MASKMED: DECOUPLED MASK AND CLASS PREDICTION FOR MEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical image segmentation methods typically use point-wise convolutional heads that map output channels to predefined semantic classes. While effective, this formulation imposes a rigid channel-to-class correspondence that limits flexible spatial-semantic reasoning and feature reuse across classes. We propose *MaskMed*, a volumetric segmentation framework that decouples mask generation from class prediction. Instead of directly predicting multi-class logits, MaskMed produces class-agnostic binary masks together with their semantic labels using a shared set of object queries. This formulation removes fixed class-channel bindings and enables more adaptive semantic reasoning over anatomical structures. To improve multi-scale feature integration, we introduce a *Full-Scale Deformable Transformer* that enables deformable cross-scale interaction between encoder and decoder features across the full feature hierarchy while preserving spatial alignment. We further design a masked multi-scale segmentation head with shared query propagation and finest-scale bipartite supervision for consistent coarse-to-fine refinement across decoder stages. Extensive experiments on AMOS22, BTCV, ACDC, and BraTS23 demonstrate state-of-the-art performance. MaskMed improves Dice scores over nnU-Net by +2.0 and +1.3 points on AMOS22 Task 1 and Task 2, respectively, as well as by +1.7 on ACDC, +0.9 on BTCV, and +1.0 on BraTS23.

1 INTRODUCTION

Medical image segmentation plays a critical role in many clinical applications, including organ delineation, tumor quantification, and treatment planning. Most modern segmentation frameworks, such as UNet Ronneberger et al. (2015) and its variants (e.g., nnUNet Isensee et al. (2021), UNETR Hatamizadeh et al. (2022), nnFormer Zhou et al. (2023), and SwinUNETR Tang et al. (2022)), follow a common design paradigm: an encoder–decoder architecture combined with a lightweight segmentation head. Typically, the segmentation head applies a point-wise convolution to map high-dimensional features to voxel-wise class logits optimized using cross-entropy and Dice losses.

Despite its strong empirical performance, this formulation imposes a rigid inductive bias on semantic prediction. The point-wise segmentation head can be interpreted as a learnable linear projection in which each output channel is permanently associated with a predefined semantic category. Under this design, semantic classes are tied to fixed channel indices regardless of context. While efficient, such a formulation restricts flexible spatial-semantic reasoning and adaptive feature reuse across classes.

This limitation is particularly relevant in medical image segmentation, where anatomical structures are often spatially adjacent, morphologically variable, partially overlapping in appearance, or modality-dependent. Abdominal organs, tumor subregions, and cardiac structures frequently require flexible spatial-semantic reasoning rather than independent class-wise channel prediction.

Such a design introduces two key limitations. First, tying semantic categories to fixed output channels restricts the model's ability to learn flexible, data-driven relationships between spatial patterns and semantic concepts. Although backbone features are shared across classes, the segmentation head itself does not explicitly model interactions among semantic categories. Second, the predefined channel-to-class correspondence limits contextual adaptation and feature reuse, which are particularly important when segmenting anatomically similar or spatially overlapping structures.

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.

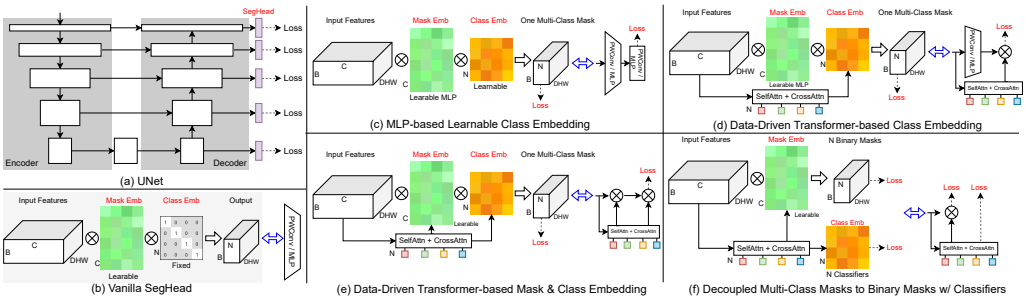


Figure 1: Comparison of segmentation head architectures, illustrating the progression from conventional UNet-based designs (a,b) to our proposed decoupled mask and class embedding framework (f).

To address these limitations, we propose a unified decoupled segmentation head that predicts binary masks together with their associated semantic labels without enforcing predefined class-channel bindings. Instead of directly producing multi-class logits, the model learns class-agnostic mask embeddings and semantic class embeddings that interact through shared object queries. This formulation enables more flexible compositional reasoning and allows the model to adaptively associate spatial patterns with semantic concepts. From this perspective, conventional point-wise segmentation heads can be viewed as a special case of our formulation in which semantic classes are fixed identity vectors and mask prediction reduces to channel-wise linear projection.

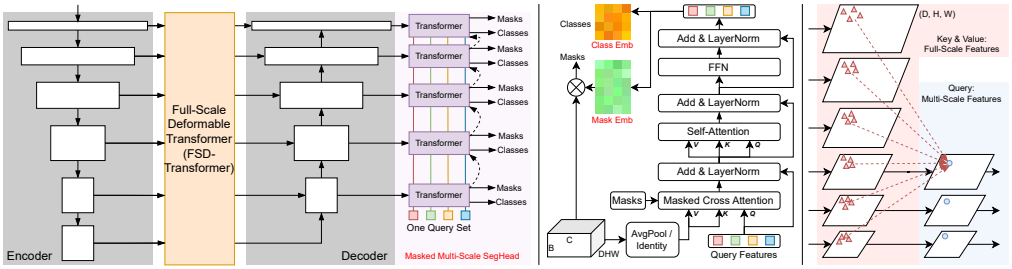
Another key observation is that although spatial resolution changes across decoder stages, the semantic queries responsible for mask and class prediction should preserve consistent semantic correspondence across scales. Motivated by this insight, we introduce a hierarchical query propagation mechanism in which shared object queries are progressively refined through masked cross-attention across decoder stages. Predictions from lower-resolution stages serve as spatial priors to guide higher-resolution attention, enabling coarse-to-fine refinement of segmentation results. This design leads to our *Masked Multi-Scale Segmentation Head*, which integrates multi-scale features while maintaining semantic consistency across decoder stages.

To further enhance encoder–decoder interaction, we introduce the *Full-Scale Deformable Transformer (FSD-Transformer)*. Unlike conventional skip connections that fuse only same-scale features, the proposed module enables deformable cross-scale interaction across the full encoder hierarchy. Specifically, low-resolution encoder features are used as queries to interact with multi-scale encoder features through deformable attention, enabling spatially aligned aggregation of global semantic and fine structural information.

Our main contributions are summarized as follows. i) We propose *MaskMed*, a unified segmentation framework that generalizes conventional point-wise segmentation heads into a fully decoupled formulation in which class-agnostic mask prediction and semantic classification are learned separately yet optimized jointly. ii) We introduce 3D bipartite matching between predicted mask–class pairs and ground-truth annotations, and leverage the finest-scale matching assignments to supervise all decoder stages, enabling effective UNet-style deep supervision. iii) We further design a *Masked Multi-Scale Segmentation Head* with shared query propagation across decoder stages for semantically consistent coarse-to-fine refinement through masked attention. iv) In addition, we propose the *Full-Scale Deformable Transformer*, which performs deformable cross-scale fusion over the full encoder hierarchy while preserving spatial alignment. v) Extensive experiments on AMOS22, BTCV, ACDC, and BraTS23 demonstrate state-of-the-art performance, with MaskMed improving Dice scores over nnUNet by +2.0 and +1.3 points on AMOS22 Task 1 and Task 2, respectively, as well as by +1.7 on ACDC, +0.9 on BTCV, and +1.0 on BraTS23.

2 Related Work

Deep learning has significantly advanced medical image segmentation, with fully convolutional architectures such as U-Net Ronneberger et al. (2015) forming the foundation of many modern approaches. Building on this architecture, nnU-Net Isensee et al. (2021) automated architecture configuration and training pipelines, achieving strong performance across numerous 3D benchmarks.



(a) Architecture overview (b) Transformer in SegHead (c) FSD-Transformer

Figure 2: **Model Architecture Overview.** (a) Overall architecture: an encoder–decoder framework enhanced with a Full-Scale Deformable Transformer (FSD-Transformer) that bridges multi-scale encoder features and decoder inputs. (b) Masked Multi-Scale Segmentation Head: a shared set of queries predicts mask and class embeddings through transformer layers. (c) FSD-Transformer: deformable attention across the full feature hierarchy using multi-scale queries and full-scale key/value features.

More recently, transformer-based models, including UNETR Hatamizadeh et al. (2022), Swin UNETR Hatamizadeh et al. (2021), and nnFormer Zhou et al. (2023), have improved global context modeling. Despite these advances, most methods still rely on point-wise convolutional segmentation heads with fixed channel-to-class assignments, limiting flexible spatial–semantic reasoning and inter-class interaction.

Mask classification offers an alternative formulation by predicting class-agnostic binary masks with associated semantic labels. MaskFormer Cheng et al. (2021) and Mask2Former Cheng et al. (2022) show that decoupled mask and class prediction with bipartite matching enables flexible and compositional segmentation in natural images. However, directly applying this paradigm to volumetric medical segmentation leaves important mismatches with UNet-style medical architectures, including deep supervision across decoder stages, same-scale prediction from decoder features, and full-scale encoder–decoder fusion. MaskMed addresses these mismatches through finest-scale matching supervision shared across stages, query propagation over same-scale decoder features, and deformable fusion over the full encoder hierarchy.

Deformable mechanisms improve spatial adaptability. Deformable Convolutional Networks Dai et al. (2017) and deformable attention models such as Deformable DETR Zhu et al. (2021) enable content-adaptive feature aggregation across spatial locations. In medical imaging, deformable modules have been explored primarily in registration Chen et al. (2022), while their use in segmentation remains limited. We introduce a *Full-Scale Deformable Transformer* to facilitate spatially aligned cross-scale interaction between encoder and decoder for medical image segmentation.

3 The Proposed Method

3.1 Rethinking Segmentation Heads

Medical image segmentation is commonly formulated with convolutional encoder–decoder architectures such as U-Net, as illustrated in Fig. 1(a). These models attach point-wise convolutional segmentation heads (SegHeads) at one or multiple decoder stages (Fig. 1(b)). The encoder extracts hierarchical representations, while the decoder reconstructs a feature map $\mathbf{F} \in \mathbb{R}^{B\times C\times D\times H\times W}$, where B is the batch size, C is the channel dimension, and (D, H, W) are spatial dimensions. Each SegHead projects $\mathbf{F}$ from C channels to N class logits using a point-wise convolution, producing an output tensor in $\mathbb{R}^{B\times N\times D\times H\times W}$, followed by softmax normalization over classes.

As illustrated in Fig. 1(b), a point-wise convolutional SegHead can be interpreted as a linear projection parameterized by a weight matrix of size $C \times N$. This operation maps decoder features to logit masks, each tied to a semantic class. Conceptually, this formulation introduces an implicit identity class embedding $\mathbf{I}_{\text{cls}} \in \mathbb{R}^{N\times N}$ that enforces a one-to-one correspondence between output channels and class labels, so that the i-th output channel exclusively represents the i-th semantic category.

Let $\mathbf{Y}_i \in \{0,1\}^{D\times H\times W}$ denote the ground-truth binary mask for class or region i, and $y_i \in \{1,\ldots,K\}$ its semantic label. The conventional channel-wise formulation applies a softmax over

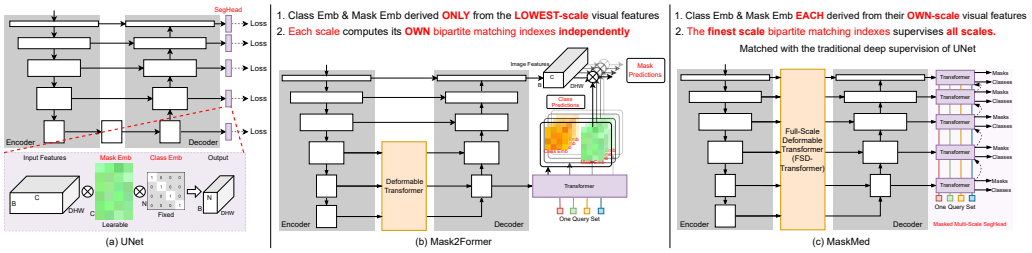


Figure 3: **Comparison with UNet and Mask2Former.** (a) UNet: class and mask embeddings are produced independently at each decoder scale. (b) Mask2Former: class and mask embeddings are generated from the lowest-resolution scale. (c) MaskMed: class and mask embeddings are produced at each scale, while bipartite matching computed at the finest scale supervises all scales.

class channels and supervises the output using Dice and Binary Cross Entropy (BCE) losses:

$$\mathcal{L} = \sum_{i=1}^{N} \mathcal{L}_{\text{Dice+BCE}}(\text{softmax}([\mathbf{F} \cdot \mathbf{M}_{\text{mask}} \cdot \mathbf{I}_{\text{cls}}]_i), \mathbf{Y}_i). \tag{1}$$

We progressively relax this rigid formulation through the segmentation head variants in Fig. 1.

Learnable Class Embedding via Point-wise Projection (Fig. 1(c)). We first replace the fixed identity class embedding with a learnable matrix $\mathbf{M}_{\text{cls}}$, implemented as a point-wise $(1 \times 1 \times 1)$ projection. This allows the model to learn how mask channels contribute to semantic classes rather than enforcing a fixed one-hot mapping:

$$\mathcal{L} = \sum_{i=1}^{N} \mathcal{L}_{\text{Dice+BCE}}(\text{softmax}([\mathbf{F} \cdot \mathbf{M}_{\text{mask}} \cdot \mathbf{M}_{\text{cls}}]_i), \mathbf{Y}_i). \tag{2}$$

Data-Driven Class Embedding via Transformer (Fig. 1(d)). To make class embeddings input-adaptive, we introduce a transformer module composed of self-attention, cross-attention, and feed-forward layers. Initialized with learnable queries $\mathbf{E}_Q \in \mathbb{R}^{N \times E}$, the module generates class embeddings conditioned on the feature map:

$$\mathcal{L} = \sum_{i=1}^{N} \mathcal{L}_{\text{Dice+BCE}}\big(\text{softmax}\big([\mathbf{F} \cdot \mathbf{M}_{\text{mask}} \cdot f(\mathbf{E}_Q, \mathbf{F})_{\text{cls}}]_i\big), \mathbf{Y}_i\big). \tag{3}$$

Data-Driven Mask and Class Embedding via Transformer (Fig. 1(e)). We further replace the fixed mask projection with a transformer-generated mask embedding, so that both mask and class embeddings are dynamically conditioned on the input:

$$\mathcal{L} = \sum_{i=1}^{N} \mathcal{L}_{\text{Dice+BCE}}\big(\text{softmax}\big([\mathbf{F} \cdot f(\mathbf{E}_Q, \mathbf{F})_{\text{mask}} \cdot f(\mathbf{E}_Q, \mathbf{F})_{\text{cls}}]_i\big), \mathbf{Y}_i\big). \tag{4}$$

Decoupled Binary Masks and Class Prediction (Fig. 1(f)). Finally, we introduce a fully decoupled SegHead that separates spatial mask generation from semantic classification. Instead of directly predicting multi-class logits, the model generates N class-agnostic binary masks, each representing a candidate region, and predicts a semantic label for each mask. Binary masks are supervised with Dice and BCE losses, while mask labels are supervised with cross-entropy.

To remove the rigid channel-to-class mapping, we adopt 3D bipartite matching during training. Each predicted mask–class pair, represented by an object query, is matched to the ground truth that yields the highest similarity. Given the matching function σ, the loss is:

$$\mathcal{L} = \sum_{i=1}^{N} \Big[\mathcal{L}_{\text{Dice+BCE}}\Big(\text{sigmoid}\big([\mathbf{F} \cdot f(\mathbf{E}_Q, \mathbf{F})_{\text{mask}}]_{\sigma(i)}\big), \mathbf{Y}_i\Big) + \mathcal{L}_{\text{CE}}\Big([f(\mathbf{E}_Q, \mathbf{F})_{\text{cls}}]_{\sigma(i)}, y_i\Big)\Big]. \tag{5}$$

This formulation subsumes the preceding segmentation heads as special cases. The classical channel-wise softmax head is recovered by using one query per class, fixed identity class embeddings,

Dataset	Method	Spleen	R.Kd	L.Kd	GB	Eso.	Liver	Stom.	Aorta	IVC	Panc.	RAG	LAG	Duo.	Blad.	Pros.	Avg.
AMOS-CT Task 1	UNETR Hatamizadeh et al. (2022)	0.928	0.913	0.903	0.719	0.763	0.955	0.849	0.922	0.838	0.766	0.663	0.663	0.662	0.815	0.744	0.807
	nnFormer Zhou et al. (2023)	0.950	0.948	0.944	0.789	0.784	0.967	0.914	0.931	0.868	0.828	0.654	0.695	0.759	0.865	0.773	0.845
	SwinUNETR Hatamizadeh et al. (2021)	0.954	0.954	0.950	0.819	0.852	0.972	0.919	0.955	0.911	0.875	0.775	0.801	0.816	0.895	0.812	0.884
	SwinUNETRv2 He et al. (2023)	0.959	0.962	0.958	0.842	0.867	0.976	0.933	0.957	0.920	0.889	0.783	0.812	0.843	0.913	0.836	0.897
	3D UX-Net Lee et al. (2022)	0.955	0.956	0.953	0.826	0.858	0.972	0.922	0.955	0.915	0.881	0.781	0.809	0.820	0.902	0.823	0.889
	nnU-Net Isensee et al. (2021)	0.951	0.961	0.956	0.826	0.869	0.973	0.931	0.957	0.923	0.880	0.784	0.809	0.846	0.898	0.827	0.893
	MaskSAM Xie et al. (2025)	0.963	**0.973**	**0.969**	0.872	0.876	**0.982**	0.940	**0.962**	0.922	0.888	0.794	0.813	0.851	0.920	0.854	0.905
	MaskFormer Cheng et al. (2021)	0.946	0.959	0.953	0.790	0.835	0.969	0.914	0.948	0.905	0.853	0.735	0.671	0.814	0.865	0.787	0.863
	Mask2Former Cheng et al. (2022)	0.954	0.966	0.962	0.819	0.855	0.972	0.934	0.954	0.917	0.874	0.760	0.779	0.844	0.891	0.809	0.886
	MaskMed	**0.971**	0.970	**0.969**	**0.880**	**0.890**	0.981	**0.950**	0.961	**0.929**	**0.904**	**0.801**	**0.826**	**0.875**	**0.931**	**0.859**	**0.913**
AMOS-CTMRI Task 2	UNETR Hatamizadeh et al. (2022)	0.925	0.938	0.923	0.737	0.776	0.948	0.854	0.918	0.836	0.803	0.672	0.684	0.723	0.814	0.757	0.821
	nnFormer Zhou et al. (2023)	0.945	0.946	0.942	0.773	0.779	0.971	0.903	0.931	0.856	0.826	0.696	0.726	0.740	0.837	0.741	0.840
	SwinUNETR Hatamizadeh et al. (2021)	0.953	0.963	0.955	0.795	0.828	0.976	0.919	0.949	0.891	0.858	0.733	0.764	0.793	0.888	0.785	0.869
	SwinUNETRv2 He et al. (2023)	0.951	0.960	0.949	0.815	0.838	0.962	0.911	0.940	0.887	0.862	0.740	0.777	0.809	0.879	0.801	0.872
	3D UX-Net Lee et al. (2022)	0.948	0.950	0.945	0.814	0.832	0.972	0.915	0.945	0.887	0.869	0.738	0.772	0.802	0.873	0.806	0.870
	nnU-Net Isensee et al. (2021)	0.955	0.961	0.957	0.823	0.841	0.976	0.919	0.952	0.894	0.877	0.744	0.787	0.819	0.880	0.816	0.880
	MaskFormer Cheng et al. (2021)	0.942	0.952	0.944	0.759	0.794	0.970	0.903	0.937	0.856	0.826	0.692	0.722	0.743	0.857	0.725	0.841
	Mask2Former Cheng et al. (2022)	0.949	0.960	0.952	0.784	0.808	0.974	0.914	0.941	0.875	0.847	0.711	0.749	0.777	0.882	0.752	0.858
	MaskMed	**0.963**	**0.967**	**0.960**	**0.851**	**0.855**	**0.978**	**0.939**	**0.953**	**0.907**	**0.889**	**0.765**	**0.794**	**0.843**	**0.917**	**0.825**	**0.893**

Table 1: Official AMOS22 leaderboard results on CT-only Task 1 and CT+MRI Task 2. All methods use 5-fold cross-validation ensembling.

Method	Spl.	R.Kd	L.Kd	GB	Eso.	Liv.	Stom.	Aorta	IVC	Veins	Panc.	AG	DSC
UNETR Hatamizadeh et al. (2022)	0.968	0.924	0.941	0.750	0.766	0.971	0.913	0.890	0.847	0.788	0.767	0.741	0.856
SwinUNETR Hatamizadeh et al. (2021)	**0.971**	0.936	0.943	**0.794**	0.773	**0.975**	0.921	0.892	0.853	**0.812**	0.794	**0.765**	0.869
3D UX-Net Lee et al. (2022)	0.946	0.942	0.943	0.593	0.722	0.964	0.734	0.872	0.849	0.722	0.809	0.671	0.814
nnU-Net Isensee et al. (2021)	0.970	0.953	0.953	0.635	0.775	0.974	0.891	0.901	0.885	0.790	0.871	0.752	0.863
nnFormer Zhou et al. (2023)	0.935	0.949	0.950	0.641	**0.795**	0.968	0.901	0.897	0.859	0.778	0.856	0.739	0.856
MaskFormer Cheng et al. (2021)	0.963	0.946	0.950	0.561	0.755	0.969	0.892	0.894	0.858	0.738	0.826	0.674	0.835
Mask2Former Cheng et al. (2022)	0.965	0.949	0.950	0.626	0.772	0.970	0.897	0.897	0.865	0.752	0.835	0.706	0.849
MaskMed	0.970	**0.952**	**0.956**	0.680	0.792	0.974	**0.922**	0.915	**0.887**	0.799	**0.875**	0.742	**0.872**

Table 2: Comparison with SOTA methods on BTCV. MaskFormer and Mask2Former are adapted to 3D.

Method	RV ↑	Myo ↑	LV ↑	DSC ↑
UNETR Hatamizadeh et al. (2022)	88.51	88.35	92.78	89.88
SwinUNETR Hatamizadeh et al. (2021)	89.97	89.07	93.12	90.72
3D UX-Net Lee et al. (2022)	89.13	88.53	92.94	90.20
nnU-Net Isensee et al. (2021)	90.59	89.52	93.86	91.32
nnFormer Zhou et al. (2023)	92.09	90.74	94.40	92.41
MaskFormer Cheng et al. (2021)	88.22	87.39	92.15	89.25
Mask2Former Cheng et al. (2022)	90.05	87.99	92.86	90.30
MaskMed	**92.46**	**91.59**	**94.97**	**93.01**

Table 3: Comparison on ACDC using 5-fold ensembling.

deterministic class-to-channel assignment, and coupled mask/class projection. The variants in Fig. 1(c–e) progressively relax these assumptions through learnable or input-conditioned embeddings, but mask prediction and class prediction remain implicitly coupled. In contrast, our final formulation explicitly separates spatial mask generation from semantic classification, enabling data-driven assignment between predicted regions and semantic labels.

3.2 Masked Multi-Scale Segmentation Head

Existing mask-classification frameworks such as MaskFormer and Mask2Former derive mask and class embeddings mainly from the lowest-resolution feature map and perform bipartite matching independently at each decoder layer. We find this design suboptimal for volumetric medical segmentation, where UNet-style deep supervision remains highly effective. Two UNet principles are particularly relevant: predictions are generated from decoder features at the same resolution, and the semantic correspondence of prediction channels remains consistent across scales, as illustrated in Fig. 3(a).

Motivated by these observations, we design the segmentation head so that mask and class embeddings at each stage are produced from same-scale decoder features, while the bipartite matching computed at the finest resolution is shared across all decoder stages for supervision. This differs from Mask2Former, where embeddings and matching are stage-independent, as shown in Fig. 3(b,c). Sharing the finest-scale assignment provides a stable query-to-structure target for coarse-to-fine refinement; in contrast, independent matching at each scale can introduce inconsistent assignments across decoder stages and weaken deep supervision.

Building on this design, we propose a *Masked Multi-Scale Segmentation Head* with shared object queries across decoder stages. A single learnable query set is initialized at the lowest-resolution stage and progressively refined as it propagates to higher resolutions, enforcing semantic consistency across scales. Predicted masks from lower-resolution stages are further used as spatial priors in masked cross-attention at subsequent stages. Each stage outputs mask and class embeddings through lightweight MLPs: mask embeddings interact with the corresponding decoder feature map to generate dense masks, while class embeddings predict semantic labels.

3.3 Full-Scale Deformable Transformer

We next improve encoder-decoder fusion, which is essential for preserving fine spatial details in high-resolution volumetric segmentation. A direct approach is to enhance skip-connected encoder features with transformer attention. However, global self-attention over dense 3D feature maps

Model	Dice ↑				HD95 ↓			
	ET	TC	WT	Avg.	ET	TC	WT	Avg.
UNETR Hatamizadeh et al. (2022)	78.50	84.77	92.27	85.18	27.93	10.76	5.02	14.57
3D UX-Net Lee et al. (2022)	81.20	85.85	92.37	86.47	20.44	11.88	4.98	12.43
nnFormer Zhou et al. (2023)	81.22	85.99	92.24	86.48	19.01	10.05	4.56	11.21
SwinUNETR Hatamizadeh et al. (2021)	81.25	86.51	92.60	86.79	19.94	9.67	4.56	11.39
nnU-Net Isensee et al. (2021)	81.40	87.46	92.61	87.16	19.82	7.55	4.54	10.64
MaskFormer Cheng et al. (2021)	79.86	85.44	91.78	85.69	25.82	14.60	10.92	17.11
Mask2Former Cheng et al. (2022)	81.42	85.23	92.36	86.34	21.47	9.50	5.46	12.14
MaskMed	**83.68**	**88.12**	**92.83**	**88.21**	**14.75**	**9.42**	**4.39**	**9.52**

Table 4: Official BraTS23 validation leaderboard results using 5-fold cross-validation ensembling.

is computationally prohibitive, restricting global reasoning to low-resolution stages. Deformable attention reduces this cost by attending to sparse sampling locations, but incorporating high-resolution encoder features into cross-scale fusion remains challenging.

Our key observation is that encoder features at different resolutions remain spatially aligned: corresponding locations represent the same anatomical structures at different semantic levels. Based on this observation, we introduce the *Full-Scale Deformable Transformer* (FSD-Transformer), illustrated in Fig. 2(c). Queries are extracted from the lower-resolution encoder stages, while keys and values are constructed from encoder features across all resolutions. Through deformable attention, each query selectively aggregates information from the full encoder hierarchy using sparse sampling locations with query-specific offsets and attention weights predicted by lightweight MLPs. This design enables spatially aligned full-scale feature aggregation, allowing the decoder to combine global semantic context with fine structural details.

3.4 The Proposed Architecture

As illustrated in Fig. 2(a), **MaskMed** follows a hierarchical encoder–decoder architecture with full-scale feature fusion and a unified query-based segmentation head. Multi-scale encoder features are progressively decoded through upsampling blocks and skip connections. To bridge the semantic gap between encoder and decoder representations, the *Full-Scale Deformable Transformer (FSD-Transformer)* performs deformable interaction across the full encoder hierarchy. At the decoder output, the *Masked Multi-Scale Segmentation Head* uses shared queries and hierarchical attention refinement to predict class-agnostic masks and their semantic labels in a decoupled manner, replacing fixed class–channel mappings with flexible mask–class prediction.

Matching and Losses. Following Cheng et al. (2021); Carion et al. (2020), the training objective combines cross-entropy loss for class prediction with binary cross-entropy and Dice losses for mask prediction. To establish a one-to-one correspondence between predictions and ground-truth segments, we perform bipartite matching Cheng et al. (2021); Carion et al. (2020) between predicted mask–class pairs and ground-truth annotations at the finest segmentation stage. The resulting assignment is computed once per forward pass and shared across all decoder stages, providing consistent supervision for multi-scale predictions.

Formally, the predicted output at the finest stage is represented as $z = \{(p_i, m_i^{\text{final}})\}_{i=1}^{N}$, where $p_i \in \Delta^{K+1}$ denotes the class distribution over K semantic classes and an additional *no-object* label $\varnothing$, and $m_i^{\text{final}} \in [0,1]^{D\times H\times W}$ is the predicted binary mask. The ground-truth set is $z^{gt} = \{(y_i, \mathbf{Y}_i)\}_{i=1}^{N^{gt}}$, where $y_i \in \{1,\dots,K\}$ and $\mathbf{Y}_i \in \{0,1\}^{D\times H\times W}$. Since $N \geq N^{gt}$, the ground-truth set is padded with *no-object* labels to enable one-to-one matching.

Given the optimal assignment σ from the finest stage, the multi-scale training loss is

$$\mathcal{L}_{\text{mask-cls}} = \sum_{l=1}^{L} w_l \sum_{j=1}^{N} \Big[-\log p^l_{\sigma(j)}(y_j) + \mathbb{1}_{y_j \neq \varnothing} \mathcal{L}^l_{\text{mask}}\big(m^l_{\sigma(j)},\, \mathbf{Y}^l_j\big)\Big], \tag{6}$$

where $\mathcal{L}^l_{\text{mask}}$ combines Dice and BCE losses at stage l, and $\mathbf{Y}^l_j$ denotes the ground-truth mask resized to the corresponding stage resolution. The weights w_l decrease by a factor of two at each lower-resolution stage, e.g., $w_2 = \frac{1}{2}w_1$ and $w_3 = \frac{1}{4}w_1$, and are normalized to sum to one.

Method	DSC
nnU-Net (fixed CLS Emb)	0.878
nnU-Net w/ Larger Decoder	0.890
MLP-based CLS Emb	0.880
Attention-based CLS Emb	0.885
Attention-based CLS & Mask Emb	0.891
Decouple CLS & Mask Emb	0.903

Table 5: SegHead Variants on AMOS22 Task 1.

Method	DSC
One query per stage	0.903
One query per model (One_Q / M)	0.906
One_Q / M + Masked Attn	0.909
One_Q / M + Masked Attn + StandardTrans	0.901
One_Q / M + Masked Attn + DeformableTrans	0.907
One_Q / M + Masked Attn + FSD-Transformer	0.913

Table 6: Different proposed modules on AMOS22 Task 1.

CNN : Transformer	DSC
1:1	collapse
1:0.5	0.822
1:0.1	0.913
1:0.05	0.908

Table 7: Learning rate ratios.

# Object Query	DSC
$1\times$ N	0.913
$2\times$ N	0.904
$3\times$ N	0.900
$4\times$ N	0.891

Table 8: # Object query.

$\lambda_{\text{Class}} : \lambda_{\text{Mask}_{\text{BCE}}} : \lambda_{\text{Mask}_{\text{Dice}}}$	DSC
$2:5:5$	0.902
$4:5:5$	0.889
$2:10:10$	0.913
$4:10:10$	0.910

Table 9: Loss ratios on AMOS22 Task 1.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate MaskMed on four public 3D medical image segmentation benchmarks: AMOS22 Ji et al. (2022), BTCV Landman et al. (2015), ACDC Bernard et al. (2018), and BraTS23 Kazerooni et al. (2023).

AMOS22. AMOS22 Ji et al. (2022) contains two abdominal multi-organ segmentation tasks. Task 1 is CT-only segmentation with 300 annotated training scans and 200 test scans. Task 2 is joint CT+MRI segmentation, using the same 300 annotated CT scans plus 60 annotated MRI scans for training, and 200 CT plus 40 MRI scans for testing. Both tasks involve 16 abdominal organs and are evaluated using the official leaderboard metrics.

BTCV. BTCV Landman et al. (2015) contains 30 abdominal CT scans with annotations for 13 organs. Following the standard protocol used in prior work Hatamizadeh et al. (2021), we use 24 cases for training and 6 cases for validation. Performance is measured using the average Dice Similarity Coefficient (DSC) across organs.

ACDC. ACDC Bernard et al. (2018) consists of cardiac MRI scans with annotations for the right ventricle (RV), myocardium (MYO), and left ventricle (LV). The dataset provides 200 labeled training cases and 100 testing cases. We report segmentation performance using DSC.

BraTS23. BraTS23 Kazerooni et al. (2023) targets multi-modal brain tumor segmentation from T1, T1c, T2, and FLAIR MRI scans, with annotations for enhancing tumor (ET), tumor core (TC), and whole tumor (WT). It contains 1,291 training scans and 219 validation scans. We evaluate on the official validation set using Dice and 95% Hausdorff distance (HD95) reported on the challenge leaderboard.

4.2 Comparison with State-of-the-Art Methods

Across four benchmarks spanning abdominal CT, multimodal CT/MRI, cardiac MRI, and brain tumor MRI, MaskMed consistently improves over strong CNN-, transformer-, and mask-classification-based baselines. The gains are especially clear on anatomically complex or heterogeneous settings, including small abdominal organs, cardiac structures, tumor subregions, and CT+MRI segmentation, suggesting that decoupled mask–class prediction and full-scale feature fusion improve robustness across structures and modalities.

AMOS22. Tab. 1 reports official leaderboard results on AMOS-CT and AMOS-CTMRI. On AMOS-CT, MaskMed achieves the highest average Dice of **0.913**, outperforming nnU-Net by +2.0 points and 3D Mask2Former by +2.7 points, with the best scores on 12 of 15 organs. On AMOS-CTMRI, MaskMed achieves the top average Dice of **0.893**, surpassing nnU-Net by +1.3 points and 3D Mask2Former by +3.5 points, demonstrating robustness under multimodal CT+MRI segmentation.

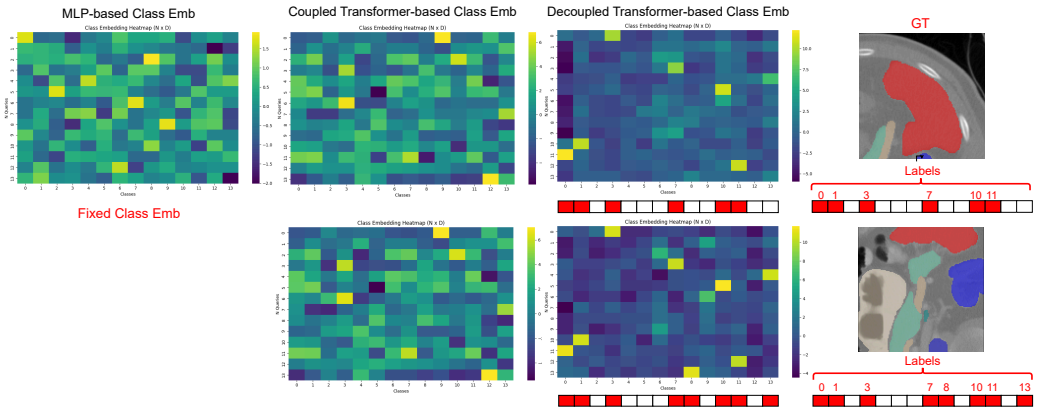


Figure 4: Visualization of class embeddings produced by different segmentation heads.

BTCV. As shown in Tab. 2, MaskMed achieves the best average Dice of **0.872**, outperforming 3D MaskFormer and 3D Mask2Former by +3.7 and +2.3 points, respectively. It achieves the best performance on 5 of 13 organs, including challenging small structures such as the pancreas, indicating robust performance across anatomical scales.

ACDC. Tab. 3 shows that MaskMed achieves the highest average Dice of **93.01%**, surpassing nnU-Net by +1.7 points and nnFormer by +0.6 points. It obtains the best Dice across all three cardiac structures (RV, MYO, and LV), showing consistent performance on cardiac MRI segmentation.

BraTS23. Tab. 4 summarizes results on the official BraTS23 validation leaderboard. MaskMed achieves the highest average Dice of **88.21** and the lowest average HD95 of **9.52**, outperforming nnU-Net and recent transformer-based methods. It attains the best Dice on all three tumor subregions (ET, TC, and WT), demonstrating strong performance on heterogeneous multi-region tumor segmentation.

4.3 Ablation Study on Architecture

Tab. 5 and Tab. 6 isolate the main architectural components of MaskMed. The results show a coherent progression: decoupling mask and class prediction provides the main representational gain, shared multi-scale query propagation improves semantic consistency across decoder stages, and the FSD-Transformer adds full-scale cross-resolution context.

Segmentation Head Variants. Tab. 5 compares different segmentation head designs. Replacing fixed class embeddings with MLP-based projections yields only a marginal improvement (0.878→0.880 DSC), while attention-based class embeddings improve performance to 0.885. Generating both class and mask embeddings with attention further increases DSC to 0.891. The fully decoupled mask–class formulation achieves the best result of 0.903 DSC, confirming that separating spatial mask prediction from semantic classification substantially improves segmentation performance.

Masked Multi-Scale Segmentation Head. As shown in Tab. 6, sharing queries across decoder stages improves DSC from 0.903 to 0.906. Adding inter-stage masked attention further increases performance to 0.909. These gains indicate that propagating shared queries and spatial priors across scales stabilizes query-to-structure correspondence and improves coarse-to-fine segmentation consistency.

Full-Scale Deformable Transformer. Tab. 6 also evaluates encoder–decoder fusion strategies. A standard transformer bridge underperforms (0.901 DSC), likely due to limited scale coverage and high memory cost in volumetric features. Replacing standard attention with deformable attention improves DSC to 0.907. The proposed FSD-Transformer achieves the best performance of 0.913 DSC by enabling efficient deformable interaction across the full encoder hierarchy.

Impact of Model Capacity. To separate architectural gains from parameter scaling, we increase the nnU-Net decoder capacity to match our segmentation head (Tab. 5). This improves DSC only from 0.878 to 0.890, while the decoupled segmentation head reaches 0.903. This suggests that the gains are mainly due to the proposed architecture rather than increased model capacity.

4.4 Ablation on Optimization Sensitivity

We analyze MaskMed's optimization sensitivity with respect to three factors: the learning-rate ratio between CNN and Transformer components, the number of object queries, and the relative weights of class and mask losses. Results are summarized in Tab. 7, Tab. 8, and Tab. 9.

Learning Rate Ratio. As shown in Tab. 7, using identical learning rates for the CNN encoder and Transformer/query modules leads to unstable training, with the $1{:}1$ setting causing divergence. This is expected because convolutional backbones and transformer query modules exhibit different optimization dynamics in 3D volumetric settings, and the query-based components are more sensitive to large updates. The best performance (DSC = 0.913) is achieved when the Transformer learning rate is set to one-tenth of the CNN learning rate ($1{:}0.1$). Further reducing it to $1{:}0.05$ slightly degrades performance to 0.908 DSC, suggesting slower or insufficient optimization.

Number of Object Queries. Unlike MaskFormer and Mask2Former on natural images, which typically over-provision object queries, medical semantic segmentation benefits from a compact query set aligned with the label space. As shown in Tab. 8, using one query per semantic class achieves the best performance (DSC = 0.913), while increasing the number of queries to $2N$, $3N$, or $4N$ progressively degrades performance. This suggests that unnecessary queries introduce assignment ambiguity rather than improving coverage.

Loss Ratio Sensitivity. We further examine the relative weights for classification (λ_{CLS}), binary cross-entropy mask loss (λ_{BCE}), and Dice mask loss (λ_{Dice}). As reported in Tab. 9, the best configuration is $2{:}10{:}10$ (DSC = 0.913). Lower mask-loss weights reduce spatial supervision, while increasing the classification weight degrades performance, indicating that accurate mask learning should dominate optimization while class prediction provides semantic assignment.

4.5 Visualization of Class Embeddings

Fig. 4 visualizes the learned class embeddings for two representative inputs, together with the corresponding ground truth. Since the MLP-based class embeddings become fixed after training, they are shown only once. The x-axis represents class-embedding dimensions, while the y-axis denotes query indices.

The visualization provides qualitative evidence consistent with more structured semantic organization under the decoupled formulation. MLP-based class embeddings appear nearly uniform after training, with moderate activations across most dimensions for each query. In contrast, the proposed Transformer-based embeddings are input-adaptive, and the decoupled formulation produces more localized activation patterns, where specific queries show stronger responses along particular semantic dimensions. As shown in Fig. 4, the prominent activation blocks align with distinct classes, suggesting that decoupled mask–class prediction encourages more structured class embeddings than the MLP-based baseline.

5 Conclusion

We present *MaskMed*, a volumetric medical image segmentation framework that replaces fixed class–channel mappings with decoupled mask–class prediction. MaskMed predicts class-agnostic masks and semantic labels through shared object queries, and combines a *Masked Multi-Scale Segmentation Head* with a *Full-Scale Deformable Transformer* to enable semantically consistent decoder-stage predictions and full-hierarchy cross-scale feature fusion. Extensive experiments demonstrate state-of-the-art performance across AMOS22 Task 1 (91.3% Dice), AMOS22 Task 2 (89.3% Dice), ACDC (93.0% Dice), BTCV (87.2% Dice), and BraTS23 (88.21% Dice), validating the effectiveness of decoupled mask–class prediction and full-scale deformable feature aggregation.

While effective, MaskMed introduces additional architectural complexity compared with simpler nnU-Net-style baselines and shows some sensitivity to learning-rate and loss-weight settings. The current evaluation is also limited to public benchmarks; prospective or external clinical validation remains important future work. Finally, query-based decoupling under incomplete, hierarchical, or institution-specific label sets remains an open direction.

Reproducibility Statement

All datasets used in this work (AMOS22, BTCV, ACDC) are publicly available. Training and inference hyper-parameters, the deep-supervision objective, the data augmentation pipeline and the hardware configuration are given in Appendix F; the architecture of MaskMed is specified in Appendix E. Code will be released upon acceptance.

Use of Large Language Models

Large language models were used only for language editing of author-written text. They were not used for research idea, experimental design, implementation, data analysis, or the generation of any reported result. The authors take full responsibility for all content in this paper.

References

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, volume 12346 of *Lecture Notes in Computer Science*, pp. 213–229. Springer, 2020.

Junyu Chen, Eric C. Frey, Yufan He, William P. Segars, Ye Li, and Yong Du. Transmorph: Transformer for unsupervised medical image registration. *Medical Image Analysis*, 82:102615, 2022.

Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17864–17875, 2021.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1290–1299, 2022.

Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 764–773, 2017.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pp. 272–284. Springer, 2021.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 574–584, 2022.

Yufan He, Vishwesh Nath, Dong Yang, Yucheng Tang, Andriy Myronenko, and Daguang Xu. Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14223 of *Lecture Notes in Computer Science*, pp. 416–426. Springer, 2023.

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.

Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in Neural Information Processing Systems*, 35: 36722–36732, 2022.

Anahita Fathi Kazerooni, Nastaran Khalili, Xinyang Liu, Debanjan Haldar, Zhifan Jiang, Syed Muhammed Anwar, Jake Albrecht, Maruf Adewole, Udunna Anazodo, Hannah Anderson, et al. The brain tumor segmentation (brats) challenge 2023: Focus on pediatrics (cbtn-connect-dipgr-asnr-miccai brats-peds). *arXiv preprint arXiv:2305.17033*, 2023.

B. Landman, Z. Xu, J. E. Iglesias, M. Styner, T. Langerak, and A. Klein. Miccai multi-atlas labeling beyond the cranial vault – workshop and challenge. In *Proceedings of MICCAI Multi-Atlas Labeling Beyond Cranial Vault – Workshop and Challenge*, 2015.

Ho Hin Lee, Shunxing Bao, Yuankai Huo, and Bennett A. Landman. 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. *arXiv preprint arXiv:2209.15076*, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241. Springer, 2015.

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 (CVPR)*, pp. 20730–20740, 2022.

Bin Xie, Hao Tang, Bin Duan, Dawen Cai, Yan Yan, and Gady Agam. Masksam: Auto-prompt sam with mask classification for volumetric medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24423–24433, 2025.

Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE Transactions on Image Processing*, 2023.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2021.

Table 10: Efficiency comparison of different 3D medical image segmentation models. We report the number of parameters, FLOPs, inference latency (averaged over 100 runs), peak GPU memory, and DSC.

Model	Params (M)	FLOPs (G)	Latency (ms)	Peak Mem (MB)	DSC ↑
UNETR	92.81	81.53	28.99	920.59	0.807
nnFormer	150.69	105.58	25.10	809.99	0.845
SwinUNETR	62.19	319.97	73.86	1572.25	0.884
3D UX-Net	53.01	579.64	72.32	1314.50	0.889
SwinUNETRv2	72.76	345.83	76.58	1615.14	0.897
nnUNet	102.36	304.17	32.13	1225.39	0.893
MaskMed w/o FSAD	122.49	328.47	67.38	1345.27	0.909
MaskMed (Ours)	**124.99**	**498.24**	**182.04**	**4946.58**	**0.913**

A Efficiency Analysis.

Table 10 below reports parameters, FLOPs, latency (100 runs), and peak memory. Starting from the nnUNet baseline, MaskMed introduces two components: (1) the unified Masked Multi-Scale Segmentation Head and (2) the FSAD-Transformer. Replacing the conventional segmentation head with the unified Masked Multi-Scale Segmentation Head increases parameters from 102.36M to 122.49M (+20.1M) while improving DSC from 0.893 to 0.909 (+1.6%). Adding the FSAD-Transformer further introduces only 2.5M parameters (122.49M $\rightarrow$ 124.99M) but increases FLOPs, latency, and peak memory because it operates on full-resolution 3D features. This yields an additional 0.4% DSC improvement (0.909 $\rightarrow$ 0.913). Although our model is computationally more expensive than nnUNet, the added cost primarily stems from processing high-resolution volumetric features, while achieving state-of-the-art segmentation accuracy.

B Statistical Significance

These results demonstrate that MaskMed provides robust and statistically validated improvements over the nnUNet baseline.

We performed paired statistical significance tests between **MaskMed** and **nnUNet** using the Wilcoxon signed-rank test. The 95% confidence intervals (CI) were estimated for Dice scores and performance differences. Effect sizes were measured using rank-biserial correlation (r).

For **ACDC**, **AMOS22-Task1**, **AMOS22-Task2**, and **BraTS23**, the official test datasets are not publicly accessible. Therefore, we performed statistical significance analysis on the validation datasets using the 5-fold cross-validation protocol. Specifically, predictions from all validation cases across the five folds were aggregated and paired between MaskMed and nnUNet. Since each case is evaluated once as a validation sample across the five folds, the total number of evaluated cases is equivalent to the full dataset size.

For **BTCV**, due to the limited dataset size, we evaluated model performance using three independent random seeds. Paired statistical significance tests were performed separately for each seed between MaskMed and nnUNet, resulting in a total of 3×6 cases for statistical comparison.

B.0.1 Overall Dice Comparison

Table 11: Overall Dice comparison between MaskMed and nnUNet across five datasets.

Dataset	#Cases	MaskMed Dice (95% CI)	nnUNet Dice (95% CI)	Improvement	Wilcoxon p-value	Effect size (r)
BraTS23	1251	0.9127 [0.9068, 0.9185]	0.9108 [0.9048, 0.9166]	+0.0019 [0.0001, 0.0040]	0.000854	0.109
AMOS22-Task1	300	0.8997 [0.8939, 0.9052]	0.8856 [0.8793, 0.8916]	+0.0141 [0.0122, 0.0160]	2.496×10^{-43}	0.919
AMOS22-Task2	360	0.8923 [0.8855, 0.8984]	0.8707 [0.8629, 0.8779]	+0.0216 [0.0195, 0.0238]	5.13×10^{-56}	0.959
BTCV	3×6	0.8720 [0.8558, 0.8863]	0.8625 [0.8482, 0.8752]	+0.0094 [0.0061, 0.0129]	0.000145	0.918
ACDC	200	0.9286 [0.9243, 0.9326]	0.9114 [0.9053, 0.9169]	+0.0172 [0.0140, 0.0210]	4.781×10^{-26}	0.861

B.0.2 Overall Analysis

Across all five datasets, MaskMed consistently achieves statistically significant improvements over the nnUNet baseline.

- **AMOS22 datasets demonstrate the largest improvements**, with Dice gains of:
  - +1.41% on AMOS22-Task1
  - +2.16% on AMOS22-Task2

  accompanied by very large effect sizes ($r = 0.919$ and $r = 0.959$), indicating strong practical significance.
- On **ACDC**, MaskMed improves overall Dice by **1.72 percentage points** with a large effect size ($r = 0.861$), demonstrating improved cardiac structure segmentation.
- On **BTCV**, despite the limited sample size (18 cases), MaskMed still achieves statistically significant improvement ($p = 0.000145$), indicating robustness under small-scale evaluation.
- On **BraTS23**, where nnUNet already achieves competitive performance, MaskMed further improves Dice by 0.19 percentage points with statistical significance ($p = 0.000854$).

Overall, these results demonstrate that MaskMed provides consistent and statistically validated improvements across diverse medical image segmentation tasks.

B.1 Per-class Statistical Significance Analysis

For class-level comparisons, Holm correction was applied to control the family-wise error rate caused by multiple comparisons. A class is considered statistically significant when:

$$p_{\text{Holm}} < 0.05$$

B.1.1 BraTS23

Table 12: Per-class results on BraTS23.

Class	#Cases	MaskMed Dice (95% CI)	nnUNet Dice (95% CI)	Improvement	Wilcoxon p-value	Holm p-value	Effect size (r)
WT	1251	0.9382 [0.9339, 0.9422]	0.9379 [0.9337, 0.9419]	+0.0002 [-0.0017, 0.0020]	0.001746	0.003491	0.102
TC	1251	0.9197 [0.9117, 0.9274]	0.9181 [0.9100, 0.9260]	+0.0016 [-0.0012, 0.0043]	0.0001321	0.0003963	0.125
ET	1251	0.8803 [0.8701, 0.8899]	0.8762 [0.8656, 0.8861]	+0.0041 [0.0000, 0.0089]	0.002696	0.003491	0.099

Analysis. MaskMed achieves statistically significant improvements on all three tumor regions after Holm correction. Although the absolute gains are relatively small due to the strong nnUNet baseline, MaskMed consistently improves WT, TC, and ET segmentation.

B.1.2 AMOS22-Task1

Table 13: Per-class results on AMOS22-Task1.

Class	#Cases	MaskMed Dice (95% CI)	nnUNet Dice (95% CI)	Improvement	Wilcoxon p-value	Holm p-value	Effect size (r)
Spleen	300	0.9656 [0.9546, 0.9735]	0.9581 [0.9470, 0.9665]	+0.0074 [0.0042, 0.0106]	8.543e-31	1.111e-29	0.771
Right kidney	300	0.9646 [0.9603, 0.9681]	0.9522 [0.9449, 0.9582]	+0.0125 [0.0079, 0.0175]	2.678e-16	1.071e-15	0.547
Left kidney	300	0.9591 [0.9499, 0.9665]	0.9437 [0.9303, 0.9550]	+0.0154 [0.0075, 0.0251]	1.182e-13	1.182e-13	0.494
Gallbladder	300	0.8589 [0.8329, 0.8823]	0.8410 [0.8146, 0.8657]	+0.0179 [0.0042, 0.0311]	3.114e-20	2.803e-19	0.630
Esophagus	300	0.8635 [0.8537, 0.8727]	0.8463 [0.8355, 0.8567]	+0.0172 [0.0129, 0.0216]	4.91e-19	2.946e-18	0.595
Liver	300	0.9760 [0.9723, 0.9791]	0.9712 [0.9670, 0.9748]	+0.0048 [0.0027, 0.0071]	1.598e-38	2.397e-37	0.865
Stomach	300	0.9229 [0.9050, 0.9382]	0.9050 [0.8860, 0.9218]	+0.0178 [0.0129, 0.0234]	9.014e-35	1.262e-33	0.828
Aorta	300	0.9568 [0.9530, 0.9601]	0.9531 [0.9484, 0.9569]	+0.0037 [0.0024, 0.0052]	3.613e-15	1.084e-14	0.524
Postcava	300	0.9147 [0.9087, 0.9202]	0.9083 [0.9020, 0.9142]	+0.0065 [0.0048, 0.0082]	5.158e-17	2.579e-16	0.558
Pancreas	300	0.8742 [0.8622, 0.8847]	0.8652 [0.8529, 0.8760]	+0.0090 [0.0062, 0.0121]	7.587e-24	7.587e-23	0.671
Right adrenal	300	0.7978 [0.7833, 0.8109]	0.7854 [0.7716, 0.7980]	+0.0124 [0.0062, 0.0176]	8.781e-15	1.756e-14	0.518
Left adrenal	300	0.8070 [0.7917, 0.8209]	0.7903 [0.7745, 0.8050]	+0.0166 [0.0125, 0.0210]	2.248e-19	1.574e-18	0.602
Duodenum	300	0.8411 [0.8263, 0.8542]	0.8193 [0.8040, 0.8334]	+0.0217 [0.0176, 0.0260]	6.17e-29	7.404e-28	0.745
Bladder	300	0.9225 [0.9092, 0.9342]	0.8978 [0.8791, 0.9145]	+0.0247 [0.0167, 0.0341]	7.436e-28	8.179e-27	0.736
Prostate/uterus	300	0.8705 [0.8503, 0.8886]	0.8469 [0.8261, 0.8666]	+0.0235 [0.0160, 0.0318]	3.311e-20	2.803e-19	0.626

Analysis. All 15 anatomical structures show statistically significant improvements after Holm correction. The largest improvements are observed on difficult organs with complex boundaries, including bladder (+2.47%), prostate/uterus (+2.35%), and duodenum (+2.17%).

B.1.3 AMOS22-Task2

Analysis. All 15 anatomical structures achieve statistically significant improvements. MaskMed provides particularly strong improvements for challenging structures, including gallbladder (+3.91%), duodenum (+3.70%), bladder (+3.61%), and prostate/uterus (+2.97%).

Table 14: Per-class results on AMOS22-Task2.

Class	#Cases	MaskMed Dice (95% CI)	nnUNet Dice (95% CI)	Improvement	Wilcoxon p-value	Holm p-value	Effect size (r)
Spleen	360	0.9651 [0.9558, 0.9724]	0.9549 [0.9454, 0.9628]	+0.0102 [0.0066, 0.0139]	9.518e-49	1.333e-47	0.896
Right kidney	360	0.9590 [0.9500, 0.9658]	0.9450 [0.9356, 0.9532]	+0.0140 [0.0089, 0.0197]	2.585e-37	2.068e-36	0.780
Left kidney	360	0.9569 [0.9474, 0.9646]	0.9401 [0.9281, 0.9503]	+0.0168 [0.0099, 0.0246]	4.752e-38	4.752e-37	0.785
Gallbladder	360	0.8531 [0.8288, 0.8752]	0.8139 [0.7873, 0.8390]	+0.0391 [0.0235, 0.0550]	2.157e-34	1.294e-33	0.765
Esophagus	360	0.8362 [0.8226, 0.8484]	0.8086 [0.7944, 0.8217]	+0.0276 [0.0235, 0.0321]	2.753e-35	1.927e-34	0.756
Liver	360	0.9755 [0.9711, 0.9787]	0.9683 [0.9639, 0.9721]	+0.0071 [0.0045, 0.0098]	1.872e-52	2.809e-51	0.927
Stomach	360	0.9150 [0.8983, 0.9296]	0.8916 [0.8738, 0.9073]	+0.0234 [0.0181, 0.0294]	1.273e-43	1.406e-42	0.847
Aorta	360	0.9493 [0.9421, 0.9548]	0.9437 [0.9371, 0.9488]	+0.0056 [0.0041, 0.0071]	1.223e-29	2.446e-29	0.688
Postcava	360	0.9077 [0.8998, 0.9144]	0.8967 [0.8888, 0.9037]	+0.0110 [0.0091, 0.0129]	5.598e-38	5.038e-37	0.784
Pancreas	360	0.8742 [0.8626, 0.8845]	0.8555 [0.8429, 0.8669]	+0.0187 [0.0159, 0.0219]	9.16e-47	1.191e-45	0.875
Right adrenal	360	0.7724 [0.7571, 0.7865]	0.7501 [0.7350, 0.7640]	+0.0223 [0.0179, 0.0265]	5.033e-32	1.51e-31	0.719
Left adrenal	360	0.7835 [0.7664, 0.7992]	0.7588 [0.7416, 0.7747]	+0.0248 [0.0191, 0.0305]	3.463e-33	1.385e-32	0.736
Duodenum	360	0.8235 [0.8092, 0.8366]	0.7865 [0.7712, 0.8007]	+0.0370 [0.0321, 0.0422]	1.172e-43	1.406e-42	0.845
Bladder	360	0.9260 [0.9116, 0.9390]	0.8899 [0.8691, 0.9093]	+0.0361 [0.0243, 0.0501]	1.422e-33	7.112e-33	0.806
Prostate/uterus	360	0.8872 [0.8697, 0.9030]	0.8575 [0.8375, 0.8758]	+0.0297 [0.0212, 0.0397]	8.535e-26	8.535e-26	0.710

Table 15: Per-class results on BTCV.

Class	#Cases	MaskMed Dice (95% CI)	nnUNet Dice (95% CI)	Improvement	Wilcoxon p-value	Holm p-value	Effect size (r)
Spleen	18	0.9701 [0.9678, 0.9723]	0.9651 [0.9610, 0.9687]	+0.50%	7.629e-06	9.155e-05	1.000
Right kidney	18	0.9520 [0.9469, 0.9574]	0.9509 [0.9461, 0.9560]	+0.10%	2.645e-01	5.661e-01	0.310
Left kidney	18	0.9560 [0.9518, 0.9603]	0.9517 [0.9470, 0.9566]	+0.43%	7.629e-06	9.155e-05	1.000
Gallbladder	18	0.6805 [0.5214, 0.8132]	0.6610 [0.5021, 0.7966]	+1.95%	7.058e-02	4.235e-01	0.485
Esophagus	18	0.7916 [0.7627, 0.8196]	0.7925 [0.7716, 0.8151]	-0.09%	8.986e-01	8.986e-01	0.041
Liver	18	0.9743 [0.9725, 0.9759]	0.9723 [0.9707, 0.9739]	+0.20%	1.526e-05	1.373e-04	0.988
Stomach	18	0.9173 [0.9008, 0.9308]	0.9135 [0.9019, 0.9250]	+0.39%	8.143e-02	4.235e-01	0.474
Aorta	18	0.9157 [0.9114, 0.9201]	0.9058 [0.8950, 0.9158]	+0.98%	5.386e-02	3.770e-01	0.520
Inferior vena cava	18	0.8876 [0.8730, 0.9005]	0.8679 [0.8530, 0.8815]	+1.97%	7.629e-06	9.155e-05	1.000
Portal vein and splenic vein	18	0.7999 [0.7897, 0.8111]	0.7903 [0.7801, 0.8018]	+0.96%	1.415e-01	5.661e-01	0.404
Pancreas	18	0.8769 [0.8665, 0.8882]	0.8567 [0.8438, 0.8694]	+2.03%	7.690e-03	6.152e-02	0.696
Adrenal gland	18	0.7418 [0.7328, 0.7511]	0.7225 [0.7012, 0.7417]	+1.93%	1.415e-01	5.661e-01	0.404

B.1.4 BTCV

Analysis. Due to the limited number of BTCV cases (18), statistical power is reduced. After Holm correction, four classes remain significant: spleen, left kidney, liver, and inferior vena cava. Nevertheless, MaskMed achieves consistent positive improvements across most structures.

B.1.5 ACDC

Table 16: Per-class results on ACDC.

Class	#Cases	MaskMed Dice (95% CI)	nnUNet Dice (95% CI)	Improvement	Wilcoxon p-value	Holm p-value	Effect size (r)
RV (Right Ventricle)	200	0.9203 [0.9122, 0.9276]	0.8913 [0.8799, 0.9016]	+2.89%	1.114e-18	2.228e-18	0.720
MYO (Myocardium)	200	0.9136 [0.9098, 0.9172]	0.8986 [0.8938, 0.9031]	+1.50%	1.112e-30	3.336e-30	0.939
LV (Left Ventricle)	200	0.9519 [0.9459, 0.9575]	0.9443 [0.9374, 0.9508]	+0.76%	5.925e-17	5.925e-17	0.682

Analysis. All three cardiac structures achieve statistically significant improvements after Holm correction. The largest improvement is observed for the right ventricle (+2.89%), demonstrating stronger capability for challenging cardiac segmentation.

B.2 Conclusion

Across five datasets, including BraTS23, AMOS22, BTCV, and ACDC, MaskMed consistently outperforms nnUNet.

- All datasets achieve statistically significant overall improvements.
- Improvements remain significant across the majority of anatomical structures after multiple-comparison correction.
- The strongest gains are observed on challenging organs with small size, irregular shape, or complex boundaries.

C Visualization of Mask and Class Embeddings

Figures 5 and 6 visualize class embeddings and mask embeddings for two representative inputs, shown in the first and second rows, respectively. The rightmost column shows the corresponding ground truth (GT). For clarity, we visualize a single slice together with its labels. Since the MLP-based class and mask embeddings become fixed after training, they are shown only once. In Fig. 5,



Figure 5: Visualization of class embeddings.

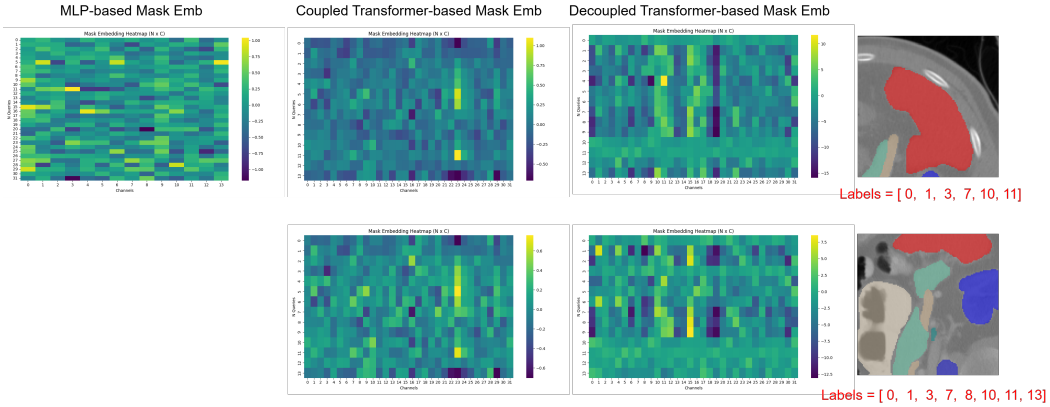


Figure 6: Visualization of mask embeddings.

the x-axis represents class-embedding dimensions, while in Fig. 6 it represents mask-embedding channels; in both figures, the y-axis denotes query indices.

C.1 Visualization of Class Embeddings

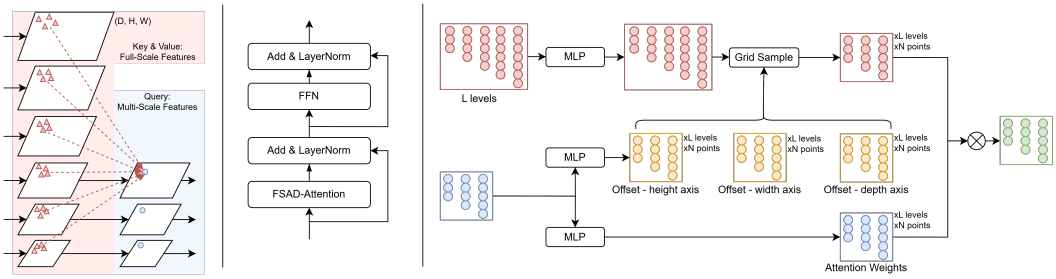
Fig. 5 provides qualitative evidence that the proposed decoupled formulation yields more structured class embeddings. For the MLP-based baseline, the learned embeddings become nearly uniform after training, and each query exhibits moderate activations across most dimensions. In contrast, the Transformer-based class embeddings are input-adaptive and show more localized activation patterns. Under the decoupled formulation, specific queries exhibit stronger responses along particular semantic dimensions, suggesting more structured semantic organization than the MLP-based baseline.

C.2 Visualization of Mask Embeddings

Fig. 6 visualizes the learned mask embeddings. The proposed decoupled Transformer-based mask embeddings exhibit diverse query-specific activation patterns across channels, indicating more adaptive spatial representations. This contrasts with conventional point-wise convolution heads, which rely on fixed filters and provide less flexibility in modeling spatial variation.

D Details of the Full-Scale Deformable Transformer

The proposed *Full-Scale Deformable Transformer* (FSD-Transformer) follows the structure of a standard Transformer block, but replaces self-attention with the proposed full-scale deformable attention module. Full-scale features are first processed by the deformable attention module, followed by residual addition and LayerNorm. The output is then passed through a feed-forward network (FFN), followed again by residual addition and LayerNorm.



(a) FSAD-Transformer (b) Details of FSAD-Transformer (c) Detals of FSAD-Attention

Figure 7: Details of the proposed FSD-Transformer.

Full-Scale Deformable Attention. In the deformable attention module, full-scale encoder features are used as value features, while the lowest three scales are used to form the query/key representation. Lightweight MLPs predict sampling offsets and attention weights from the query features. The value features are projected, sampled at the learned offsets via grid sampling, and aggregated using the predicted attention weights to produce the final output. This design enables efficient cross-scale feature interaction while preserving spatial alignment.

E Analysis of MaskMed

E.1 Comparison with MaskFormer and Mask2Former

In this section, we compare MaskMed with MaskFormer and Mask2Former. Fig. 8 summarizes the main architectural differences.

UNet backbone and stable 3D optimization. Directly extending Mask2Former to 3D medical segmentation often leads to unstable optimization and training collapse, partly due to the mismatch between CNN-based encoders and transformer-based decoders in volumetric settings. In addition, MaskFormer and Mask2Former rely on a Feature Pyramid Network (FPN) for progressive upsampling, whereas UNet-style architectures remain the dominant design in medical image segmentation. We therefore build MaskMed on a UNet backbone and introduce a decoupled segmentation head to predict mask and class embeddings separately. To further stabilize training, we use different learning rates for CNN-based modules (the UNet backbone) and Transformer-based modules (the segmentation head and FSD-Transformer).

Multi-scale interaction between class and mask embeddings. In MaskFormer and Mask2Former, mask predictions are produced only at the final resolution: the highest-resolution FPN features are combined with mask embeddings generated by different layers of the transformer decoder. Both class and mask embeddings are derived primarily from low-resolution features and the initial query set, creating a relatively long semantic path between feature extraction and final prediction. In contrast, MaskMed introduces stage-wise interaction between shared queries and multi-scale decoder features from the UNet. This shortens the semantic path and improves alignment between class and mask representations during coarse-to-fine refinement.

Full-scale feature aggregation via FSD-Transformer. MaskMed introduces the *Full-Scale Deformable Transformer* (FSD-Transformer), which aggregates skip-connected features from all encoder stages within a unified Transformer block. This design enables full-scale cross-resolution fusion and strengthens long-range contextual reasoning. By comparison, Mask2Former employs deformable attention over only the last few encoder stages and does not perform full-hierarchy feature aggregation.

3D design for volumetric medical segmentation. All components of MaskMed are implemented in 3D and designed specifically for volumetric medical image segmentation. In contrast, MaskFormer and Mask2Former were originally developed for 2D natural image segmentation, making a direct transfer to 3D medical settings less straightforward.

MaskMed is inspired by Mask2Former and DETR, but it is not a direct adaptation. To the best of our knowledge, prior work has not established a strong decoupled mask–class prediction framework for

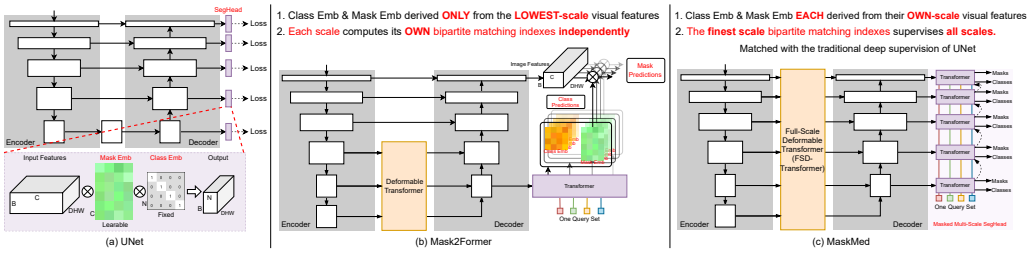


Figure 8: Comparison of MaskMed with MaskFormer and Mask2Former.

3D medical image segmentation with competitive performance. Our results show that this paradigm can be made effective in volumetric medical settings when combined with UNet-style multi-scale decoding, shared finest-scale supervision, and full-scale deformable feature fusion.

More broadly, we believe MaskMed highlights a promising direction for 3D medical image segmentation. Beyond the improvements demonstrated here, it also raises several interesting questions for future work, including how decoupled query-based prediction behaves under more complex label structures and broader clinical settings.

E.2 Necessity of the Decoupled Framework

The proposed decoupled framework is useful for medical image segmentation for three main reasons.

Flexibility. By separating mask prediction from semantic classification, MaskMed can model spatial regions and class labels independently. This differs from conventional softmax-based segmentation heads, where all categories are predicted through fixed output channels. The decoupled formulation allows the model to adaptively associate predicted regions with semantic classes, making it more flexible across datasets, modalities, and label configurations.

Interpretability. The decoupled design also provides a more explicit separation between semantic and spatial representations. As shown in the supplementary visualizations, class embeddings exhibit structured semantic activation patterns, while mask embeddings capture query-specific spatial patterns. Although these visualizations are qualitative, they suggest that decoupled mask–class prediction offers a useful way to inspect how the model organizes semantic and spatial information.

Performance. The empirical results show that the decoupled formulation is not only conceptually motivated but also practically effective. Combined with the Masked Multi-Scale Segmentation Head and FSD-Transformer, MaskMed consistently achieves strong performance across multiple benchmarks, supporting the value of decoupled prediction for volumetric medical segmentation.

Overall, MaskMed provides a flexible and high-performing alternative to fixed channel-based segmentation heads, while also offering a more structured representation of semantic and spatial prediction.

F Implementation Details

We build on the nnU-Net training framework and modify only the network architecture, keeping the remaining training configuration and data augmentation strategy consistent with nnU-Net. The initial learning rate is set to 0.001, with polynomial decay:

$$lr_i(e) = \lambda_i \cdot lr_{\text{init}} \cdot \left(1 - \frac{e}{E_{\max}}\right)^{0.9}, \quad i \in \{\text{CNN}, \text{Transformer}\}, \tag{7}$$

where e denotes the current epoch and $E_{\max} = 1000$. Each epoch contains 250 iterations. The coefficient λ_i controls the learning-rate ratio between the CNN and Transformer modules. We set $\lambda_{\text{CNN}} = 1.0$ and $\lambda_{\text{Transformer}} = 0.1$ in all experiments. We use SGD with momentum 0.99 and weight decay 3×10^{-5}. The training loss combines cross-entropy, Dice, and binary cross-entropy terms as described in Sec. 3.