Learning Generalizable 3D Medical Image Representations from Mask-Guided Self-Supervision

Supplementary Material

6. Supplementary Method

6.1. Annotation-Free Mask Generation with SAM2

Pipeline Overview. We adapt SAM2 [44], a foundation model trained on natural images, to generate class-agnostic 3D masks from medical volumes without any human annotation. Our pipeline consists of five stages: (1) *Multi-channel preprocessing* converts single-channel medical volumes into 3-channel representations using complementary intensity mappings to preserve tissue contrasts compatible with SAM2's RGB input expectations, (2) *Slice sampling with automatic axis selection* identifies the highest-resolution imaging axis and samples a sparse set of 2D slices based on physical spacing to ensure consistent coverage, (3) *2D mask generation* applies SAM2's automatic segmentation mode with dense point prompts to systematically explore each slice, prioritizing diversity and coverage to generate hundreds of candidate masks per slice spanning multiple granularities, (4) *3D propagation* extends 2D masks into volumetric segmentations using SAM2's video tracking capability, and (5) *Post-processing* refines masks through connected component filtering, removes redundant overlapping masks. The entire pipeline operates at original image resolution, generating hundreds to thousands diverse masks per volume that serve as structural proposals for representation learning.

Multi-Channel Preprocessing. SAM2 expects 3-channel RGB images with pixel values in [0, 255], while medical images are single-channel grayscale volumes with modality-specific intensity ranges. For CT images with Hounsfield Units spanning -1024 to 3071, we create three channels using complementary intensity windows that capture different tissue contrasts. For abdomen CT, we apply: (1) soft tissue window (center=60 HU, width=350 HU), (2) contrast window (center=15 HU, width=250 HU), and (3) bone window (center=150 HU, width=1200 HU). For MRI and PET images, we apply quantile-based normalization with three percentile ranges: (5-95%), (15-85%), and (1-99%). Each intensity mapping forms one channel of the final RGB representation.

We evaluated three preprocessing strategies (Table 5): (1) *Linear mapping*: map a fixed intensity range (e.g., [-400, 1500] HU for CT) to [0, 255] and replicate across three channels; (2) *Single window*: apply one clinical window (e.g., soft tissue window) and replicate across three channels; (3) *Three-channel windows*: create three distinct intensity mappings targeting different tissues. The three-channel approach (65.5% Dice) substantially outperforms linear mapping (54.7%) and single-window (64.5%) strategies. This is because complementary tissue contrasts in each channel enable SAM2 to detect boundaries across diverse anatomical structures—tissues invisible in one window may have clear boundaries in another. We adopt three-channel windows as our default strategy, though users can tune specific window parameters to optimize SAM2's sensitivity for their target regions of interest.

Table 5. Ablation on SAM2 preprocessing strategies. Evaluated on BCV 1-shot segmentation.

Preprocessing Method	Dice (%)
Linear mapping	54.7
Single window replication	64.5
Three-channel windows	**65.5**
Three-channel + multi-axis	65.4

Slice Sampling with Automatic Axis Selection. Medical images typically exhibit anisotropic resolution with high in-plane resolution but coarse through-plane spacing. We automatically select the imaging axis with highest in-plane resolution for mask generation by computing average in-plane resolutions for each axis and selecting the minimum. For approximately isotropic images (spacing ratios ≤ 1.3), we default to the first axis. This optimizes mask quality while maintaining computational efficiency.

Along the selected axis, we sample 2D slices using physical distance intervals rather than fixed slice counts, ensuring consistent spatial coverage across images with different resolutions. Given axis spacing s_{axis} and physical interval d_{mm} (default: 15mm), the slice interval is $\max(1, \lfloor d_{\text{mm}}/s_{\text{axis}} \rfloor)$. This strategy balances comprehensive coverage with computational efficiency.

We also evaluate multi-axis processing, where masks are generated independently along all three anatomical axes (axial, sagittal, coronal) and merged (Table 5). For the BCV dataset with anisotropic resolution (e.g. in-plane: 0.7×0.7mm, through-plane: 5-10mm), multi-axis processing yields 65.4% performance—marginally lower than single-axis 65.5% despite 3× computational cost. This occurs because processing lower-resolution planes (sagittal/coronal) introduces interpolation artifacts that produce degraded masks, which dilute rather than enhance the mask set quality. Based on these results, we default to single-axis processing on the highest-resolution plane for all experi-

ments. However, for images with approximately isotropic resolution, where all axes have comparable quality, multi-axis processing may provide complementary structural information and improve performance.

2D Mask Generation. For each sampled slice, we apply SAM2's automatic mask generation mode, which densely samples point prompts across the image grid and generates masks for each prompt without requiring any human annotation. This automatic mode systematically explores the entire image space to discover diverse segmentation proposals. We configure SAM2 with key parameters including points per side (typically 32), prediction IoU threshold (0.3-0.4), and stability score threshold (0.5-0.7) to control mask quality and quantity. Generated masks are filtered by minimum area to remove trivially small regions. We limit the number of masks retained per slice (typically 70) to control computation while maintaining diversity across granularities. Importantly, we prioritize diversity and coverage over individual mask accuracy—our goal is to generate masks spanning as many different anatomical and pathological structures and scales as possible, from large regions to fine-grained structures, rather than obtaining perfectly accurate segmentations. This stage generates hundreds to thousands of candidate masks per volume, providing rich structural proposals for representation learning.

3D Propagation via Video Tracking. To extend 2D masks into 3D volumes, we leverage SAM2's video prediction capability. We save the multi-channel slice sequence as JPEG frames and initialize SAM2's video predictor with this sequence. For each mask generated on a seed slice, we add it as a tracked object and perform bidirectional propagation: forward from the seed slice to the volume end, then backward from the seed slice to the volume start. This bidirectional approach ensures complete volumetric coverage while maintaining temporal consistency through SAM2's memory attention mechanism. The propagation produces per-object 3D mask volumes at the original image resolution without any resampling.

Post-Processing Each propagated 3D mask undergoes refinement: (1) *Connected component filtering*: We identify all connected components and retain only the largest component that intersects with the original seed slice mask, removing spurious disconnected regions introduced during propagation. (2) *Minimum volume filtering*: Masks smaller than a threshold are discarded to remove noise while preserving small but meaningful structures. (3) *Redundancy reduction*: Since masks are propagated from different seed slices, the same anatomical structure may be captured multiple times with high overlap. To reduce redundancy while maintaining diversity, we identify pairs of masks with $>$90% overlap (IoU > 0.9) and randomly discard one from each pair.

Mask Statistics Our pipeline generates hundreds to thousands of masks per volume depending on anatomical complexity and dataset configuration. For the BCV dataset (abdominal CT scans), we generate an average of 1462 masks per scan, with minimum 938 and maximum 2210.

Computational Cost. Mask generation is performed offline once per dataset. Typical processing time for a single CT volume (512×512×150) with our default parameters: (1) Preprocessing: 10-20 seconds; (2) 2D mask generation (10-15 slices): 2-3 minutes; (3) 3D propagation: 3-5 minutes; (4)Post-processing: 30-60 seconds; Total: 6-10 minutes per volume on a single NVIDIA H100 GPU. Once generated, masks are reused across all pretraining experiments, amortizing this one-time cost. The annotation-free mask generation pipeline can be parallelized across volumes, enabling efficient scaling to large datasets.

6.2. Visualization of SAM2 generated masks

SAM2 enables fully automatic mask generation without any human annotation, providing the structural supervision needed for MASS pretraining. Figure 3 shows 2D masks generated on seed slices for abdomen CT, head & neck CT, and abdomen MR. The class-agnostic masks demonstrate excellent coverage and diversity, capturing anatomical structures at multiple granularities—from large organs (liver, kidneys) and skeletal structures (ribs, mandible, vertebrae) to finer sub-organ regions and small pathological findings (renal cysts). This diversity is critical for MASS: by training on masks spanning different anatomical scales and tissue types, the model learns compositional visual primitives applicable to novel structures. While the masks contain noise from over-segmentation and miss some small subtle structures (e.g., tiny vessels, thin organ boundaries), they provide sufficient structural information for learning generalizable representations.

Figure 4 demonstrates the 3D propagation results using SAM2's video prediction capability. The same cases from Figure 3 show how 2D seed masks are extended into volumetric segmentations by tracking boundaries through adjacent slices. Bidirectional propagation (forward and backward from each seed slice) ensures complete volumetric coverage. While propagation successfully maintains anatomical coherence for most structures, it introduces additional noise compared to the 2D masks. This occurs because SAM2's video tracking was trained on natural videos where occlusion handling is critical. When objects move behind others in natural scenes, the tracker maintains identity across the occlusion. However, in medical imaging, this behavior can inappropriately merge anatomically separate structures that happen to be adjacent. For example, in the second column of the first row in Figure 4, we observe the heart and liver connected as a single mask despite being distinct organs. Despite these imperfections, the propagated 3D masks provide volumetrically consistent structural su-

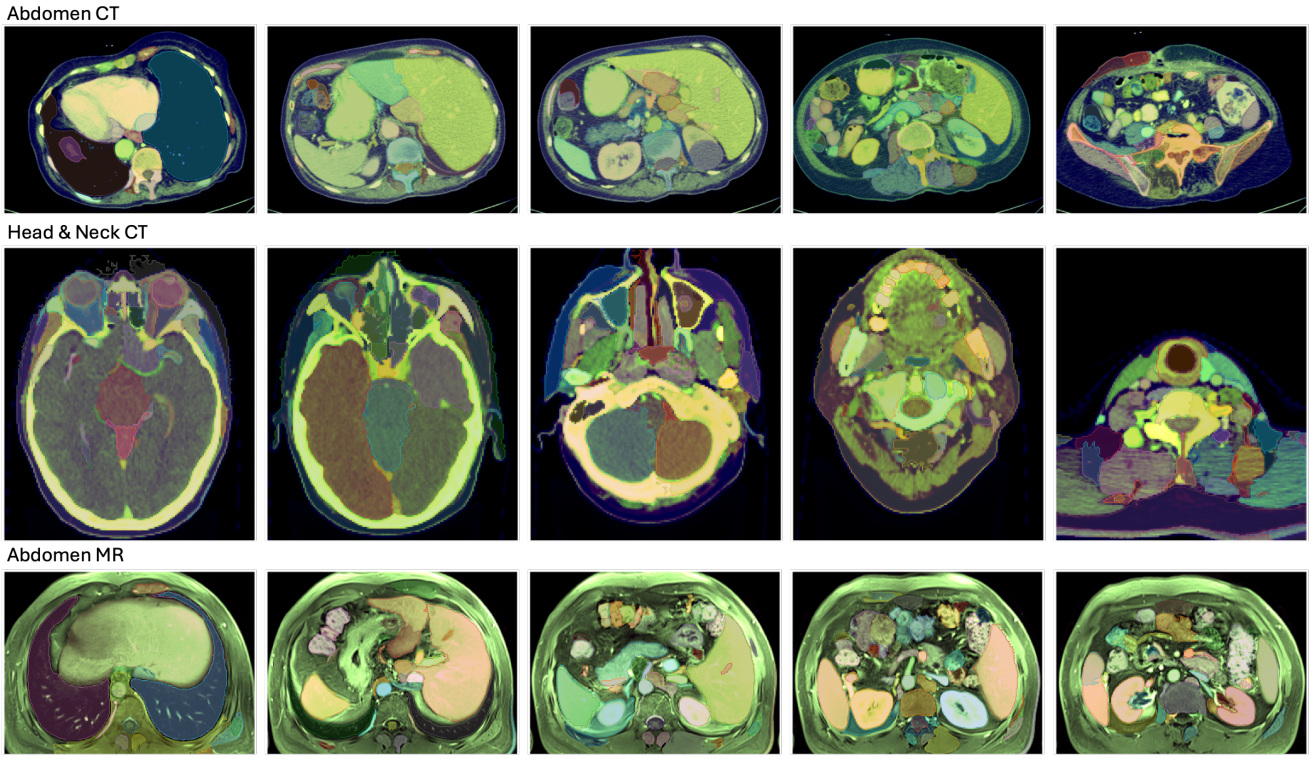

Figure 3. **SAM2 generated 2D masks on initial seed slices.** We show examples from abdomen CT, head & neck CT, and abdomen MR. Images appear in color because we create 3-channel inputs for SAM2 using three complementary intensity windows. Each tissue's color reflects its relative intensity across the three channels. SAM2 generates meaningful region proposals with good coverage of diverse anatomical structures including organs, bones, muscles, sub-organ regions, and pathologies (cysts). However, the masks also contain substantial noise from over-segmentation, missing objects and struggle with small subtle structures. Despite these imperfections, the diverse mask proposals spanning multiple granularities provide sufficient supervision for MASS to learn generalizable representations.

pervision across hundreds to thousands of diverse proposals per volume, which proves sufficient for MASS to learn broadly transferable medical imaging representations.

6.3. In-Context Segmentation Architecture

MASS adopts the Iris [21] architecture for in-context segmentation, which decouples task definition from query image inference through a lightweight task encoding module. As illustrated in Figure 1 (B), the framework consists of three components: an image encoder, a task encoding module, and a mask decoder. The details of the task encoding module is shown in Figure 5. Given a reference image-mask pair (x_s, y_s), the encoder extracts visual features F_s, which the task encoding module then processes to produce compact task embeddings that capture "what to segment." The task encoding module operates through two parallel streams: (1) *foreground feature encoding* upsamples encoder features to full resolution and applies the reference mask to preserve fine anatomical details, producing a pooled foreground embedding; (2) *contextual feature encoding* employs pixel shuffle operations for memory-efficient high-resolution feature-mask fusion, followed by cross-attention and self-attention layers with learnable query tokens to extract contextual information. The final task embedding combines both foreground and contextual representations. During inference, the mask decoder takes query image features and task embeddings as inputs, using cross-attention to enable information exchange between task-specific guidance and query features, producing the final segmentation in a single forward pass.

Critically, MASS is a flexible learning framework compatible with arbitrary encoder decoder architectures. The image encoder can be any feature extraction network without modification, e.g. convolutional architectures (ResNet, UNet), vision transformers, or hybrid models. The decoder requires only minor customization: adding a few cross-attention layers to fuse query features with task embeddings. Beyond these cross-attention mechanisms for task conditioning, the decoder architecture itself is arbitrary. Users can employ UNet-style decoders, transformer decoders, or any other segmentation head. This architectural flexibility enables MASS to leverage advances in backbone

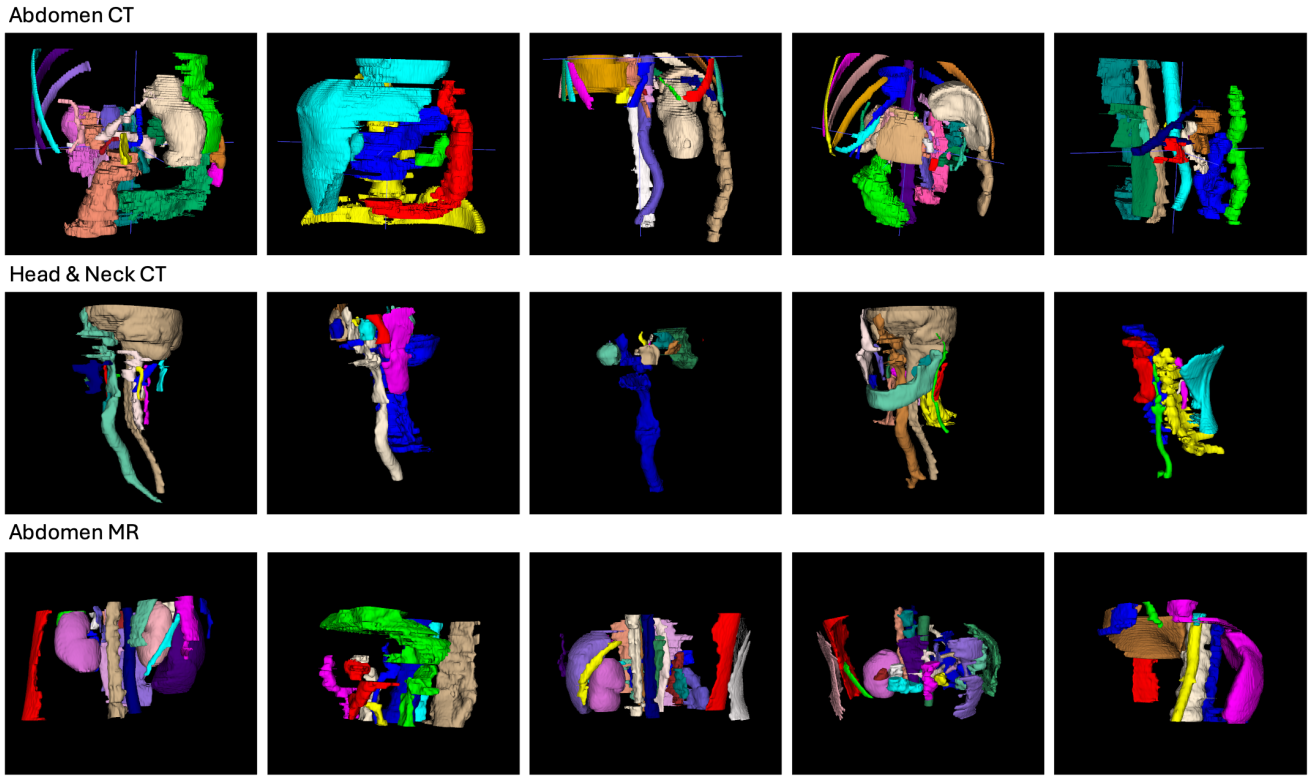

Figure 4. **SAM2 3D mask propagation results.** We show 3D masks generated by propagating the 2D seed masks from Figure 3 through the volume using SAM2's video prediction capability. SAM2 successfully converts 2D masks into volumetric segmentations by tracking boundaries across slices. While propagation maintains anatomical coherence for most structures, the results still contain noise from inconsistent boundaries and occasional tracking failures. These imperfect but volumetrically consistent masks provide the structural supervision needed for MASS pretraining.

design while maintaining its core self-supervised learning capability through mask-guided pretraining. We refer readers to the Iris paper [21] for comprehensive in-context segmentation architecture details and implementation specifics.

6.4. Implementation Details

Architecture. For most of our experiments, we employ a 3D ResUNet encoder with four downsampling stages, producing feature maps at multiple resolutions. The task encoding module follows the Iris [21] architecture. The decoder consists of four upsampling stages with skip connections from the encoder, incorporating cross-attention layers at 3 stages (lower resolution) to fuse task embeddings with query features. All models are trained with randomly initialized weights unless otherwise specified.

Training Configuration. We train MASS for 100 epochs using LAMB optimizer with an initial learning rate of 2×10^{-3}, weight decay of 1×10^{-5}, and polynomial learning rate decay with power 0.9. Training employs mixed precision (BF16) on 8 NVIDIA H100 GPUs [31] with PyTorch distributed data parallel (DDP). The effective batch size is 32 (4 per GPU). Loss combines Dice loss and binary cross-entropy with equal weighting. All 3D volumes are re-

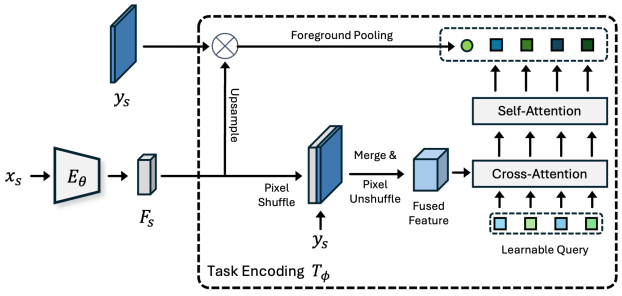

Figure 5. **Task encoding module architecture.** The module extracts compact task embeddings from reference image-mask pairs through two parallel streams: foreground feature encoding (top) captures fine anatomical details via high-resolution mask application, while contextual feature encoding (bottom) uses pixel shuffle operations and learnable query tokens with cross/self-attention to extract global context. The combined task embedding guides query image segmentation through the mask decoder. We follow [21] and refer the readers for more details in [21]

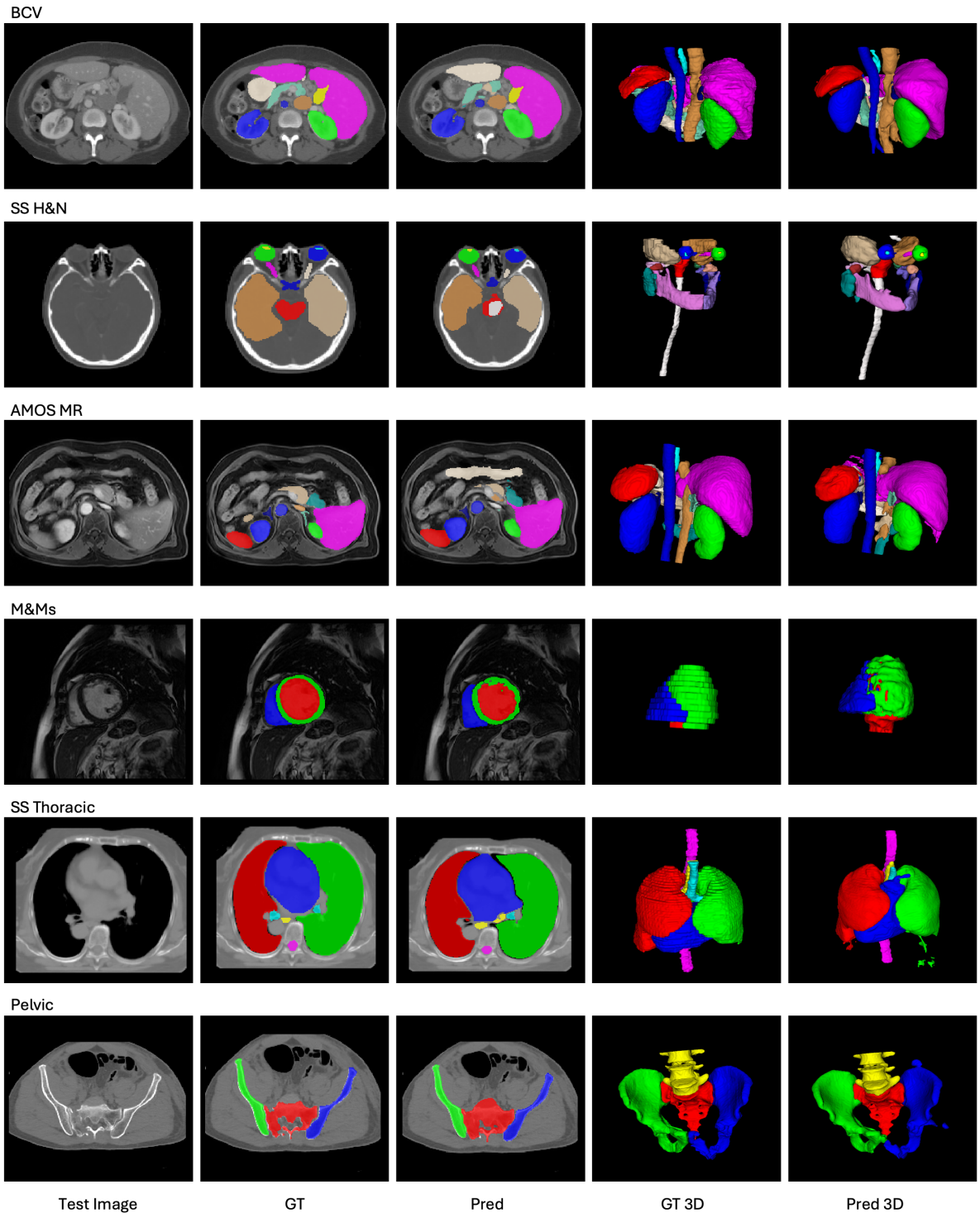

Figure 6. **Probing learned knowledge in pretraining through one-shot in-context inference.** We visualize the medical imaging understanding acquired by MASS during self-supervised pretraining by probing the model with one-shot in-context segmentation (no finetuning). Given a single reference image-mask pair, the pretrained model segments novel query images across diverse anatomical regions and modalities: BCV (abdominal CT), SS H&N (head & neck CT), AMOS MR (abdominal MRI), M&Ms (cardiac MRI), SS Thoracic (thoracic CT), and Pelvic (pelvic CT). The results demonstrate that MASS learns rich medical imaging knowledge including organ morphology, spatial relationships, and anatomical consistency entirely from mask-guided self-supervision without expert annotations.

sampled to consistent spacing of 1.5mm. The training 3D window size is cropped to $128 \times 128 \times 128$. Large-scale multi-modal pretraining (5K volumes) completes in 2 days.

Finetuning Protocol. For few-shot segmentation finetuning experiments, we fully finetune all parameters of all models. We train for 50 epochs using the same AdamW optimizer settings with learning rate 1×10^{-4}. Early stopping is applied based on validation performance. For in-context inference, we randomly sample k reference images from the support (training) set and get the average task embedding then make predictions when $k > 1$.

Classification Protocol. For frozen encoder classification, we attach classifier (attention pooling followed by a fully-connected layer) to the pretrained encoder. Only the classifier is trained using AdamW optimizer with learning rate 5×10^{-4} for 50 epochs.

7. Supplementary Experiments

7.1. Visualizing Learned Knowledge from Pretraining

To demonstrate that MASS acquires rich medical imaging knowledge during self-supervised pretraining, we probe the pretrained model using one-shot in-context segmentation as a visualization tool. Figure 6 shows qualitative results where the pretrained MASS model, given only a single reference example, segments novel anatomical structures across diverse body regions and modalities without any finetuning. This evaluation protocol serves as a direct probe into what anatomical concepts the model has internalized during pretraining.

The results reveal that MASS successfully learns fundamental medical imaging knowledge from mask-guided self-supervision: the model demonstrates understanding of organ morphology (shapes and sizes of anatomical structures), spatial relationships (relative positions of organs within the body), tissue boundaries (interfaces between different anatomical regions), and anatomical consistency (structures maintain coherent appearance across different patients). Critically, this knowledge emerges entirely from training on automatically generated class-agnostic masks without any expert annotations. The model has never seen expert-delineated boundaries or semantic labels during pretraining, yet it captures semantically meaningful anatomical concepts. The knowledge learned by MASS extends far beyond these shown examples—the visualization is inherently limited by our choice of reference prompts. The model has been exposed to thousands of diverse structures during pretraining through auto-generated masks, and we can only probe a small subset of this learned knowledge by selecting specific reference examples.

The visualization shows reasonable understanding across abdominal organs (BCV), head & neck structures (SS H&N), abdominal MRI (AMOS MR), cardiac structures (M&Ms), thoracic organs (SS Thoracic), and pelvic bones (Pelvic). While the predictions are not perfectly aligned with expert annotation standards, this actually validates our pretraining paradigm: MASS is like the learning process of large language models, where broad knowledge is first acquired through self-supervision, then aligned with human expert standards through minimal supervised finetuning. Table 2 quantifies this alignment process, demonstrating that just a few expert-annotated examples suffice to adapt the pretrained knowledge to downstream tasks with expert-level performance.

7.2. Feature Visualizations

To qualitatively assess what representations MASS learns during self-supervised pretraining, we visualize the learned feature maps using Principal Component Analysis (PCA). Specifically, we use the large-scale pretrained MASS model as a feature extractor without any finetuning, extract the decoder output features, and reduce the channel dimension to 3 using PCA for RGB visualization. This approach reveals the semantic structure captured by MASS's learned representations.

Figure 7 presents PCA visualizations on representative examples from abdomen CT, head & neck CT, and abdomen MR images. Several observations emerge from these visualizations:

Clear anatomical boundaries. Since MASS includes a decoder trained for segmentation, the output features maintain the original image resolution and exhibit remarkably sharp boundaries between anatomical structures. Organs such as kidneys, liver, spleen, heart, and lungs are clearly delineated, as are skeletal structures including pelvic bone, mandible, and skull. Soft tissue compartments (muscles, fat) also show distinct feature representations.

Semantic consistency. Semantically similar structures exhibit consistent feature representations (visualized as similar colors) across different spatial locations and patients. For instance, bilateral structures like left and right kidneys share similar feature patterns, indicating that MASS learns semantic concepts rather than merely spatial templates.

Multi-granular understanding. Most notably, MASS captures not only large anatomical structures but also fine-grained sub-anatomical details. The visualizations reveal clear representations of small structures such as pulmonary vasculature within the lungs, hepatic vessels within the liver, and distinct cardiac chambers. This multi-scale understanding emerges naturally from training on diverse auto-generated masks spanning multiple granularities.

These visualizations provide compelling evidence that mask-guided self-supervised pretraining enables MASS to learn rich, hierarchical medical imaging knowledge—from coarse organ-level semantics to fine sub-anatomical struc-

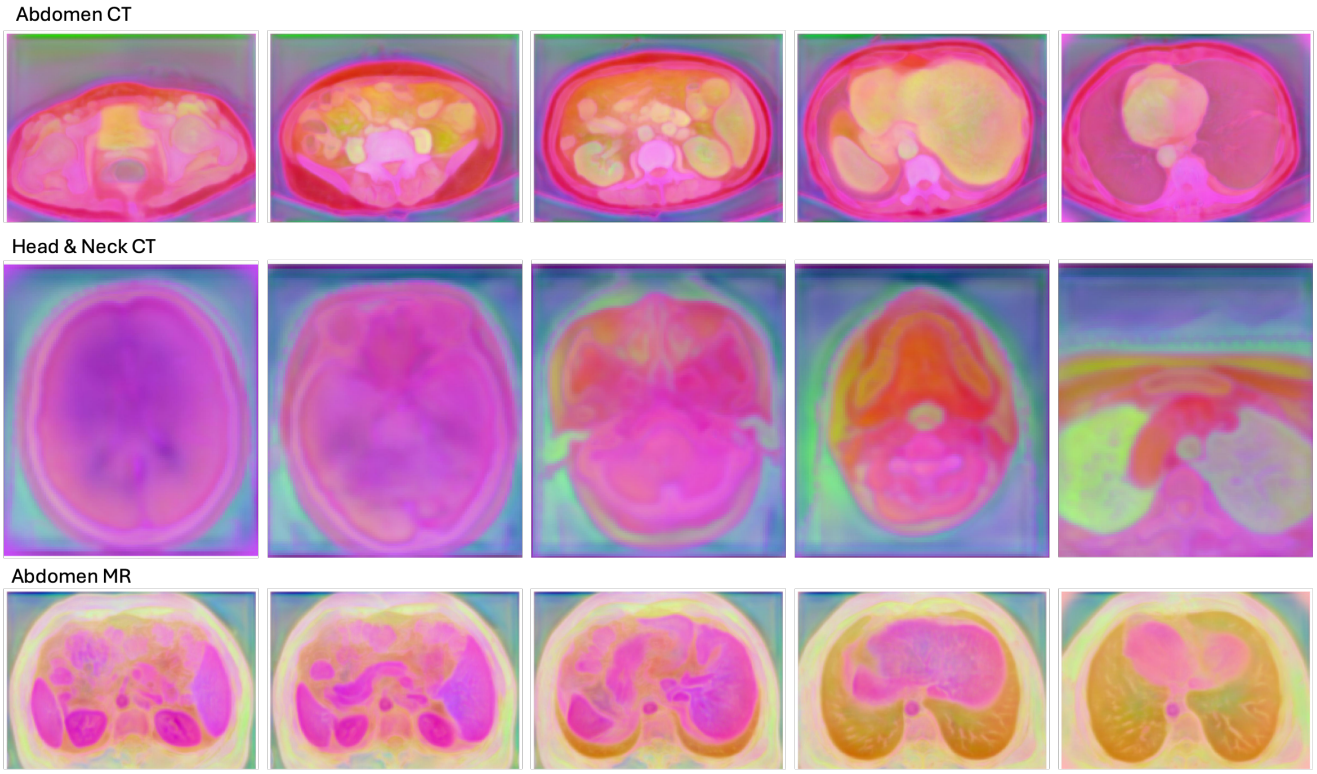

Figure 7. **PCA visualization of MASS learned features.** We extract decoder output features from the pretrained MASS model (without finetuning) and reduce to 3 channels via PCA for visualization. Examples shown include abdomen CT, head & neck CT, and abdomen MR images. The visualizations reveal that MASS learns semantically meaningful representations with clear anatomical boundaries at full image resolution. Large structures (kidneys, liver, spleen, heart, lungs, bones, muscles) are clearly delineated, while semantically similar structures share consistent feature patterns (similar colors). Notably, MASS also captures fine-grained sub-anatomical details including pulmonary vasculature, hepatic vessels, and cardiac chambers, demonstrating multi-granular anatomical understanding learned entirely through mask-guided self-supervision without expert annotations.

tures—entirely without expert annotations.

7.3. Failure Case Analysis

Tables 2 and 1 reveal a notable performance gap: MASS's in-context segmentation (without finetuning) achieves strong results on anatomical structures but struggles significantly on pathologies such as tumors. We investigate this phenomenon to understand its underlying causes and implications for MASS's learned representations.

Figure 8 visualizes a representative failure case on BraTS T1CE brain tumor data. The first row shows the reference image and mask defining the segmentation target; the second row displays the query image with ground truth tumor annotation; the bottom row presents MASS's prediction. We observe that the model tends to segment regions in spatial locations similar to where the reference mask appears, rather than identifying the actual tumor in the query image. This behavior stems from the inherent characteristics of pathological structures: unlike anatomical organs that occupy consistent spatial positions across patients, tumors exhibit substantial variance in location, size, and shape. Consequently, MASS struggles to establish correspondence between the reference and query when the target structure appears in drastically different positions.

This limitation is actually expected given MASS's pretraining objective from two main factors. First, SAM2-generated masks for pathologies are inherently noisier due to their diffuse, irregular boundaries—unlike large organs with well-defined edges, tumors often exhibit gradual intensity transitions that challenge boundary-based segmentation. Second, MASS's pretraining objective learns invariance between two augmented views of the *same* image, meaning the model is never exposed to how the same type of pathology appears across different patients. While anatomical structures maintain consistent spatial configurations across individuals (e.g., the liver is always in the right upper abdomen), pathologies can manifest anywhere within an organ with highly variable size, shape, and appearance. Thus, MASS learns strong within-patient invariance but lacks cross-patient correspondence for highly vari-

able structures.

Crucially, this does not indicate that MASS fails to learn meaningful representations of pathologies. To verify this, we conducted an analysis experiment where we use the query image (and mask) itself as the reference (with different augmentations applied to reference and query views), then segment the query image. Under this self-reference setting, MASS achieves over 70% Dice on BraTS T1CE, demonstrating that the model has indeed learned discriminative features for tumor tissues. The limitation lies not in representation quality but in cross-patient correspondence. MASS recognizes tumor features but does not inherently know that tumor features in different patients represent the same semantic concept.

This analysis explains why minimal finetuning with expert annotations dramatically improves pathology segmentation performance (Tables 2 and 1). The few labeled examples teach MASS that diverse-appearing pathological features across patients belong to the same semantic category, effectively bridging the cross-patient correspondence gap. The strong finetuning results validate that MASS's pretrained representations capture meaningful pathological features, they simply require minimal supervision to align with expert-defined semantic categories. Improving such cross-patient correspondence would be an interesting future research direction.

7.4. Additional Experiments

Baseline selection rationale. Since MASS is a self-supervised learning method, we primarily compare against other self-supervised approaches: methods pretrained on OpenMind [53] (MG, MAE, S3D, SimCLR trained on 114K multi-modal images), AnatoMix [17] (synthetic data generation), and Merlin [6] (15K CT with language and EHR supervision). We also include two supervised pretraining methods—SuPreM [35] (2.1K scans with 32 organ/tumor masks) and Iris [21] (2K multi-modal images)—to contextualize MASS's performance against methods that leverage expert annotations. Due to computational constraints, we evaluate all methods using their publicly released pretrained weights rather than retraining on our pretraining corpus. Table 7 summarizes the pretraining data characteristics of each method.

We deliberately exclude SAM-based medical imaging variants [15, 38] from our comparison for several reasons. First, these methods are finetuned in a supervised manner on very large-scale medical imaging datasets that have substantial overlap with our training and evaluation sets, making direct comparison unfair to MASS which uses no expert annotations. Second, we lack the computational resources to finetune these models on our data for a controlled comparison. Third, and most fundamentally, medical SAM methods are designed specifically for interactive segmentation with user prompts, whereas MASS aims at general-purpose representation learning that transfers beyond segmentation to tasks such as classification (Table 3). The different design objectives make these methods not directly comparable.

Despite using substantially less data (5K volumes vs. 114K for OpenMind) and requiring zero expert annotations, MASS achieves superior performance across downstream tasks, demonstrating exceptional data efficiency and the effectiveness of mask-guided self-supervised learning. This comparison, while not perfectly controlled due to different pretraining corpora, reflects realistic scenarios where practitioners must choose between available pretrained models

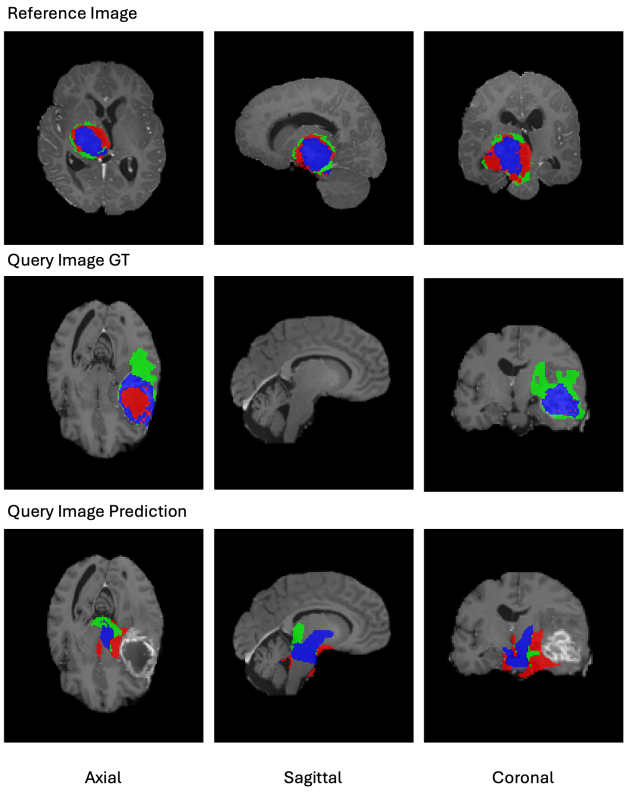

Figure 8. **Failure case analysis on pathology segmentation.** We visualize MASS's in-context segmentation on BraTS T1CE brain tumor data. Top row: reference image and mask defining the target tumor. Middle row: query image with ground truth tumor annotation. Bottom row: MASS's prediction, which incorrectly segments a region at a similar spatial location as the reference rather than the actual tumor. This failure stems from the large cross-patient variance in tumor location, size, and shape. MASS's pretraining objective learns within-image invariance but not cross-patient correspondence for highly variable pathologies. Notably, when using the query image as its own reference (with augmentation), MASS achieves $>$70% Dice, confirming that the model learns meaningful tumor representations but lacks cross-patient semantic alignment. This gap is efficiently bridged through minimal finetuning with expert annotated samples.

Table 6. Large-scale multi-modal pretraining segmentation performance (Dice %). Results shown as mean (standard deviation) over 3 runs. Numbers in brackets for "Full supervision" indicate the number of labeled samples for full supervised training. Bold: best performance; underlined: second best.

	BCV		AMOS MR		SS H&N		KiTS Tumor		LiTS Tumor		AutoPET		BraTS T1CE		ACDC		Pelvic
# shot	1	10	1	10	1	10	30	60	10	30	30	100	30	60	1	10	1
Full supervision	83.6 [23]		85.5 [38]		78.2 [39]		81.7 [159]		63.2 [99]		67.8 [983]		72.8 [241]		90.8 [70]		94.7 [80]
Scratch	27.3(3.8)	75.2(1.3)	32.9(4.0)	75.9(1.3)	51.8(3.2)	65.1(1.7)	35.7(3.6)	45.9(2.9)	42.5(2.6)	50.4(2.1)	40.1(2.4)	53.4(1.6)	54.0(1.9)	62.8(1.4)	38.7(3.4)	69.8(1.4)	57.8(3.1)
Supervised pretrain																	
SuPreM	63.9(1.9)	83.6(0.8)	55.1(2.1)	82.1(0.9)	66.1(1.7)	75.6(1.0)	64.1(1.2)	78.1(0.9)	53.9(1.6)	62.7(1.2)	48.8(1.5)	64.8(0.9)	60.3(1.2)	70.8(0.9)	55.9(2.0)	82.3(0.8)	85.4(1.3)
Iris (IC)	83.2(1.2)	85.4(0.7)	83.5(1.1)	86.4(0.7)	78.4(1.0)	80.1(0.7)	78.2(0.9)	80.2(0.7)	59.2(1.3)	63.3(1.0)	65.2(1.1)	69.5(0.8)	48.6(1.6)	79.8(0.8)	86.5(1.1)	88.2(0.7)	69.0(1.9)
Iris (FT)	83.4(1.0)	85.5(0.6)	83.6(0.9)	86.3(0.6)	78.5(0.8)	80.3(0.6)	78.3(0.8)	80.3(0.6)	59.4(1.2)	64.6(0.9)	67.2(1.0)	70.2(0.7)	60.7(1.0)	71.9(0.8)	86.9(0.9)	90.1(0.6)	86.9(1.1)
Self-supervised pretrain																	
OM-MG	49.0(2.6)	78.4(1.2)	38.8(2.9)	78.6(1.2)	61.3(2.3)	68.8(1.4)	41.1(1.9)	51.7(1.5)	48.1(1.9)	52.2(1.4)	44.6(1.7)	59.7(1.1)	58.5(1.4)	69.6(1.0)	46.8(2.5)	74.7(1.2)	76.7(2.1)
OM-MAE	48.8(2.7)	79.1(1.1)	37.9(3.0)	77.9(1.2)	59.1(2.4)	73.0(1.3)	47.4(1.7)	52.4(1.4)	40.5(2.0)	52.2(1.4)	43.4(1.8)	58.8(1.2)	56.3(1.5)	70.0(1.0)	45.4(2.6)	73.8(1.2)	72.2(2.2)
OM-S3D	46.4(2.8)	78.3(1.2)	38.9(2.9)	77.2(1.2)	59.8(2.4)	71.9(1.3)	44.3(1.8)	54.7(1.4)	42.8(1.9)	50.6(1.5)	42.3(1.8)	58.3(1.2)	59.4(1.4)	70.3(1.0)	46.3(2.6)	74.2(1.2)	73.3(2.1)
OM-SimCLR	45.6(2.9)	80.2(1.1)	37.0(3.1)	78.2(1.2)	58.8(2.5)	67.5(1.5)	46.8(1.7)	60.8(1.3)	49.2(1.8)	55.8(1.3)	45.4(1.7)	60.2(1.1)	56.0(1.5)	68.6(1.1)	48.9(2.5)	75.8(1.2)	77.0(2.0)
AnatoMix	53.1(2.4)	81.0(1.0)	35.9(3.2)	78.8(1.2)	48.3(2.8)	66.7(1.5)	40.6(2.0)	44.1(1.7)	49.9(1.8)	52.1(1.4)	46.1(1.7)	62.8(1.1)	58.7(1.4)	66.7(1.1)	42.8(2.7)	73.1(1.3)	82.2(1.8)
Merlin	50.1(2.5)	78.0(1.2)	37.9(3.0)	78.3(1.2)	62.7(2.2)	72.7(1.3)	51.1(1.6)	58.0(1.3)	49.2(1.8)	55.1(1.3)	41.8(1.9)	56.3(1.2)	53.2(1.5)	61.2(1.2)	45.8(2.6)	74.9(1.2)	79.3(1.9)
MASS (IC)	68.7(1.7)	73.6(1.5)	66.0(1.9)	71.6(1.6)	62.7(1.8)	63.5(1.7)	3.4(2.1)	4.3(1.9)	2.6(2.3)	4.5(2.0)	13.9(2.4)	18.6(2.1)	11.0(2.5)	12.0(2.3)	69.8(1.6)	75.8(1.4)	89.9(1.4)
MASS (FT)	**70.2(1.4)**	**84.2(0.8)**	**74.3(1.6)**	**85.0(0.7)**	**70.0(1.3)**	**78.9(0.9)**	**68.5(1.1)**	**79.1(0.8)**	**56.1(1.3)**	**64.5(1.0)**	**50.2(1.2)**	**65.2(0.8)**	**63.0(1.1)**	**72.3(0.9)**	**75.7(1.3)**	**90.0(0.7)**	**92.8(1.1)**

Table 7. Comparison of training data of large-scale pretraining baseline methods.

Method	**# Samples**	**Modalities**	**Supervision**
OpenMind	114K	24 MR	Self-supervised
Merlin	15K	CT	Report & EHR
AnatoMix	120K	Synthetic	Synthetic labels
SuPreM	2.1K	CT	32 organ/tumor masks
Iris	2K	CT, MR, PET	Expert annotated masks
MASS	5K	CT, MR, PET	Self-supervised

for their applications.

Ablation on augmentation strategy. Table 9 ablates augmentation design. Both spatial and appearance augmentations are necessary: spatial transformations maintain image-mask correspondence while appearance variations force semantic learning. Among magnitude levels, medium strength achieves optimal balance (65.5%)—small augmentations provide insufficient variation while excessive augmentations introduce instability. The augmentation parameters we used for training MASS is: affine transformations with probability 0.8 including scaling (± 0.3), rotation (± 30), and shear (± 0.1); appearance transformations with probability 0.2 including brightness (multiplicative $[0.8, 1.3]$, additive std 0.15), gamma ($[0.8, 1.3]$), contrast ($[0.8, 1.3]$), blur (sigma $[0.7, 1.5]$), and noise (std 0.04). Spatial augmentations are applied jointly to images and masks to maintain correspondence; appearance augmentations are applied only to images.

Inference Efficiency. MASS is a pretraining framework that does not impose significant inference overhead compared to standard segmentation models. Table 8 reports inference benchmarks using the 3D ResUNet architecture on a single NVIDIA H100 GPU with BF16 precision. Full in-context inference (including reference encoding and query segmentation) takes 88.84ms per volume, achieving 11.26 FPS. Notably, the reference encoding step (39.72ms) only needs to be computed once per task—when segmenting multiple query volumes for the same target structure, only the query inference (49.00ms, 20.41 FPS) is required per volume. This makes MASS practical for batch processing scenarios. Peak memory usage is 6.3GB for input size 128^3, comparable to standard 3D segmentation networks.

Model Parameters. Table 8 also presents the parameter breakdown. The full model contains 120.13M parameters, distributed across three components: the encoder (37.66M, 31.3%), decoder (46.00M, 38.3%), and task encoding module (36.35M, 30.3%). The encoder and decoder together comprise a standard segmentation backbone (83.66M), while the task encoding module adds 36.35M parameters to enable in-context learning. This represents a moderate overhead ($\sim$43%) compared to the base segmentation architecture. Importantly, the task encoding module is only used during reference processing; query inference primarily utilizes the encoder and decoder, resulting in efficient per-query computation. Furthermore, for downstream deployment with fixed target classes, the task encoding module can be entirely discarded. Only the encoder and decoder (83.66M) are required for standard segmentation finetuning. For classification tasks, only the encoder (37.66M) is needed as a feature extractor. This modular design allows practitioners to select the appropriate subset of components based on their deployment requirements. The overall model size remains comparable to typical 3D segmentation networks such as nnUNet and SwinUNETR, making MASS practical for clinical deployment.

7.5. Dataset Details

This section provides comprehensive information about the datasets used in our experiments. We organize datasets into two categories: upstream training datasets used for pretraining MASS, and downstream evaluation datasets used to assess generalization capabilities. Table 10 summarizes all datasets.

Table 8. Model specifications and inference benchmark on NVIDIA H100 GPU with BF16 precision. Input size: 128^3.

Metric	Value
Inference Time	
Full in-context inference	88.84 ± 0.13 ms
Reference encoding	39.72 ± 0.07 ms
Query inference	49.00 ± 0.02 ms
Query inference FPS	20.41
Memory	
Peak memory	6.3 GB
Parameters	
Encoder	37.66M (31.3%)
Decoder	46.00M (38.3%)
Task encoding module	36.35M (30.3%)
Total	120.13M

Table 9. Ablation study on augmentation strategies. All experiments evaluated on BCV 1-shot segmentation.

Aug. Strategy	Dice (%)	Aug. Magnitude	Dice (%)
Spatial only	60.7	Small	61.3
Appearance only	54.6	Medium	**65.5**
Both	**65.5**	Large	62.9

Table 10. Dataset statistics. Upper section: upstream training datasets. Lower section: downstream evaluation datasets held out entirely from pretraining or reserved modalities.

Dataset	Body Region	Modality	Clinical Target	#Cls	Size
AMOS CT [29]	Abdomen	CT	Organs	15	300
AMOS MR [29]	Abdomen	MRI	Organs	13	60
AutoPET [22]	Whole body	CT + PET	Lesions	1	1014
BCV [23]	Abdomen	CT	Organs	13	30
BraTS [39]	Brain	T1/T2/FLAIR	Tumors	3	213
KiTS [26]	Abdomen	CT	Kidney & Tumor	2	210
LiTS [5]	Abdomen	CT	Liver & Tumor	2	131
M&Ms [10]	Cardiac	cineMRI	Structures	3	320
StructSeg H&N [34]	Head & Neck	CT	Organs	22	50
StructSeg Tho [34]	Thorax	CT	Organs	6	50
TotalSeg CT [55]	Whole body	CT	Anatomies	104	1229
TotalSeg MR [55]	Whole body	MR	Anatomies	50	299
BraTS T1CE [39]	Brain	T1CE MRI	Tumors	3	213
ACDC [4]	Cardiac	cineMRI	Structures	3	100
Pelvic [37]	Pelvic	CT	Bones	4	103
RSNA ICH [19]	Head	CT	Hemorrhage	5	21744
RSNA Trauma [27]	Abdomen	CT	Trauma	2	4710

Data Leakage Prevention. We implement strict protocols to ensure no data leakage between pretraining and evaluation. For datasets used in both phases (BCV, AMOS, StructSeg, KiTS, LiTS, AutoPET), we partition data into non-overlapping train/validation/test splits at the patient level, using only training splits for pretraining and reserving test splits exclusively for evaluation. For BraTS, we additionally hold out the T1CE modality entirely from pretraining to evaluate cross-sequence generalization. Five datasets (ACDC, Pelvic, RSNA ICH, RSNA Trauma, and BraTS T1CE) are completely excluded from upstream training to serve as out-of-distribution evaluations. No test data from any dataset is used during MASS pretraining.

7.5.1. Upstream Training Datasets

Multi-organ Abdominal Collection (AMOS). AMOS [29] is a multi-modal dataset featuring 500 CT and 100 MRI scans from 600 patients with abdominal abnormalities, acquired across eight scanner platforms. The dataset provides annotations for 15 anatomical structures in CT and 13 structures in MRI. We use both modalities in upstream training with the official training set (200 CT, 40 MRI), implementing a 95%/5% split for training/validation. The official validation set (100 CT, 20 MRI) is reserved exclusively for downstream evaluation.

Whole-body PET/CT Collection (AutoPET). AutoPET [22] comprises 1,014 whole-body FDG-PET/CT studies, balanced between 501 cases with confirmed malignancies (lymphoma, melanoma, NSCLC) and 513 negative controls. We use both PET and CT modalities in upstream training with a 75%/5%/20% patient-level split for training, validation, and testing.

Abdominal CT from Multi-Atlas (BCV). The BCV [23] collection consists of 30 abdominal CT scans with annotations for 13 abdominal organs. We implement a 75%/5%/20% split (23/2/5 scans) for training, validation, and testing.

Brain Tumor Segmentation (BraTS). BraTS [39] features multi-parametric MRI scans (T1, T1CE, T2, FLAIR) from 213 glioma patients with annotations for three tumor subregions. For upstream training, we use only T1, T2, and FLAIR modalities with a 75%/5%/20% patient-level split. The T1CE modality is completely excluded from pretraining and reserved for downstream evaluation to test cross-sequence generalization.

Kidney Tumor Dataset (KiTS). KiTS19 [26] comprises 210 contrast-enhanced CT scans from kidney cancer patients with annotations for kidney and tumor regions. We use a 75%/5%/20% split for training, validation, and testing.

Liver Tumor Segmentation (LiTS). LiTS [5] contains 131 abdominal CT scans with annotations for liver parenchyma and tumor lesions. We employ a 75%/5%/20% split for training, validation, and testing.

Multi-Centre Cardiac Segmentation (M&Ms). M&Ms [10] features 320 cardiac cine-MRI scans from multiple centers and vendors with annotations for left ventricle, right ventricle, and myocardium. We use a 95%/5% split for training and validation. This dataset is used only for upstream pretraining; ACDC serves as the held-out cardiac evaluation dataset.

Radiation Treatment Planning (StructSeg). StructSeg [34] comprises CT imaging for radiation therapy plan-

ning. StructSeg H&N includes 50 scans with annotations for 22 head & neck organs-at-risk. StructSeg Tho contains 50 scans with annotations for 6 thoracic organs. We implement a 75%/5%/20% split for both components.

TotalSegmentator. TotalSegmentator [55] provides whole-body segmentation datasets: 1,229 CT scans with 104 anatomical structures and 298 MR scans with 50 structures. We use an 80%/10%/10% split for training, validation, and testing. TotalSegmentator is only used for upstream pretraining due to its great anatomy coverage.

Upstream Training Corpus Summary. For large-scale multi-modal pretraining, we combine training splits from all upstream datasets, totaling approximately 5K 3D volumes spanning CT, MRI, and PET modalities. This corpus covers diverse anatomical regions (whole-body, abdomen, cardiac, brain, head & neck, thorax) and clinical targets (organs, tumors, lesions), simulating real-world clinical repositories.

7.5.2. Downstream Evaluation Datasets

The following datasets are held out entirely from upstream training or represent reserved modalities, used exclusively to evaluate out-of-distribution generalization.

Brain Tumor Segmentation - T1CE (BraTS T1CE). We use the T1CE modality from BraTS [39], which was completely excluded from upstream training (only T1, T2, FLAIR were used). This evaluates cross-sequence generalization on the same patients but an unseen MRI contrast. We use the same 75%/5%/20% patient-level split as upstream BraTS.

Automated Cardiac Diagnosis Challenge (ACDC). ACDC [4] consists of 100 cardiac cine-MRI scans from a different institution (University Hospital of Dijon) than M&Ms, with different scanner configurations (Siemens 1.5T/3.0T). This dataset is entirely excluded from upstream training to serve as an out-of-distribution cardiac evaluation. We use a 75%/5%/20% split.

Pelvic Bone Segmentation (Pelvic). Pelvic1K [37] contains 103 CT scans annotated for four pelvic skeletal structures. This dataset is entirely excluded from upstream training to test generalization to novel anatomical regions not seen during pretraining. We employ a 75%/5%/20% split.

RSNA Intracranial Hemorrhage Detection (RSNA ICH). RSNA ICH [19] comprises 21,744 non-contrast head CT scans with multi-label annotations for five hemorrhage subtypes. This dataset is entirely excluded from upstream training and evaluates transfer to pathology classification, a different task (classification vs. segmentation) on an unseen pathology. We split into 15,220/2,174/4,350 for train/validation/test, evaluating frozen encoder performance with 5%/30%/100% of training data tuning.

RSNA Abdominal Trauma Detection (RSNA Trauma). RSNA Trauma [27] contains 4,710 contrast-enhanced abdominal CT scans with annotations for trauma injury in abdomen organs. We use the severe trauma on three solid organs, liver, kidney, and spleen, for evaluation. This dataset is entirely excluded from upstream training. Although abdominal CT appears in pretraining, trauma detection represents a novel clinical task distinct from anatomical or tumor segmentation. We split into 65%/5%/30% for train/validation/test, evaluating frozen encoder performance with 5%/30%/100% of training data across three organ-specific binary classification tasks.