±1	0.98±0.01	0.01±0.01	8±1	139±1	0.50±0.09
	CT and MRI							
Mean	0.13±0.11	40±9	100±6	0.90±0.03	0.12±0.07	36±7	42±3	0.64±0.04
Pearson correlation	-	0.25	0.32	-	-	0.26	0.37	-

geometries. Low inter-slice cosine similarities (Table 3) indicate that both backbones embed data injectively, providing empirical evidence to our theoretical results (Section 3.4) and ruling out a lack of spatial mapping capacity as the failure reason.

Instead, performance differences stem from angular separability in latent space. We computed *structure anchors* as the mean latent vector of patch tokens corresponding to each target class within a slice (Figure S2), both before and after processing by the conditioning decoder. As reported in Table 3, the angular separation between DINOv2 structure anchors increases significantly after passing through the two-way attention decoder, whereas SAM 2 anchors exhibit minimal angular spread. A positive Pearson correlation between structure anchor angles and mIoU indicates that DINOv2's latent space possesses higher orthogonal separability, allowing the pixel decoder to disentangle overlapping anatomical features and passing them to the pixel decoder more effectively.

6 Conclusion

We have demonstrated that state-of-the-art biomedical semantic segmentation can be achieved by conditioning a UNet pixel decoder on frozen ViT representations via a multi-stage, multi-scale mechanism. By fusing the global representational capacity of pre-trained ViTs with the spatial inductive biases of convolutional decoders, ViTC-UNet addresses core challenges of biomedical imaging such as complex target morphology, low signal-to-noise ratios, and data scarcity.

Beyond empirical gains, this work provides both theoretical and practical insights into pixel decoder design, representation learning and latent space geometry. By decoupling semantic identity from the spatial output layer, ViTC-UNet performs multi-class, open-vocabulary and OOD segmentation within a single-channel formulation. This flexibility enables seamless label-space expansion via new learnable structure tokens or vision-language embeddings without requiring architectural modifications or head retraining. Furthermore, achieving these results using frozen backbones offers a computationally light and sample-efficient alternative to full transformer fine-tuning. Ultimately, our framework demonstrates that frozen natural image priors can be effectively decoded to bridge the gap between large foundation models and specialized biomedical applications.

7 Limitations

Despite its strengths, several limitations present opportunities for future research. First, because the used ViT backbones were pre-trained on 2D natural images, our pipeline operates on 2D slices before reassembling 3D volumes. Further work is needed to find the optimal way of transferring the 2D general representations to the 3D pixel decoder. Second, while we establish a link between latent orthogonal separability and segmentation accuracy, the downstream impact of other geometric and topological properties in ViT feature spaces remains to be fully explored. Finally, while ViTC-UNet currently conditions on semantic text and class tokens; extending the architecture to support interactive spatial prompts (e.g., points or bounding boxes) would further establish the UNet as a high-fidelity alternative to lightweight mask decoders.

8 Reproducibility Statement

To ensure complete reproducibility, all source code, weights, and configuration files will be publicly released upon acceptance. All reported metrics reflect the mean across five independent experimental runs with distinct random seeds, unless explicitly noted otherwise, guaranteeing this way statistical stability. Detailed training and optimization protocols are provided in Section 4, full mathematical proofs are included in Appendix A.1 and licenses from used models are found in Appendix A.4.

9 AI use disclosure

In this work, we used generative AI tools for improving readability, searching information and identify relevant literature. We have not used generative AI tools for any part of the implementation or evaluation. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Sikandar Afridi, Atif Jan, Muhammad Abeer Irfan, Muhammad Irfan Khattak, and Taimur Ahmed Khan. 3d-vit-unet: 3d vision transformer based unet-like model for volumetric brain tumor segmentation. *PLOS Digital Health*, 5(3):1–27, 03 2026.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Bram van Ginneken, Michel Bilello, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc J. Gollub, Stephan H. Heckers, Henkjan Huisman, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Jennifer S. Golia Pernicka, Kawal Rhode, Catalina Tobon-Gomez, Eugene Vorontsov, James A. Meakin, Sebastien Ourselin, Manuel Wiesenfarth, Pablo Arbeláez, Byeonguk Bae, Sihong Chen, Laura Daza, Jianjiang Feng, Baochun He, Fabian Isensee, Yuanfeng Ji, Fucang Jia, Ildoo Kim, Klaus Maier-Hein, Dorit Merhof, Akshay Pai, Beomhee Park, Mathias Perslev, Ramin Rezaiifar, Oliver Rippel, Ignacio Sarasua, Wei Shen, Jaemin Son, Christian Wachinger, Liansheng Wang, Yan Wang, Yingda Xia, Daguang Xu, Zhanwei Xu, Yefeng Zheng, Amber L. Simpson, Lena Maier-Hein, and M. Jorge Cardoso. The medical segmentation decathlon. *Nature Communications*, 13(1):4128, 2022.

Anwai Archit, Luca Freckmann, Sushmita Nair, Nabeel Khalid, Paul Hilt, Vikas Rajashekar, Marei Freitag, Carolin Teuber, Melanie Spitzner, Constanza Tapia Contreras, Genevieve Buckley, Sebastian von Haaren, Sagnik Gupta, Marian Grade, Matthias Wirth, Günter Schneider, Andreas Dengel, Sheraz Ahmed, and Constantin Pape. Segment anything for microscopy. *Nature Methods*, 22(3): 579–591, 2025.

Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

Ujjwal Baid, Satyam Ghodasara, Michel Bilello, Suyash Mohan, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe Kitamura, Sarthak Pati, Luciano Prevedello, Jeffrey Rudie, Chiharu Sako, Russell Shinohara, Tim Bergquist, Rong Chai, James Eddy, Julia Elliott, Walter Reade, and Spyridon Bakas. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification, 07 2021.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S. Kirby, John B. Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific Data*, 4 (1):170117, 2017.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2026. URL https://arxiv.org/abs/2511.16719.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation, 2022. URL https://arxiv.org/abs/2112.01527.

Weicheng Chi, Lin Ma, Junjie Wu, Mingli Chen, Weiguo Lu, and Xuejun Gu. Deep learning-based medical image segmentation with limited labels. *Phys Med Biol*, 65(23), Nov 2020.

Ezequiel de la Rosa, Mauricio Reyes, Sook-Lei Liew, Alexandre Hutton, Roland Wiest, Johannes Kaesmacher, Uta Hanning, Arsany Hakim, Richard Zubal, Waldo Valenzuela, David Robben,

Diana M. Sima, Vincenzo Anania, Arne Brys, James A. Meakin, Anne Mickan, Gabriel Broocks, Christian Heitkamp, Shengbo Gao, Kongming Liang, Ziji Zhang, Md Mahfuzur Rahman Siddiquee, Andriy Myronenko, Pooya Ashtari, Sabine Van Huffel, Hyunsu Jeong, Chiho Yoon, Chulhong Kim, Jiayu Huo, Sebastien Ourselin, Rachel Sparks, Albert Clèrigues, Arnau Oliver, Xavier Lladó, Liam Chalcroft, Ioannis Pappas, Jeroen Bertels, Ewout Heylen, Juliette Moreau, Nima Hatami, Carole Frindel, Abdul Qayyum, Moona Mazher, Domenec Puig, Shao-Chieh Lin, Chun-Jung Juan, Tianxi Hu, Lyndon Boone, Maged Goubran, Yi-Jui Liu, Susanne Wegener, Florian Kofler, Ivan Ezhov, Suprosanna Shit, Moritz R. Hernandez Petzsche, Michael Müller, Bjoern Menze, Jan S. Kirschke, and Benedikt Wiestler. Deepisles: a clinically validated ischemic stroke segmentation model from the isles'22 challenge. *Nature Communications*, 16(1):7357, 2025a.

Ezequiel de la Rosa, Mauricio Reyes, Sook-Lei Liew, Alexandre Hutton, Roland Wiest, Johannes Kaesmacher, Uta Hanning, Arsany Hakim, Richard Zubal, Waldo Valenzuela, David Robben, Diana M. Sima, Vincenzo Anania, Arne Brys, James A. Meakin, Anne Mickan, Gabriel Broocks, Christian Heitkamp, Shengbo Gao, Kongming Liang, Ziji Zhang, Md Mahfuzur Rahman Siddiquee, Andriy Myronenko, Pooya Ashtari, Sabine Van Huffel, Hyunsu Jeong, Chiho Yoon, Chulhong Kim, Jiayu Huo, Sebastien Ourselin, Rachel Sparks, Albert Clèrigues, Arnau Oliver, Xavier Lladó, Liam Chalcroft, Ioannis Pappas, Jeroen Bertels, Ewout Heylen, Juliette Moreau, Nima Hatami, Carole Frindel, Abdul Qayyum, Moona Mazher, Domenec Puig, Shao-Chieh Lin, Chun-Jung Juan, Tianxi Hu, Lyndon Boone, Maged Goubran, Yi-Jui Liu, Susanne Wegener, Florian Kofler, Ivan Ezhov, Suprosanna Shit, Moritz R. Hernandez Petzsche, Michael Müller, Bjoern Menze, Jan S. Kirschke, and Benedikt Wiestler. Deepisles: a clinically validated ischemic stroke segmentation model from the isles'22 challenge. *Nature Communications*, 16(1):7357, 2025b.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Andriy Fedorov, Clare Tempany, and Fiona Fennessy. Data from QIN-PROSTATE-Repeatability, 2019.

Jingliu He, Yuqi Ma, Mingyue Yang, Wensong Yang, Chunming Wu, and Shanxiong Chen. Tac-unet: transformer-assisted convolutional neural network for medical image segmentation. *Quant Imaging Med Surg*, 14(12):8824–8839, Dec 2024.

Andrew Hoopes, Neel Dey, Victor Ion Butoi, John V. Guttag, and Adrian V. Dalca. Voxelprompt: A vision agent for end-to-end medical image analysis, 2025. URL https://arxiv.org/abs/2410.08397.

Fabian Isensee, Paul F. Jaeger, Peter M. Full, Philipp Vollmuth, and Klaus H. Maier-Hein. nnu-net for brain tumor segmentation, 2020. URL https://arxiv.org/abs/2011.00848.

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

Fabian Isensee, Constantin Ulrich, Tassilo Wald, and Klaus H. Maier-Hein. Extending nnu-net is all you need, 2022. URL https://arxiv.org/abs/2208.10791.

Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F. Jäger. nnu-net revisited: A call fornbsp;rigorous validation innbsp;3d medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024: 27th International Conference, Marrakesh, Morocco, October 6–10, 2024, Proceedings, Part IX*, pp. 488–498, Berlin, Heidelberg, 2024. Springer-Verlag.

Yuanfeng Ji, Haotian Bai, Chongjian GE, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, and Ping Luo. AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=Vk4-HUnkEak.

A. Emre Kavur, N. Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, Bora Baydar, Dmitry Lachinov, Shuo Han, Josef Pauli, Fabian Isensee, Matthias Perkonigg, Rachana Sathish, Ronnie Rajan, Debdoot Sheet, Gurbandurdy Dovletov, Oliver Speck, Andreas Nürnberger, Klaus H. Maier-Hein, Gözde Bozdağı Akar, Gözde Ünal, Oğuz Dicle, and M. Alper Selver. Chaos challenge - combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, April 2021. ISSN 1361-8415. doi: 10.1016/j.media.2020.101950. URL http://dx.doi.org/10.1016/j.media.2020.101950.

Ali Emre Kavur, M. Alper Selver, Oğuz Dicle, Mustafa Barış, and N. Sinem Gezer. CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data, April 2019. URL https://doi.org/10.5281/zenodo.3362844.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. URL https://arxiv.org/abs/2304.02643.

Valentin Koch, Sophia J. Wagner, Salome Kazeminia, Ece Sancar, Matthias Hehr, Julia A. Schnabel, Tingying Peng, and Carsten Marr. DinoBloom: A Foundation Model for Generalizable Cell Embeddings in Hematology . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15012. Springer Nature Switzerland, October 2024.

Lorenz Lamm, Simon Zufferey, Ricardo D. Righetto, Wojciech Wietrzynski, Kevin A. Yamauchi, Alister Burt, Ye Liu, Hanyi Zhang, Antonio Martinez-Sanchez, Sebastian Ziegler, Fabian Isensee, Julia A. Schnabel, Benjamin D. Engel, and Tingying Peng. Membrain v2: an end-to-end tool for the analysis of membranes in cryo-electron tomography. *bioRxiv*, 2024.

Mart G. F. Last, Lenard M. Voortman, and Thomas H. Sharp. Scaling data analyses in cellular cryoet using comprehensive segmentation. *bioRxiv*, 2025.

Feng Li, Lianzhong Sun, Kwok-Yan Lam, Songbo Zhang, Zhongming Sun, Bao Peng, Hongzeng Xu, and Libo Zhang. Segmentation of human aorta using 3d nnu-net-oriented deep learning. *Rev Sci Instrum*, 93(11):114103, Nov 2022.

Lei Li, Fuping Wu, Sihan Wang, Xinzhe Luo, Carlos Martín-Isla, Shuwei Zhai, Jianpeng Zhang, Yanfei Liu, Zhen Zhang, Markus Ankenbrand, Haochuan Jiang, Xiaoran Zhang, Linhong Wang, Tewodros Arega, Elif Altunok, Zhou Zhao, Feiyan Li, Jun Ma, Xiaoping Yang, and Xiahai Zhuang. Myops: A benchmark of myocardial pathology segmentation combining three-sequence cardiac magnetic resonance images. *Medical Image Analysis*, 87:102808, 04 2023.

Mengyu Li, Guoyao Shen, Chad W. Farris, and Xin Zhang. Few-shot deployment of pretrained MRI transformers in brain imaging tasks. aug 2025.

Xin Li, Wenhui Zhu, Xuanzhao Dong, Oana M. Dumitrascu, and Yalin Wang. Evit-unet: U-net like efficient vision transformer for medical image segmentation on mobile and edge devices, 2024. URL https://arxiv.org/abs/2410.15036.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Yang Liu, Yufei Yin, Chenchen Jing, Muzhi Zhu, Hao Chen, Yuling Xi, Bo Feng, Hao Wang, Shiyu Li, and Chunhua Shen. Unified open-world segmentation with multi-modal prompts, 2025. URL https://arxiv.org/abs/2510.10524.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Jun Ma, Yixin Wang, Xingle An, Cheng Ge, Ziqi Yu, Jianan Chen, Qiongjie Zhu, Guoqiang Dong, Jian He, Zhiqiang He, Tianjia Cao, Yuntao Zhu, Ziwei Nie, and Xiaoping Yang. Toward data-efficient learning: A benchmark for covid-19 ct lung and infection segmentation. *Med Phys*, 48(3): 1197–1210, Mar 2021.

Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, Shucheng Cao, Qi Zhang, Shangqing Liu, Yunpeng Wang, Yuhui Li, Jian He, and Xiaoping Yang. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Trans Pattern Anal Mach Intell*, 44(10):6695–6714, Oct 2022.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

Christos Matsoukas, Johan Fredin Haslum, Moein Sorkhei, Magnus Söderberg, and Kevin Smith. What makes transfer learning work for medical images: Feature reuse other factors, 2022. URL https://arxiv.org/abs/2203.01825.

Christos Matsoukas, Johan Haslum, Magnus Soderberg, and Kevin Smith. Pretrained vits yield versatile representations for medical images, 03 2023.

Mehwish Mehmood, Shahzaib Iqbal, Tariq Khan, Ivor Spence, and Muhammad Fahim. Lvs-net: A lightweight vessels segmentation network for retinal image analysis, 01 2025.

Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Çağatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, José António Mariz, Raphael Meier, Sérgio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015.

Yves Meyer. *Oscillating Patterns in Image Processing and Nonlinear Evolution Equations: The Fifteenth Dean Jacqueline B. Lewis Memorial Lectures*, volume 22 of *University Lecture Series*. American Mathematical Society, United States, 2001. ISBN 0821829203.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation, 2016. URL https://arxiv.org/abs/1606.04797.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Debojyoti Pal, Kaushik Dutta, Daniel R. Ludwig, and Kooresh Shoghi. PanNet: a feature-based attention aggregation model for segmenting pancreatic ductal adenocarcinoma on contrast-enhanced CT images of the abdomen. In Susan M. Astley and Axel Wismüller (eds.), *Medical Imaging 2025: Computer-Aided Diagnosis*, volume 13407, pp. 134071R. International Society for Optics and Photonics, SPIE, 2025.

Tristan Payer, Faraz Nizamani, Meinrad Beer, Michael Götz, and Timo Ropinski. Medical volume segmentation by overfitting sparsely annotated data. *J Med Imaging (Bellingham)*, 10(4):044007, Jul 2023.

Gašper Podobnik, Primož Strojan, Primož Peterlin, Bulat Ibragimov, and Tomaž Vrtovec. Han-seg: The head and neck organ-at-risk ct and mr segmentation dataset. *Med Phys*, 50(3):1917–1927, Mar 2023.

Lukas Radl, Yuan Jin, Antonio Pepe, Jianning Li, Christina Gsaxner, Fen-Hua Zhao, and Jan Egger. Aortic vessel tree (AVT) CTA datasets and segmentations, January 2022.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12179–12188, October 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Ha6RTeWMd0.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. URL https://arxiv.org/abs/1505.04597.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Spreading vectors for similarity search. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkGuG2R5tm.

Yudi Sang, Yanzhen Liu, Sutuke Yibulayimu, Yunning Wang, Benjamin D. Killeen, Mingxu Liu, Ping-Cheng Ku, Ole Johannsen, Karol Gotkowski, Maximilian Zenk, Klaus Maier-Hein, Fabian Isensee, Peiyan Yue, Yi Wang, Haidong Yu, Zhaohong Pan, Yutong He, Xiaokun Liang, Daiqi Liu, Fuxin Fan, Artur Jurgas, Andrzej Skalski, Yuxi Ma, Jing Yang, Szymon Płotka, Rafał Litka, Gang Zhu, Yingchun Song, Mathias Unberath, Mehran Armand, Dan Ruan, S. Kevin Zhou, Qiyong Cao, Chunpeng Zhao, Xinbao Wu, and Yu Wang. Benchmark of segmentation techniques for pelvic fracture in ct and x-ray: Summary of the pengwin 2024 challenge. *IEEE Transactions on Medical Imaging*, 45(5):2212–2228, May 2026. ISSN 1558-254X. doi: 10.1109/tmi.2025.3650126. URL http://dx.doi.org/10.1109/TMI.2025.3650126.

Wei Kwek Soh and Jagath C Rajapakse. Hybrid unet transformer architecture for ischemic stoke segmentation with mri and ct datasets. *Front Neurosci*, 17:1298514, 2023.

Karen-Helene Støverud, David Bouget, André Pedersen, Håkon Olav Leira, Tore Amundsen, Thomas Langø, and Erlend Fagertun Hofstad. Aeropath: An airway segmentation benchmark dataset with challenging pathology and baseline method. *PLoS One*, 19(10):e0311416, 2024.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20730–20740, 2022.

Joel Valdivia Ortega, Lorenz Lamm, Franziska Eckardt, Benedikt Schworm, Marion Jasnin, and Tingying Peng. Randomized-MLP regularization improves domain adaptation and interpretability in DINOv2. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=Rzqs4q9ooa.

Jasper W. van der Graaf, Miranda L. van Hooff, Constantinus F. M. Buckens, Matthieu Rutten, Job L. C. van Susante, Robert Jan Kroeze, Marinus de Kleuver, Bram van Ginneken, and Nikolas Lessmann. Lumbar spine segmentation in mr images: a dataset and a public benchmark. *Scientific Data*, 11(1):264, 2024.

Jasper Willem van der Graaf, Miranda L van Hooff, Constantinus F M Buckens, Matthieu Rutten, Job L C van Susante, Robert Jan Kroeze, Marinus de Kleuver, Bram van Ginneken, and Nikolas Lessmann. SPIDER - lumbar spine segmentation in MR images: a dataset and a public benchmark, 2023.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Andrew W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Jakob Wasserthal. Dataset with segmentations of 117 important anatomical structures in 1228 CT images, 2023.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiol Artif Intell*, 5(5):e230024, Sep 2023.

Meng Wei, Miaojing Shi, and Tom Vercauteren. Enhancing surgical instrument segmentation: integrating vision transformer insights with adapter. *International Journal of Computer Assisted Radiology and Surgery*, 19(7):1313–1320, 2024.

Qiaoyi Xu, Afzan Adam, Azizi Abdullah, Tao Chen, Xinglin Zhang, Adam Shephard, Patsy Ng Pei Sze, Noraidah Masir, Lei Li, and Reena Rahayu. Robust multi-domain digital pathology image segmentation via joint balancing representation learning. *Expert Systems with Applications*, 320: 132093, 2026.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2025. URL https://arxiv.org/abs/2303.00915.

Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Sid Kiblawi, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, Christine Moung-Wen, Brian Piening, Carlo Bifulco, Mu Wei, Hoifung Poon, and Sheng Wang. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature Methods*, 22(1):166–176, November 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02499-w. URL http://dx.doi.org/10.1038/s41592-024-02499-w.

Ziheng Zhao, Yao Zhang, Chaoyi Wu, Xiaoman Zhang, Xiao Zhou, Ya Zhang, Yanfeng Wang, and Weidi Xie. Large-vocabulary segmentation for medical images with text prompts. *npj Digital Medicine*, 8(1):566, 2025.

Qikui Zhu, Bo Du, and Pingkun Yan. Boundary-weighted domain adaptive neural network for prostate mr image segmentation. *IEEE Trans Med Imaging*, 39(3):753–763, Mar 2020.

A Technical Appendices and Supplementary Material

A.1 Mathematical Definitions and Proofs

A.1.1 Compactness

We provide some useful topological definitions and mathematical demonstrations to back up our decision to use the output of intermediate attention blocks instead of only the output of the last one. To do this, it is useful to study the topology of our latent spaces and the mathematical properties that our models have.

Since the ViTs we are considering are constrained to imaging spaces, we can assume their domains are bounded. Additionally, given the discrete properties of acquisition of devices, we can further understand our imaging spaces as discrete and thus, closed. This way, let us re-introduce a very famous property of sets in topology, namely *compactness*, and a well known result for them.

Definition 1. *Given X a topological space and $A \subseteq X$, we say that A is compact if for every open cover of A, i.e. $\{U_i\}_{i\in I}$ such $A \subseteq \bigcup_{i\in I}$ where U_i is open for every $i \in I$, there is $\{i_0, \cdots, i_n\} \subseteq I$ such that $A \subseteq \bigcup_{n=0}^{N}$.*

Theorem 2. *(Heine-Borel) A subset in $\mathbb{R}^n$ is compact if and only if it is closed and bounded.*

A.1.2 Enforced homeomorphism through prompting-based segmentation

Some compactness propertied being introduced, we will propose an understanding of prompt-based segmentation as the ability to disconnect sets and will show that the use of these prompts can also be used as a way of enforcing injectivity in the image encoder although the segmentation function or model remains non-injective. To begin this, let us begin with some well established definitions for clarity.

Definition 3. *Let $f : X \to Y$ be a function. We say that f is injective if for every $\{x, y\} \subseteq X$ such that $x \neq y$, we have that $f(x) \neq f(y)$.*

Definition 4. *Given we have a function $f : X \to Y$ with X and Y being topological spaces, we say that f is continuous if, for every $U \subseteq Y$ open, its inverse image under f, i.e. $f^{-1}[U]$, is open.*

Definition 5. *Given we have a function $f : X \to Y$ with X and Y being topological spaces, we say that f is open if, for every $U \subseteq Y$ open, its image under f, i.e. $f[U]$, is open.*

Definition 6. *Given we have a function $f : X \to Y$ with X and Y being topological spaces, we say that f is an homeomorphism if f is continuous and invertible and its inverse, $f^{-1} : Y \to X$ is also continuous.*

Having these concepts in mind, we propose some proofs to classic results in topology.

Remark 7. *If $f : X \to Y$ is an homeomorphism if and only if f is a continuous, open and invertible function.*

Proof. Since we already have invertibility and continuity by hypothesis, we just need to prove openness and continuity for the inverse.

This way, let us assume that f is an homeomorphism and prove that f needs to be open. Then, U being open, we have that $(f^{-1})^{-1}[U]$ is open due to f^{-1} being continuous. Furthermore, $(f^{-1})^{-1}[U] = f[U]$.

$\therefore f$ is open.

Let us now prove that f being open in this scenario implies that f^{-1} is continuous. If f is an open function and $U \subseteq X$ is an open set, we have that $f[U]$ is open while at the same time $f[U] = (f^{-1})^{-1}[U]$.

$\therefore f^{-1}$ is continuous. □

Lemma 8. *If a function $f : X \to Y$ is injective and continuous with compact domain and Hausdorff codomain, f is an homeomorphism on $f[X]$.*

Proof. Assuming the hypothesis from the lemma and using Remark 7, the only thing remaining to show is that f is open.

Thing being this way, let us take an open set $U \subseteq X$. Then $X \backslash U$ is closed which turns it into a compact set since X is compact. Thus, by continuity, $f[X \backslash U]$ is compact, which makes it closed since Y is Hausdorff, forcing $f[U]$ to be open.

□

With this equivalence in mind, let us introduce some helpful definitions as well as Yves-Meyer decomposition of functions and Tietze extension theorem to that will help us formulate our main result of the section.

Definition 9. *Given $u \in L_1(\Psi)$, we define the total variation of u in Ψ as*

$$V(u, \Psi) := \sup \left\{ \int_{\Psi} u(x) div\phi(x) dx : \phi \in C_c^1(\Psi, \mathbb{R}^m), \|\phi\|_{L^\infty(\psi)} \leqslant 1 \right\}.$$

Thus, we define the space functions of of bounded variation as

$$BV(\Psi) = \left\{ u \in L^1(\Psi) : V(u, \Psi) < \infty \right\}.$$

Definition 10. *Given a compact domain $\Psi \in \mathbb{R}^n$, we will say that a function $f : \Psi \to \mathbb{R}^m$ has an Yves-Meyer decomposition (Meyer, 2001) if there are*

$$u, v, w : \Psi \to \mathbb{R}^m$$

functions such that u is piece-wise smooth, $v \in BV(\Psi)$ and w is stochastic and

$$f = u + v + w.$$

This being the case, we say that u represents the texture, v the cartoon and w the noise of the function.

Having this definitions, we can make the next observation.

Remark 11. *Let*

$$X = \left\{ f : \Psi \to \mathbb{R}^3 : f \text{ is an image and admits an Yves-Meyer decomposition} \right\}.$$

Thus, being $p \leqslant 1$, the relation

$$u + v + w = f \sim g = \alpha + \beta + \gamma \Leftrightarrow \mathbb{E}[L_p(v - \beta)] = 0$$

is an equivalence relation.

Proof. Clearly, $\forall f \in X, f \sim f$ out of L_p being a norm.

$\therefore \sim$ is reflective

Further, taking $f, g \in X$ such that v and β are their cartoon components in the Yves-Mayer decomposition and assuming $f \sim g$, then $\mathbb{E}[L_p(v - \beta)] = 0$ from which we have symmetry follows given L_p is a norm.

$\therefore \sim$ is symmetric.

Lastly, taking $f, g, h \in X$ such that $f \sim g$, $g \sim h$ and v, β and κ are their cartoon components in the Yves-Meyer decomposition respectively, we have that $\mathbb{E}[L_p(v - \beta)] = 0 = \mathbb{E}[L_p(\beta - \phi)]$.

Thus, computing,

$$\mathbb{E}[L_p(u-\phi)] = \mathbb{E}[L_p(u-\beta+\beta-\phi)] \leqslant \mathbb{E}[L_p(u-\beta)] + \mathbb{E}[L_p(\beta-\phi)]$$

since L_p is a norm. This way, we have that

$$\mathbb{E}[L_p(u-\phi)] \leqslant 0$$

from the linearity of the expectation operator. Nevertheless, all norms are non-negative by definition, so $\mathbb{E}[L_p(u-\phi) \geqslant 0$. Then,

$$\mathbb{E}[L_p(u-\phi)] = 0$$

$\therefore \sim$ is transitive. □

Given we will discuss what we will understand by a prompt-based segmentation function, let us propose a definition for it and provide a result to justify our previous claim that this can be understood as the ability to separate closed prompt sets.

Definition 12. *Let us consider*

$$\Omega = \left\{ f : \Psi \to \mathbb{R}^3 : f \text{ is an image admitting an Yves-Meyer decomposition} \right\},$$

and a function $F : \Omega \times \mathcal{P}(\Psi) \times \mathcal{P}(\Psi) \to [0,1]^{|\Psi|}$ taking an image as its first argument while positive and negative prompts as its second and third ones, where $\mathcal{P}(\Psi)$ is the power set of Ψ.

Taking $f \in \Omega$, let it be v its cartoon component in the Yves-Meyer decomposition and $P_f \subseteq \Psi$ and $N_f \subseteq \Psi$ positive and negative prompts respectively.

Then we will say that F is an optimal segmentation function if, $\forall f \in \Omega$, $\forall P_f \subseteq \Psi$ and $\forall N_f \subseteq \Psi$,

$$F(A_v, P_f, N_f) = 1 \qquad \text{and} \qquad F(B_v, P_f, N_f) = 0$$

where A_v and B_V are the minimal set under contention such that $P_f \subseteq A_v$ and $N_f \subseteq B_v$ and contain at least one region with no empty interior and whose boundary is contained in the set of discontinuities of v.

Definition 13. *Let X be a topological space. We say X is normal if, given disjoint closed sets A and B, there are neighborhoods U of A and V of B that are also disjoint.*

Theorem 14. *(Tietze extension theorem) If X is a normal space, $A \subset X$ is closed, $f : A \to \mathbb{R}$ is continuous and $\mathbb{R}$ has the standard topology, there is a continuous function $F : X \to \mathbb{R}$ such that $F(a) = f(a)$ for all $a \in A$. Furthermore, F can be chosen to be bounded and to have the same bound as f.*

Lemma 15. *Being an optimal segmentation function is equivalent to being able to separate all closed positive and negative prompts with neighborhoods.*

Proof. Let us proof first how being an optimal segmentation function provides the function with the ability to separate closed positive from closed negative prompts. For this, let us take an optimal segmentation function $h : \Psi \to \mathbb{R}$ and some positive and negative prompts $P \subseteq \Psi$ and $N \subseteq \psi$ respectively.

Since h is optimal, we have $h[P, P, N] = 1$ and $h[N, P, N] = 0$ where $P \cup N$ is closed for being a finite union of closed sets. Further, P and N are disjoint since a point cannot be a positive and a negative prompt at the same time, turning $h[\cdot\,, P, N]$ into a continuous function. Thus, we finish by Tietze extension theorem.

$\therefore$ Optimal segmentation functions can separate all closed positive and negative prompts with neighborhoods.

Now, to prove that being able to separate positive and negative prompts with neighborhoods makes a function h an optimal segmentation function, we just need to remember that the positive and negative prompts are closed, which assures us that the minimal set will exist and will be contained in the neighborhoods that the hypothesis provides.

$\therefore h$ is an optimal segmentation function

□

Let us notice that the previous Lemma puts our understanding of a prompt-based segmentation function as one capable distinguishing the discontinuities of the cartoon component of the image to later use them to separate the positive versus negative prompts, proposing this way a mathematical formalization for this sort of models.

Theory being developed, we are able to formulate our main result of the section.

Theorem 16. *Let $\Psi \subseteq \mathbb{R}^n$ compact and*

$$\Omega' = \left\{ f : \Psi \to \mathbb{R}^3 : f \text{ is an image admitting an Yves-Meyer decomposition} \right\}.$$

Let us define $\Omega := \Omega'/\sim$ where $\sim$ is defined as in Remark 11.

Then, if $f : \Psi \to Y$ is an image encoder, $g : \mathcal{P}(\Psi) \to Y$ is a prompt encoder and

$$h : Y \times \mathcal{P}(Y) \times \mathcal{P}(Y) \to [0,1]^{|\Psi|}$$

is such that $h(f(\cdot), g(\cdot), g(\cdot))$ an optimal segmentation function, then f is injective.

Proof. Let us take $\alpha, \beta \in \Omega$ such that $\alpha \neq \beta$ and consider $g[\mathcal{P}(\mathcal{P}(\Psi))]$ the encoding through g of all possible prompts on our space. We can also observe that although that set might have a big cardinality, it will still be finite since Ψ is finite since being a domain of an function representing an image.

Thus, if we assume

$$h[f(\alpha), g[\mathcal{P}(\mathcal{P}(\Psi))], g[\mathcal{P}(\mathcal{P}(\Psi))]] = h[f(\beta), g[\mathcal{P}(\mathcal{P}(\Psi))], g[\mathcal{P}(\mathcal{P}(\Psi))]],$$

it would imply that for every combination of prompts, the segmentation masks was the same for α and β. Nevertheless, since $h(f(\cdot), g(\cdot), g(\cdot))$ is an optimal segmentation function, this means that it is able to distinguish all the cartoon components of α and β via prompts, but if $h[f(\alpha), g[\mathcal{P}(\mathcal{P}(\Psi))], g[\mathcal{P}(\mathcal{P}(\Psi))]] = h[f(\beta), g[\mathcal{P}(\mathcal{P}(\Psi))], g[\mathcal{P}(\mathcal{P}(\Psi))]]$, then we would have $\alpha \sim \beta$, which is a contraction.

$\therefore h[f(\alpha), g[\mathcal{P}(\mathcal{P}(\Psi))], g[\mathcal{P}(\mathcal{P}(\Psi))]] \neq h[f(\beta), g[\mathcal{P}(\mathcal{P}(\Psi))], g[\mathcal{P}(\mathcal{P}(\Psi))]]$.

However, h is a well-defined function, so the inequality implies that at least one of the arguments should be different. Nevertheless, the second and third are the same, leading the only remaining possibility to be $f(\alpha) \neq f(\beta)$.

$\therefore f$ is injective. □

Having this formulation and theorem, we can now prove the result that we need for our current work.

Corollary 17. *Assuming SAM2 is an optimal segmentation function, its image encoder is an homeomorphism.*

Proof. Let us notice SAM2 is defined, trained and used for inference as a function taking positive and negative prompts as well as images from a fixed image domain in which an Yves-Mayer decomposition is possible. Thus, using Theorem 16, its image encoder is injective.

Furthermore, this image encoder is continuous for being a composition of continuous functions, i.e. the transformer blocks that build it. Moreover, its domain is compact for its closed, bounded and metric and its codomain is Hausdorff for being metric. Thus, we finish using Lemma 8.

$\therefore$ SAM2's image encoder is an homeomorphism. □

A.1.3 HOMEOMORPHISMS UNDER COMPOSITION

In this subsection, we will provide some results on the implications of a function being the result of a composition and its relationship with being an homeomorphism.

Lemma 18. *If $f : X \to Y$ is an homeomorphism and $f = f_n \circ \cdots \circ f_1$, then f_i is an homeomorphism for every i.*

Proof. We will show this results by mathematical induction. The initial step, i.e. $f = f_1$ is true by hypothesis. Thus, let us take our induction hypothesis to be that our statement is true for $g = f_n \circ \cdots \circ f_1$ for $n \in \mathbb{N}$.

Without loss of generality, let us assume $f_i : X \to X$ for every $i \in \{1, \cdots n+1\}$ and take $f := f_{n+1} \circ g$ to be a homeomorphism. This way, let us begin by proving that f_{n+1} is injective.

Since f_i is an homeomorphism for every $i \in \{1, \cdots n\}$, g is an homeomorphism as well. Thus, let us take $\{x, y\} \subseteq X$ such that $x \neq y$. Then $\exists \{u, w\} \subseteq X$ such that $g(u) = x$ and $g(w) = y$ since g is surjective. In addition, $u \neq w$ due to g being a function and so it follows that $f(u) \neq f(w)$ because of f being an homeomorphism. This way we have that

$$f_{n+1}(x) = f_{n+1}(g(u)) = f(u) \neq f(w) = f_{n+1}(g(w)) = f_{n+1}(y).$$

$\therefore f_{n+1}$ is injective.

To prove surjectivity, let us take $y \in X$. Then $x \in X$ such that $f(x) = y$, but then $f_{n+1}(g(x)) = y$ where $g(x) \in X$.

$\therefore f_{n+1}$ is surjective.

Moving on for continuity, let us tke $U \subseteq X$ open $\Rightarrow f^{-1}[U]$ is open $\Rightarrow g[f^{-1}[U]]$ is open, where we have that

$$g[f^{-1}[u]] = g[(f_{n+1} \circ g)^{-1}[U]] = g \circ g^{-1} \circ f_{n+1}^{-1}[U] = f_{n+1}^{-1}[U].$$

$\therefore f_{n+1}$ is continuous.

Finally, to prove that f_{n+1} is open, let us take $U \subseteq X$ open $\Rightarrow g^{-1}[U]$ is open out of continuity from $g \Rightarrow f[g^{-1}[U]]$ is open out of f being open. This way we arrive to

$$f \circ g^{-1}[U] = f_{n+1} \circ g \circ g^{-1}[U] = f_{n+1}[U].$$

$\therefore f_{n+1}$ is open.

$\therefore f_{n+1}$ is an homeomorphism.

□

Remark 19. *If a function $f : X \to Y$ is such that $f = f_1 \circ f_2 \circ \cdots \circ f_n$, then f is an homeomorphism if an only if f_i is an homeomorphism for every i.*

Proof. The proof that if f_i being an homeomorphism for every i implies that f is an homeomorphism too comes directly from the fact that injectivity, surjectivity and being continuous are properties preserved by the finite composition of functions.

The proof that f being an homeomorphism forces f_i to be an homeomorphism for every finite i follows form Lemma 18. □

A.2 SUPPLEMENTARY TABLES

In this section, we provide supplementary Tables containing references from the datasets used in this paper as well as quantitative results regarding the performance of the different models tested in this work.

Table S1: Datasets, references they were released in and nnU-Net performance.

Dataset	Reference	nnU-Net performance reference
	CT	
Abdomen1K	Ma et al. (2022)	Ma et al. (2022)
Aorta	Radl et al. (2022)	Li et al. (2022)
AirwayTree	Støverud et al. (2024)	Støverud et al. (2024)
HaN-Seg	Podobnik et al. (2023)	Podobnik et al. (2023)
Lungs	Ma et al. (2021)	Ma et al. (2021)
TotalSeg	Wasserthal (2023)	Wasserthal et al. (2023)
	MRI	
AMOS	Ji et al. (2022)	Isensee et al. (2022)
BraTS	Baid et al. (2021); Bakas et al. (2017); Menze et al. (2015)	Isensee et al. (2020)
CHAOS-T1	Kavur et al. (2019)	Kavur et al. (2021)
Heart_ACDC	Li et al. (2023)	Li et al. (2023)
ISLES ADC/DWI	de la Rosa et al. (2025a)	de la Rosa et al. (2025b)
LeftAtrium	Antonelli et al. (2022)	Isensee et al. (2021)
ProstateADC/T2	Antonelli et al. (2022)	Isensee et al. (2021)
QIN-Lesion	Fedorov et al. (2019)	Zhu et al. (2020)
Spider	van der Graaf et al. (2023)	van der Graaf et al. (2024)

Table S2: Quantitative analysis of semantic segmentation performance across CT and MRI modalities. We report the mean mIoU $\pm$ standard deviation across five independent experimental runs for each encoder-decoder-pixel decoder configuration. nnU-Net (Isensee et al., 2021) was taken from literature. Bold and underlined indicates the models that performed the best and second best in mean per dataset respectively. † indicates statistical significance over all other models ($p<0.05$, Mann-Whitney U test (Virtanen et al., 2020)) Datasets are taken from the compilation made by BiomedParse (Zhao et al., 2024).

	DINOv2			SAM 2			
Dataset	Linear	ViT-UNet hybrid	ViTC-UNet	Linear	ViT-UNet hybrid	ViTC-UNet	nnU-Net
			CT				
Abdomen1K	0.04±0.03	0.07±0.05	**0.90±0.07**†	0.03±0.03	0.07±0.05	0.69±0.01†	0.85±0.05
Aorta	0.13±0.08	0.33±0.21	**0.98±0.01**†	0.04±0.03	0.35±0.21	0.75±0.04†	0.96
AirwayTree	0.01±0.01	0.01±0.01	**0.93±0.04**†	0.01±0.01	0.03±0.01	0.68±0.04†	**0.93±0.03**
HaN-Seg	0.01±0.01	0.06±0.01	**0.97±0.01**†	0.01±0.01	0.16±0.01	0.76±0.01†	0.77±0.07
Lungs	0.63±0.21	0.77±0.26	**0.90±0.03**†	0.63±0.21	0.78±0.26	0.85±0.06	0.67±0.07
TotalSeg	0.12±0.05	0.36±0.13	**0.87±0.06**†	0.03±0.02	0.33±0.16	0.55±0.04†	**0.87**
			MRI				
AMOS	0.30±0.12	0.56±0.25	**0.90±0.03**†	0.19±0.07	0.52±0.25	0.48±0.03	0.76
BraTS	0.08±0.13	0.19±0.21	**0.93±0.03**	0.07±0.10	0.13±0.16	0.43±0.04†	0.75
CHAOS-T1	0.57±0.01	0.71±0.03	**0.91±0.01**†	0.56±0.01	0.75±0.02	0.26±0.07	0.90±0.01
Heart_ACDC	0.18±0.17	0.22±0.22	**0.92±0.02**†	0.20±0.18	0.22±0.23	0.46±0.01†	0.62±0.22
ISLES ADC/DWI	0.01±0.01	0.01±0.01	**0.99±0.01**†	0.01±0.01	0.01±0.01	0.72±0.01†	0.69±0.08
LeftAtrium	0.28±0.01	0.63±0.03	**0.98±0.01**†	0.36±0.08	0.61±0.09	0.50±0.09	0.87
ProstateADC/T2	0.59±0.07	0.73±0.06	0.73±0.02	0.62±0.09	0.71±0.06	**0.92±0.05**†	0.71
QIN-Lesion	0.01±0.01	0.01±0.01	0.80±0.06†	0.01±0.01	0.01±0.01	**0.99±0.01**†	0.79
Spider	0.12±0.05	0.56±0.10	**0.84±0.05**†	0.09±0.05	0.44±0.08	0.61±0.08†	0.82±0.02

Table S3: Segmentation performance (mIoU, mean ± std over 5 runs) comparing ViTC-UNet with learnable structure tokens against fine-tuned hybrid (ViT-UNet (Valdivia Ortega et al., 2026), TransUNet (Chen et al., 2021), Swin UNETR (Tang et al., 2022)) and text-promptable (SAM 3 (Carion et al., 2026)) baselines. SAM 3's image and text encoder were frozen during training. All models were fine-tuned independently on each dataset from the BiomedParse benchmark compilation (Zhao et al., 2024) and evaluated on a shared holdout test set.

	Hybrid models			Promptable	Hybrid and promptable	
Dataset	ViT-UNet hybrid	TransUNet	Swin UNETR	SAM3	ViTC-UNet (SAM2)	ViTC-UNet (DINOv2)
			CT			
Abdomen1K	0.07±0.05	0.35±0.02	0.04±0.04	0.03±0.03	0.69±0.01	**0.90±0.07**
Aorta	0.33±0.21	0.33±0.01	0.06±0.06	0.05±0.01	0.75±0.04	**0.98±0.01**
AirwayTree	0.01±0.01	0.25±0.01	0.01±0.01	0.01±0.01	**0.93±0.04**	**0.93±0.04**
HaN-Seg	0.06±0.01	0.01±0.01	0.06±0.01	0.01±0.01	0.76±0.01	**0.97±0.01**
Lungs	0.77±0.26	0.70±0.07	0.61±0.11	0.14±0.07	0.85±0.06	**0.90±0.03**
TotalSeg	0.36±0.13	0.05±0.01	0.17±0.14	0.05±0.02	0.55±0.04	**0.87±0.06**
			MRI			
AMOS	0.56±0.25	0.10±0.01	0.21±0.14	0.04±0.01	0.48±0.03	**0.90±0.03**
BraTS	0.19±0.21	0.24±0.02	0.23±0.20	0.17±0.10	0.43±0.04	**0.93±0.03**
CHAOS-T1	0.71±0.03	0.31±0.01	0.43±0.07	0.07±0.08	0.26±0.07	**0.91±0.01**
Heart_ACDC	0.22±0.22	0.32±0.07	0.14±0.14	0.16±0.16	0.46±0.01	**0.92±0.02**
ISLES ADC/DWI	0.01±0.01	0.01±0.01	0.01±0.01	0.01±0.01	0.72±0.01	**0.99±0.01**
LeftAtrium	0.63±0.03	0.53±0.01	0.62±0.08	0.01±0.01	0.50±0.09	**0.98±0.01**
ProstateADC/T2	0.73±0.06	0.51±0.03	0.54±0.15	0.14±0.01	**0.92±0.05**	0.73±0.02
QIN-Lesion	0.01±0.01	0.20±0.01	0.01±0.01	0.01±0.01	**0.99±0.01**	0.80±0.06
Spider	0.56±0.10	0.32±0.01	0.16±0.05	0.10±0.04	0.61±0.08	**0.85±0.05**

Table S4: Ablation study of ViTC-UNet decoder variants using a frozen DINOv2 backbone and learnable structure tokens. Evaluated architectures are detailed in Section 4 and illustrated in Figure S1. Performance is reported as mIoU (mean ± std over 5 independent runs). Bold indicates top performance, underline indicates second best, and † denotes statistical significance over all other variants (p<0.05, two-sided Mann-Whitney U test (Virtanen et al., 2020)).

Dataset	Addition	Last shallow	Last deep	One-way att. t2e	One-way att. e2t	Unconditioned	ViTC-UNet
				CT			
Abdomen1K	0.80±0.08	0.66±0.06	0.73±0.06	**0.81±0.08**	0.80±0.08	0.04±0.04	**0.90±0.07**†
Aorta	0.77±0.01	0.56±0.02	0.53±0.05	**0.98±0.01**	**0.98±0.01**	0.08±0.01	**0.98±0.01**
AirwayTree	0.78±0.01	0.69±0.01	0.32±0.01	**0.95±0.01**	**0.95±0.01**	0.01±0.01	0.93±0.04
HaN-Seg	0.96±0.01	0.96±0.01	0.96±0.01	0.96±0.01	0.96±0.01	0.04±0.01	**0.97±0.01**
Lungs	0.58±0.02	0.50±0.02	0.84±0.04	**0.94±0.02**†	**0.94±0.02**	0.74±0.25	0.90±0.03
TotalSeg	0.85±0.06	0.85±0.02	0.82±0.07	0.85±0.06	0.85±0.06	0.15±0.08	**0.87±0.06**†
				MRI			
AMOS	0.86±0.02	0.83±0.03	0.78±0.04	0.86±0.02	0.87±0.02	0.44±0.21	**0.90±0.03**†
BraTS	0.74±0.03	0.69±0.02	0.73±0.02	**0.93±0.03**	**0.93±0.03**	0.15±0.17	**0.93±0.03**
CHAOS-T1	0.89±0.02	0.83±0.02	0.84±0.02	**0.93±0.01**	0.92±0.01	0.58±0.05	0.91±0.01
Heart_ACDC	0.71±0.02	0.45±0.02	0.87±0.04	0.90±0.02	0.90±0.02	0.10±0.10	**0.92±0.02**†
ISLES ADC/DWI	0.98±0.01	0.98±0.01	0.98±0.01	0.98±0.01	0.98±0.01	0.01±0.01	**0.99±0.01**
LeftAtrium	0.54±0.04	0.74±0.01	0.35±0.01	**0.98±0.01**	**0.98±0.01**	0.51±0.02	**0.98±0.01**
ProstateADC/T2	0.63±0.07	0.33±0.01	0.32±0.01	0.70±0.02	0.69±0.02	0.50±0.09	**0.73±0.02**
QIN-Lesion	0.01±0.01	0.62±0.01	0.48±0.10	0.80±0.01	**0.82±0.01**	0.01±0.01	0.80±0.06
Spider	0.78±0.05	0.73±0.04	0.77±0.05	**0.84±0.04**	**0.84±0.04**	0.30±0.08	**0.84±0.05**
				CT and MRI			
Mean	0.72±0.03	0.69±0.02	0.69±0.03	0.89±0.02	0.89±0.02	0.24±0.07	**0.90±0.02**

Table S5: Ablation study of ViTC-UNet decoder variants using a frozen SAM 2 backbone and learnable structure tokens. Evaluated architectures are detailed in Section 4 and illustrated in Figure S1. Performance is reported as mIoU (mean ± std over 5 independent runs). Bold indicates top performance, underline indicates second best, and † denotes statistical significance over all other variants ($p<0.05$, two-sided Mann-Whitney U test (Virtanen et al., 2020)).

Dataset	Addition	Last shallow	Last deep	One-way att. t2e	One-way att. e2t	Unconditioned	ViTC-UNet
				CT			
Abdomen1K	0.42±0.06	**0.54±0.04**	0.22±0.07	0.11±0.01	0.23±0.02	0.04±0.04	0.69±0.01†
Aorta	0.49±0.03	0.68±0.01	0.28±0.01	**0.97±0.01**	0.84±0.02	0.01±0.01	0.75±0.04
AirwayTree	0.15±0.08	0.70±0.01	0.39±0.05	**0.90±0.01**	0.88±0.01	0.03±0.01	0.93±0.04†
HaN-Seg	0.21±0.01	0.63±0.01	0.11±0.01	0.04±0.01	0.01±0.01	0.12±0.01	**0.76±0.01**
Lungs	0.33±0.03	0.68±0.03	0.33±0.04	0.73±0.03	0.75±0.04	0.74±0.25	**0.85±0.06**
TotalSeg	0.08±0.01	0.50±0.04	0.22±0.01	0.19±0.01	0.15±0.02	0.16±0.09	**0.55±0.04**†
				MRI			
AMOS	0.05±0.01	0.42±0.02	0.18±0.01	0.05±0.01	0.02±0.01	0.43±0.20	**0.48±0.03**
BraTS	0.35±0.01	**0.61±0.02**	0.21±0.01	0.60±0.03	0.29±0.02	0.15±0.10	0.43±0.04
CHAOS-T1	0.57±0.04	0.40±0.01	0.21±0.01	0.45±0.01	0.40±0.01	**0.64±0.06**	0.26±0.07
Heart_ACDC	0.01±0.01	0.72±0.04	0.28±0.06	**0.55±0.08**†	0.41±0.01	0.10±0.10	0.46±0.01
ISLES ADC/DWI	0.26±0.04	0.57±0.02	0.41±0.03	0.07±0.01	0.01±0.01	0.29±0.05	**0.72±0.01**†
LeftAtrium	0.40±0.03	0.68±0.01	0.37±0.01	0.87±0.01	**0.94±0.02**†	0.44±0.02	0.50±0.09
ProstateADC/T2	0.26±0.09	0.55±0.03	0.18±0.01	**0.93±0.02**†	0.90±0.02	0.49±0.09	0.92±0.05
QIN-Lesion	0.37±0.13	0.65±0.01	0.22±0.01	0.92±0.01	0.75±0.03	0.01±0.01	**0.99±0.01**
Spider	0.33±0.07	0.46±0.02	0.05±0.01	0.41±0.01	0.33±0.03	0.25±0.08	**0.61±0.08**†
				CT and MRI			
Mean	0.29±0.04	0.58±0.02	0.24±0.02	0.52±0.01	0.46±0.01	0.26±0.07	**0.64±0.03**

A.3 SUPPLEMENTARY FIGURES

In this section, we provide more qualitative examples on the performance of our proposed ViTC-UNet when using DINOv2 as encoder for segmenting anatomical structures on CT and MRI modalities.



(a) ViTC-UNet (b) Unconditioned (c) Last deep

(d) Last shallow (e) Addition (f) Oneway att. t2e

(g) Oneway att. e2t (h) One-way attention block (i) Two-way attention block

Figure S1: Overview of decoder variants evaluated in the ablation study. Modified components are highlighted in color, while unchanged modules remain in grayscale. Variants include: (b) Unconditioned, omitting structure tokens; (c) Last deep and (d) Last shallow, routing only the final decoder state to the deepest or shallowest UNet layers, respectively; (e) Addition, adding structure tokens directly to image embeddings via self-attention; and (f) Oneway att. t2e and (g) Oneway att. e2t, substituting two-way cross-attention with directional one-way attention.

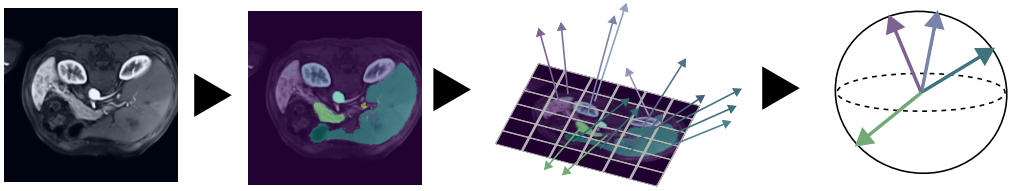


Figure S2: Computation of structure anchors and their angular separation. Given an input image and its ground-truth annotations, patch tokens intersecting each target class mask are extracted from the ViT embeddings and averaged to form a *structure anchor*. Following L_2 normalization, pairwise angles between the resulting vectors are computed to measure latent class separability.

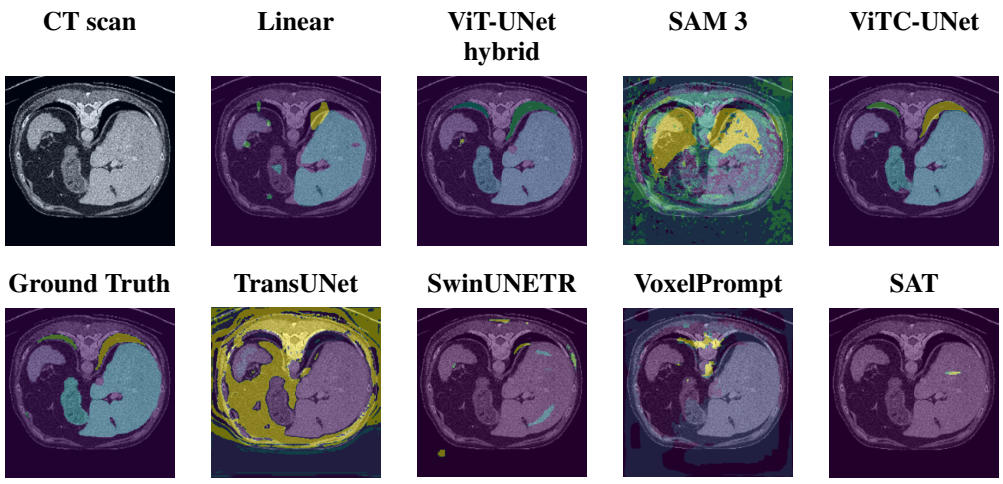


Figure S3: Comparison of segmentation performance between a linear baseline, the ViT-UNet hybrid (Valdivia Ortega et al., 2026), and our proposed ViTC-UNet utilizing DINOv2 (Oquab et al., 2024) as encoder for semantic segmentation on the Totalseg (Wasserthal, 2023) CT dataset.

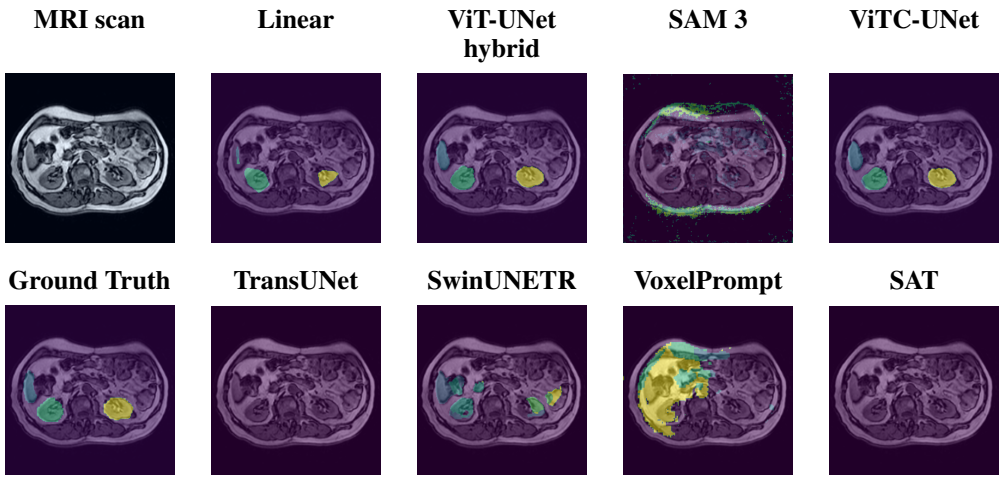


Figure S4: Comparison of segmentation performance between a linear baseline, the ViT-UNet hybrid (Valdivia Ortega et al., 2026), and our proposed ViTC-UNet utilizing DINOv2 (Oquab et al., 2024) as encoder for semantic segmentation on the CHAOS (Kavur et al., 2019) MRI dataset.

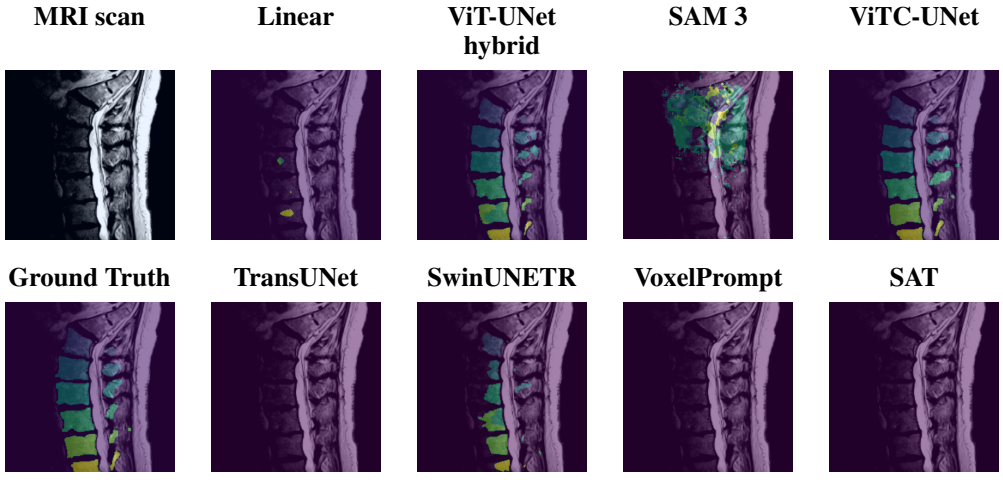


Figure S5: Comparison of segmentation performance between a linear baseline, the ViT-UNet hybrid (Valdivia Ortega et al., 2026), and our proposed ViTC-UNet utilizing DINOv2 (Oquab et al., 2024) as encoder for semantic segmentation on the Spider (van der Graaf et al., 2023) MRI dataset.

A.4 LICENSES

The datasets used in this paper where obtained as part of the compilation made by BiomedParse (Zhao et al., 2024) who release it under the Creative Commons Attribution Non Commercial Share Alike 4.0 license. This work adheres to this license. These datasets were originally released in the publications listed in Table S1.

This paper used the models DINOv2 (Oquab et al., 2024) and SAM 2 (Ravi et al., 2025) who were released under a Apache License 2.0. Their use adhered to this license in this paper.

Furthermore, the presented work was compared against SAT (Zhao et al., 2025) and Voxel-Prompt (Hoopes et al., 2025), released under a Apache License 2.0 and MIT License respectively. Their use respected these licenses.