AUDIOROUTER: AUDIO-GUIDED VISUAL COMPRESSION FOR STREAMING VIDEO LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Bringing temporal modeling from vision-language models into 3D medical image understanding offers a practical way to process full-length CT and MRI volumes and improve training stability. Existing methods still have two main limits. They often treat the spatial Z axis as a uniformly sampled time axis and ignore the true physical slice thickness, which weakens spatial perception. They also rely on external foundation models for one-way slice-by-slice mask propagation, which can accumulate errors in long sequences. We propose PhysiVLM, a physically constrained two-stage framework that explicitly separates 3D spatial reasoning from 2D mask generation. In the first stage, we use an anatomical 2.5D grid puzzle and spatial coordinate prompts to inject the prior of physical slice thickness, which allows the model to regress the global 3D depth range of a lesion at low cost and with high accuracy. In the second stage, we introduce a Spatial Prompting mechanism to build anchor-guided reasoning. The model takes the predicted depth range from the first stage, dynamically samples key slices, and then guides the VLM to perform slice-wise spatial semantic reasoning, parse lesion shape, and output structured 2D bounding boxes. We then feed these boxes into MedSAM2 to produce spatially consistent bidirectional mask propagation along the Z axis. Extensive experiments show that PhysiVLM reduces cumulative errors in long-sequence propagation, cuts redundant computation, and improves segmentation accuracy for lesions with complex cross-slice morphology. The code is at: https://anonymous.4open.science/r/PhysiVLM-1E33

1 INTRODUCTION

Medical image segmentation supports lesion assessment, anatomical delineation, treatment planning, and response evaluation Litjens et al. (2017); Isensee et al. (2021). Accurate tumor segmentation is particularly challenging because tumors often exhibit irregular boundaries and heterogeneous appearances. A practical segmentation system must jointly achieve generalizable representation, accurate 3D localization, and efficient volumetric processing.

Three factors make this problem difficult. First, medical volumes have physical spatial structures rather than temporal structures. CT and MRI slices are positioned according to slice spacing and thickness, whereas video frames represent sampled time steps. Treating the Z axis as time therefore discards physical depth cues and weakens lesion localization. Second, accuracy and efficiency are often conflicting. Conventional 3D networks capture volumetric context through 3D convolutions or attention but require substantial memory and commonly rely on resampling, cropping, or patch-based inference Chen et al. (2019); Hatamizadeh et al. (2022); Zhou et al. (2023); Xing et al. (2026); Shi et al. (2025). Efficient 2D models, by contrast, may lose spatial information when processing long slice sequences. Third, semantic reasoning and mask generation remain weakly connected. Promptable segmentation models mainly propagate masks through visual correspondence, causing error accumulation and cross-slice drift when lesion morphology changes sharply.

Vision-language models (VLMs) provide strong spatial reasoning capabilities Ma et al. (2024a); Cheng et al. (2025); Tian et al. (2025) and show promise for generalizable CT and MRI analysis Zhai et al. (2023); Hamamci et al. (2026); Xin et al. (2025). Recent 3D-as-Video methods further reduce computation by unfolding volumes into 2D slice sequences Lai et al. (2024); Xin et al. (2025); Du et al. (2023); Qinhao et al. (2025); Bai et al. (2024); Zhao et al. (2025). However, these methods rarely

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.

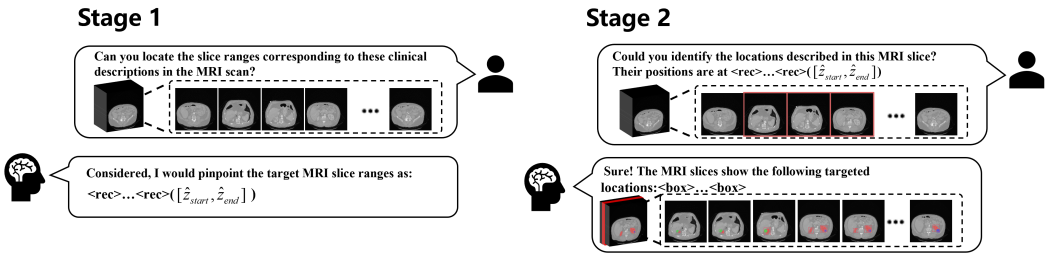


Figure 1: Progressive lesion localization in PhysiVLM. The model first regresses the physical depth interval of the target lesion along the Z axis, $[\hat{z}_{\text{start}}, \hat{z}_{\text{end}}]$, from a global view. PhysiVLM then focuses on slices within this interval and generates structured 2D bounding boxes that guide downstream mask generation.

model physical depth explicitly, and efficient sequence processing alone cannot provide sufficient semantic constraints for reliable mask propagation.

We propose PhysiVLM, a physically constrained framework that separates global 3D spatial reasoning from local 2D mask generation (Fig. 1). A VLM first localizes the lesion in physical space, after which a promptable segmentation model generates high-resolution masks only within the predicted region. This decomposition avoids end-to-end volumetric encoding while preserving anatomical context and connecting semantic reasoning with mask propagation.

In Stage 1, the Anatomy-Aware 2.5D Grid stacks neighboring slices as channels and arranges the resulting groups into a spatial grid. Local stacks preserve short-range continuity, while the grid exposes long-range anatomical context to a 2D VLM. PhysiVLM then predicts the physical lesion interval, $[\hat{z} * \text{start}, \hat{z} * \text{end}]$. In Stage 2, Spatial Coordinate Prompting predicts bounding boxes for sampled key slices within this interval. PhysiVLM selects an anatomically informative anchor slice and supplies its bounding box to MedSAM2, which propagates masks bidirectionally along the Z axis (Fig. 2). The 2.5D grid provides efficient physical localization, while explicit coordinate prompts constrain propagation and preserve fine lesion boundaries.

We evaluate PhysiVLM on challenging CT and MRI segmentation datasets. PhysiVLM improves depth localization and segmentation accuracy over video-based large-model methods while reducing error accumulation during MedSAM2 propagation, particularly for lesions with irregular cross-slice morphology. The framework also avoids the memory cost of end-to-end volumetric encoding.

Our contributions are summarized as follows:

- We propose PhysiVLM, a physically constrained framework that decomposes efficient 3D segmentation into global lesion localization and local mask generation.
- We introduce the Anatomy-Aware 2.5D Grid, which enables a 2D VLM to preserve cross-slice context and predict physical lesion intervals without volumetric encoding.
- We develop Spatial Coordinate Prompting to convert semantic reasoning into explicit bounding-box constraints for anchor selection and bidirectional mask propagation.
- Experiments on CT and MRI datasets demonstrate improved localization and segmentation accuracy, reduced propagation errors, and lower volumetric processing costs.

The remainder of this paper reviews related work, describes PhysiVLM, reports the experimental results, and discusses its limitations and implications.

2 Related Work

2.1 Multimodal Large Language Models for 3D and Video Understanding

Traditional 3D medical image segmentation methods Chen et al. (2024); Hatamizadeh et al. (2021); Zhou et al. (2022); Isensee et al. (2024) achieve strong performance in closed-task settings but often rely on task-specific architectures, 3D convolutions, or volumetric attention Li et al. (2024a;b);

Rokuss et al. (2025). Processing large CT and MRI volumes therefore requires substantial memory and preprocessing, including resampling, cropping, and patch-based inference. These operations may reduce global anatomical continuity and limit cross-domain generalization Ye et al. (2025); Yu et al. (2025).

Recent 3D-as-Video methods reduce this cost by representing volumes as 2D slice sequences and reusing video-based multimodal models Du et al. (2023); Xin et al. (2025); Wenhui Lei (2023). However, directly treating the physical (Z) axis as time ignores slice thickness and spacing, which weakens anatomical depth reasoning. PhysiVLM instead introduces an Anatomy-Aware 2.5D Grid and Spatial Coordinate Prompting to encode physical depth without using 3D convolutions.

2.2 Promptable Foundation Models in Medical Segmentation

Promptable foundation models such as SAM and DINO have advanced general-purpose segmentation Kirillov et al. (2023); Jose et al. (2024). MedSAM adapts SAM to medical images through large-scale fine-tuning Ma et al. (2024b), while MedSAM2 extends prompt-based segmentation to 3D volumes using a streaming memory bank Ma et al. (2025). Given a bounding box on one slice, MedSAM2 propagates the mask bidirectionally through cross-slice visual matching.

Although effective for pixel prediction, this propagation relies mainly on low-level feature similarity and lacks semantic control. Large appearance changes or irregular lesion boundaries can therefore cause mask drift and cumulative errors. PhysiVLM addresses this limitation through VLM-guided anchor selection and spatial reasoning, which provide high-confidence prompts and constrain MedSAM2 propagation across slices.

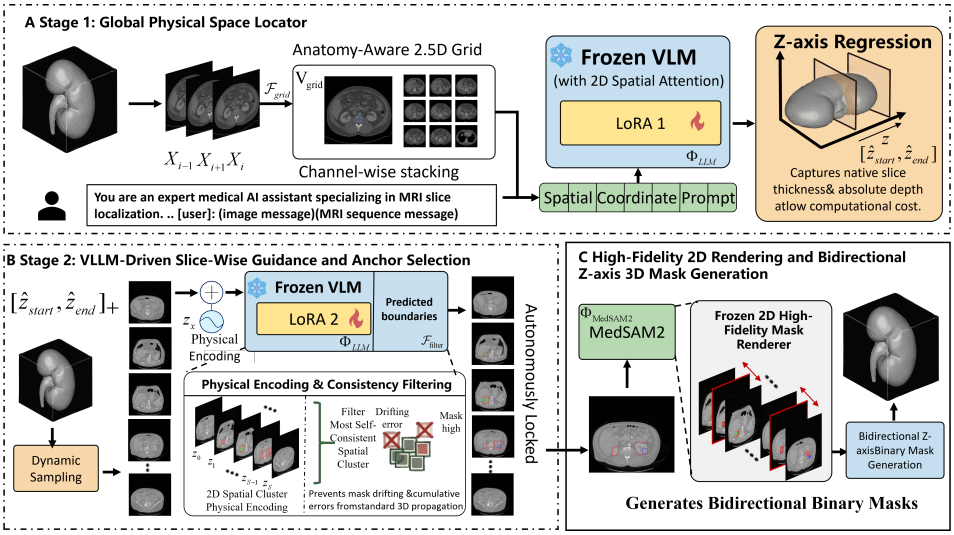


Figure 2: PhysiVLM framework. Stage 1 constructs an Anatomy-Aware 2.5D Grid and predicts the physical lesion interval, $[\hat{z} * \text{start}, \hat{z} * \text{end}]$. Stage 2 predicts bounding boxes on selected key slices, chooses an informative anchor slice, and uses its box to guide bidirectional MedSAM2 propagation. PhysiVLM therefore separates global 3D localization from local 2D mask generation.

3 Methodology

3.1 Overall Architecture

We propose PhysiVLM, a text-driven two-stage framework for 3D medical image segmentation. Directly processing a full 3D volume leads to high computational cost and often causes spatial distortion in vision-language models (VLMs). To address this issue, PhysiVLM separates the segmentation pipeline into two parts: high-level spatial reasoning and low-level mask generation. Given a 3D medical image X and a natural-language clinical query Q, the first stage serves as a global physical-space locator and predicts the global depth interval of the target lesion. The second

stage provides slice-level guidance. In this stage, we introduce a Spatial Prompting mechanism to drive spatial semantic reasoning in the VLM and use a foundation model, MedSAM2, to perform controlled spatially consistent 2D mask generation.

3.2 Stage 1: Global Physical Space Locator

Anatomy-Aware 2.5D Grid and Visual Encoding. To capture morphological gradients along the Z axis while keeping the visual encoder Φ_{visual} frozen, we introduce an Anatomy-Aware 2.5D Grid module, denoted by $\mathcal{F}_{\text{grid}}$. Before constructing the grid, we stack each grayscale slice X_i with its adjacent slices in physical space and form a three-channel image, that is, $[X_{i-1}, X_i, X_{i+1}]$. This lightweight design matches the input format of the visual encoder and adds local depth context at the same time. We then arrange the processed slices on a 2D plane through pixel-level stitching to form a global grid. The visual encoder extracts features from the grid,

$$V_{\text{grid}} = \Phi_{\text{visual}}(\mathcal{F}_{\text{grid}}(X)), \tag{1}$$

where V_{grid} denotes the visual feature map. We next concatenate V_{grid} with the physical coordinate prompt Q^{spatial} and feed the resulting multimodal sequence into the large language model Φ_{LLM}.

Target Depth Regression. The large language model produces hidden states for K predefined retrieval tokens, denoted by $Z_{\text{rec}} = \{z_1, z_2, \ldots, z_K\}$. We use a lightweight multilayer perceptron decoder Φ_{loc} to map these hidden states to the predicted depth boundaries and a foreground confidence score $\hat{p}_{\text{loc}}$,

$$[\hat{z}_{\text{start}}, \hat{z}_{\text{end}}], \hat{p}_{\text{loc}} = \Phi_{\text{loc}}(Z_{\text{rec}}). \tag{2}$$

During training, we use Hungarian matching to assign predicted intervals to ground-truth targets and optimize the model with a joint loss,

$$\mathcal{L}_{\text{loc}} = \lambda_{\text{text}}^{(1)} \mathcal{L}_{\text{text}}^{(1)} + \mathcal{L}_{\text{reg}}. \tag{3}$$

Here, $\mathcal{L}_{\text{text}}^{(1)}$ denotes the text generation loss of the large language model, which constrains the model to produce the predefined response format and the required retrieval tokens. The term $\mathcal{L}_{\text{reg}}$ denotes the interval regression loss, which refines the predicted depth coordinates and the matching confidence. It contains three terms,

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{L1}} \mathcal{L}_{\text{L1}} + \lambda_{\text{iou}} \mathcal{L}_{\text{iou}} + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}}. \tag{4}$$

Here, $\mathcal{L}_{\text{L1}}$ and $\mathcal{L}_{\text{iou}}$ denote the L1 distance loss and the 1D generalized intersection over union (1D GIoU) loss, respectively. Unlike standard 2D or 3D overlap measures in medical imaging, 1D GIoU measures the alignment of slice intervals along the z axis Lei et al. (2021). $\mathcal{L}_{\text{cls}}$ is a binary cross-entropy loss that predicts whether a matched interval belongs to the foreground or the background.

3.3 Stage 2: VLM-Driven Slice-Wise Spatial Guidance

In the second stage, PhysiVLM explicitly separates high-level spatial reasoning from low-level pixel segmentation. Specifically, the vision-language model Φ_{LLM} models anatomical variation and spatial relations across slices, while the frozen foundation segmentation model Φ_{MedSAM2} is used only for controlled mask generation. This division of roles avoids interference between long-range 3D reasoning and dense pixel prediction within a single model, and it improves structural consistency and boundary stability when lesion appearance changes substantially across slices.

Dynamic Sampling and Spatial-Prompted Reasoning. Given the depth interval predicted in Stage 1, $[\hat{z}_{\text{start}}, \hat{z}_{\text{end}}]$, we use a uniform interval sampling function $\mathcal{F}_{\text{sample}}$ to select S spatially distributed anchor slices within this range,

$$X_{\text{sub}} = \mathcal{F}_{\text{sample}}(X, [\hat{z}_{\text{start}}, \hat{z}_{\text{end}}], S)\,, \tag{5}$$

where S is the number of anchor slices used for reasoning and is set to 10 in our experiments. This design provides multiple observations with known physical spacing inside the predicted target range and turns long-range propagation from a single initialization point into shorter-range propagation constrained by multiple anchors.

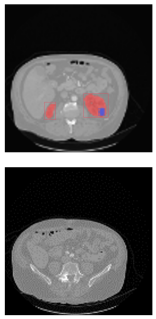


Figure 3: Semantic reasoning and target rejection in PhysiVLM. Top: PhysiVLM localizes and boxes anatomical structures and lesion regions from functional descriptions rather than explicit category names. Bottom: when the queried target is absent from the sampled slices, the model explicitly returns "target not found." This negative rejection behavior reduces forced predictions, limits hallucinations, and improves reliability in clinical use.

For each sampled slice $X_s \in X_{\text{sub}}$, we build a spatial prompt P_s, which contains the slice image, its absolute physical coordinate, and the task query. Conditioned on P_s, the VLM performs slice-wise spatial semantic reasoning and predicts a candidate bounding box,

$$B_s = \Phi_{\text{box}}(\Phi_{\text{LLM}}(X_s, P_s)), \quad \forall s \in \{1, \ldots, S\}. \tag{6}$$

In this stage, the VLM does not produce segmentation masks directly. Instead, it identifies a set of reliable anatomical anchors from spatial prompts and provides structured guidance for the downstream mask generation stage.

Anchor Consolidation and Multi-Anchor Rendering. Given the candidate box set $\{B_s\}_{s=1}^{S}$, we apply an anchor consolidation function $\mathcal{F}_{\text{filter}}$ to remove redundant predictions and retain valid anchors,

$$\mathcal{A} = \mathcal{F}_{\text{filter}}\left(\{B_s\}_{s=1}^{S}\right), \tag{7}$$

where $\mathcal{A} = \{(X_{a_k}, B_{a_k})\}_{k=1}^{K}$ denotes the retained anchor set, and $\mathcal{F}_{\text{filter}}$ includes simple post-processing steps such as IoU-based box filtering and non-maximum suppression.

We then use the anchor set $\mathcal{A}$ as a multi-point initialization for MedSAM2. For each anchor (X_{a_k}, B_{a_k}), MedSAM2 produces a local bidirectional propagation result,

$$\hat{Y}^{(k)} = \Phi_{\text{MedSAM2}}(X, B_{a_k}), \qquad k = 1, \ldots, K. \tag{8}$$

The final 3D segmentation is obtained by aggregating the predictions from all anchors,

$$\hat{Y} = \mathcal{G}\left(\{\hat{Y}^{(k)}\}_{k=1}^{K}\right), \tag{9}$$

where $\mathcal{G}(\cdot)$ denotes the final mask aggregation operation. Compared with the default single-anchor propagation in MedSAM2, this multi-anchor initialization shortens the propagation distance in each run and reduces error accumulation and mask drift along the Z axis.

Finally, we define the joint loss as

$$\mathcal{L}_{\text{spa}} = \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{text}}^{(2)}\mathcal{L}_{\text{text}}^{(2)} + \lambda_{\text{box}}\mathcal{L}_{\text{box}}. \tag{10}$$

Here, $\mathcal{L}_{\text{mask}}$ supervises pixel-level segmentation quality and is defined as

$$\mathcal{L}_{\text{mask}} = \lambda_{\text{f}}\mathcal{L}_{\text{focal}} + \lambda_{\text{d}}\mathcal{L}_{\text{dice}} + \lambda_{\text{i}}\mathcal{L}_{\text{IoU}}, \tag{11}$$

which addresses the foreground-background imbalance that is common in medical images. The loss $\mathcal{L}_{\text{box}}$ uses the standard ℓ_1 and GIoU losses to supervise box regression. The term $\mathcal{L}_{\text{text}}^{(2)}$ constrains the structured slice-level output in Stage 2, such as bounding-box responses or *target-not-found* responses, and improves the spatial localization ability of the hidden states.

Table 1: Performance comparison on KiTS23 and MSD datasets. To ensure a fair and comprehensive evaluation, baselines are categorized by their paradigms. **Ext. Data** indicates whether the method utilizes external datasets for pre-training beyond the target training set; **Sup.** indicates the supervision paradigm during evaluation (FS: Fully Supervised on target dataset, ZS: Zero-Shot, FT: Fine-Tuned); **Prompt** specifies the input prompt modality (None, Point, Box, Text).

Method	Venue	Ext. Data	Sup.	Prompt	KiTS23				MSD			
					Dice ↑	HD95 ↓	Prec. ↑	Rec. ↑	Dice ↑	HD95 ↓	Prec. ↑	Rec. ↑
Group A: Traditional 3D Volumetric Networks (Task-specific, No external data)												
nnFormer	Arxiv'21	No	FS	None	75.85	-	77.3	74.2	42.8	-	45.0	40.8
nnUNet-Res	MICCAI'24	No	FS	None	72.8	27.3	78.1	72.4	50.0	58.1	52.1	56.8
VSmTrans	MIA'24	No	FS	None	48.7	35.2	50.2	47.3	47.5	39.8	49.1	46.0
U-KAN	AAAI'25	No	FS	None	83.41	50.8	88.43	84.6	46.0	56.8	64.5	39.7
SuperLightNet	CVPR'25	No	FS	None	51.3	32.1	52.8	49.9	49.6	36.5	51.4	48.0
U-RWKV	TIP'26	No	FS	None	76.41	50.8	79.93	75.7	46.0	56.8	64.5	39.7
Group B: Promptable Visual Foundation Models												
SAM-Med3D	ECCV'24	Yes	ZS	P/B	58.6	212.2	60.1	57.2	55.4	48.7	57.3	53.6
DINOv3-L	Arxiv'25	Yes	ZS	None	32.2	240.1	31.0	35.5	51.8	52.3	53.6	50.1
MedSAM-2	arXiv'24	Yes	FT	P/B	67.5	94.3	80.2	76.9	64.3	38.9	66.1	62.6
SegVol	NIPS'24	Yes	ZS	B/T	74.8	181.5	79.7	72.2	66.6	119.3	72.9	64.7
Group C: Text-Guided Segmentation & VLM-based Pipelines												
VILA-M3	Arxiv'24	Yes	ZS	Text	59.4	44.7	79.2	54.7	33.3	70.4	37.9	32.4
M3D-LaMed	Arxiv'24	Yes	FT	Text	83.1	24.5	84.6	81.7	69.4	80.7	73.5	78.2
R1Seg-3D	MICCAI'25	Yes	FT	Text	70.5	138.1	73.4	70.8	56.4	144.0	58.5	59.4
PhysiVLM (Ours)	-	**Yes**	**FT**	**Text**	**86.3**	**124.2**	**89.2**	**88.8**	**73.2**	**143.1**	**83.9**	**78.5**

Note: Comparisons across different groups highlight the architectural and paradigm shifts in medical image segmentation. The most direct apples-to-apples comparison for our method lies within Group C, where PhysiVLM demonstrates superior performance among VLM-based pipelines.

3.4 Learnable Special Tokens for Physical Encoding and Action Spaces

To connect discrete text outputs from the large language model with continuous physical coordinates in 3D medical images, we introduce three types of task-specific learnable special tokens. During training, we update only the embeddings of these tokens, together with the LoRA parameters described above. This lightweight design gives the VLM the ability to regress spatial coordinates and encode physical structure.

Physical Coordinate Prompts (`<z>`). To correct the Z-axis distortion introduced by the 3D-as-Video setting, we do not use uniform sequence indices. Instead, we introduce absolute physical depth tokens defined by the true slice thickness. Specifically, we prepend each slice with its cumulative physical distance in anatomical space, measured in millimeters. For example, a slice located 15 mm from the starting position is assigned the token `<z_15>`. This mechanism injects non-uniform physical spacing into the attention process and allows the model to represent the true anatomical extent of the lesion.

Depth Retrieval Tokens (`<rec>`). In Stage 1, the system instructs the VLM to output retrieval tokens together with its semantic response. We extract the corresponding hidden states from the final network layer and use a lightweight decoder to regress the 1D physical depth interval of the target lesion, $[\hat{z}_{\text{start}}, \hat{z}_{\text{end}}]$.

Spatial Bounding Tokens (`<box>`). In Stage 2, for each sampled target slice, the VLM generates spatial bounding tokens. We then pass the corresponding hidden states through a box decoder to obtain continuous 2D coordinates, $[x_1, y_1, x_2, y_2]$. These coordinates serve as deterministic visual prompts for MedSAM2 and drive slice-wise segmentation. Through these special tokens, PhysiVLM casts 3D lesion localization and segmentation into an end-to-end sequence generation framework.

4 Experiments

4.1 Datasets and Implementation

We train PhysiVLM on 21 public CT and MRI segmentation datasets spanning multiple anatomical regions and lesion types, including multi-organ, organ/lesion-specific, spine, HaN-Seg, and three

Medical Segmentation Decathlon subsets Kavur et al. (2021); Ji et al. (2022); Wasserthal et al. (2023); Landman et al. (2015); Rister et al. (2020); Heller et al. (2023); Simpson et al. (2019). For Stage 1, the first and last foreground slices of each 3D mask define $[z_{\text{start}}, z_{\text{end}}]$. Following M3D, GPT-4 generates clinical prompts that describe target shape, location, and anatomy.

We train on eight NVIDIA H800 GPUs. During SFT, we freeze the visual encoder, apply LoRA to the VLM backbone, and optimize the special-token embeddings and localization decoders. We split all data at the patient level.

4.2 Main Results and Qualitative Analysis

We compare PhysiVLM with conventional 3D networks, promptable foundation models, and text-guided VLMs on KiTS23 and the more heterogeneous four-task MSD benchmark. Table 4 summarizes segmentation performance and computational characteristics. PhysiVLM obtains the highest Dice on both datasets (86.3% and 73.2%). Although U-KAN reaches 83.41% on KiTS23, U-KAN and U-RWKV fall to 46.0% on MSD; SegVol and M3D-LaMed reach 66.6% and 69.4%, respectively.

Table 2: **Size-stratified segmentation on MSD.** Small targets contain $<$ 29k voxels; large targets contain $\geq$ 29k voxels.

Method	Size	Dice ↑	HD95 ↓	Prec. ↑	Rec. ↑
VILA	Large	57.8	181.4	70.1	70.7
	Small	48.7	275.9	61.3	**71.7**
M3D-LaMed	Large	**89.8**	**22.6**	92.8	**94.4**
	Small	49.3	**138.7**	54.1	57.3
R1Seg-3D	Large	76.5	55.2	70.0	67.0
	Small	36.3	232.8	40.0	41.1
SegVol	Large	82.2	24.5	86.7	80.7
	Small	58.6	166.7	65.8	56.4
PhysiVLM	Large	85.8	31.6	**94.0**	87.8
	Small	**68.5**	165.2	**76.5**	67.3

Table 3: **Mask drift by distance to the nearest anchor.** Distance is measured in slices. C-Shift denotes centroid displacement, A-Error denotes relative area error, and Single denotes single-anchor propagation.

Method	Metric	0	1–2	3–5	6–10	>10	Slope ↓
Single	Dice (%)	72.5	60.3	53.3	54.2	69.9	
	C-Shift	2.98	11.21	14.26	19.56	33.51	0.006
	A-Error	0.249	1.04	1.66	3.25	10.62	
PhysiVLM	Dice (%)	75.8	74.2	76.0	73.5	72.0	
	C-Shift	1.38	1.78	2.80	3.30	4.61	**0.002**
	A-Error	0.305	0.252	0.222	0.247	0.283	

Table 2 explains why PhysiVLM's overall HD95 (143.1) remains higher than that of M3D-LaMed despite better Dice. PhysiVLM attains the best small-lesion Dice and precision, but slight boundary deviations or disconnected false positives on these targets disproportionately affect HD95. Table 3 nevertheless shows that drift grows much more slowly than with single-anchor propagation: at >10 slices, C-Shift drops from 33.51 to 4.61 and A-Error from 10.62 to 0.283. The remaining gap therefore reflects boundary outliers on difficult small targets rather than systematic long-range drift. We claim improved overlap and lesion coverage, not universal boundary superiority.

Qualitative results (Fig. 4) show propagation errors for MedSAM2 and inconsistent masks for thin or irregular lesions with SegVol and M3D-LaMed. PhysiVLM combines depth localization, spatial prompting, and anchor selection to reduce cross-slice drift and preserve Z-axis continuity.

Computational Efficiency Comparison.

Table 4 compares segmentation performance, supported resolution, and per-case input memory. Because the models have different parameter scales, the reported VRAM is the *peak memory used to*

Table 4: **Computational Efficiency Comparison.** We report peak evaluation memory (VRAM), segmentation performance (Dice), and maximum supported input resolution.

Method	VRAM (GB)	Dice (%)	Max Resolution
VILA-M3	0.24	33.3	384×384
R1Seg-3D	1.05	56.4	$32 \times 256 \times 256$
M3D-LaMed (3D)	1.053	69.4	$32 \times 256 \times 256$
PhysiVLM (Ours)	0.43	**73.2**	**Full Resolution**

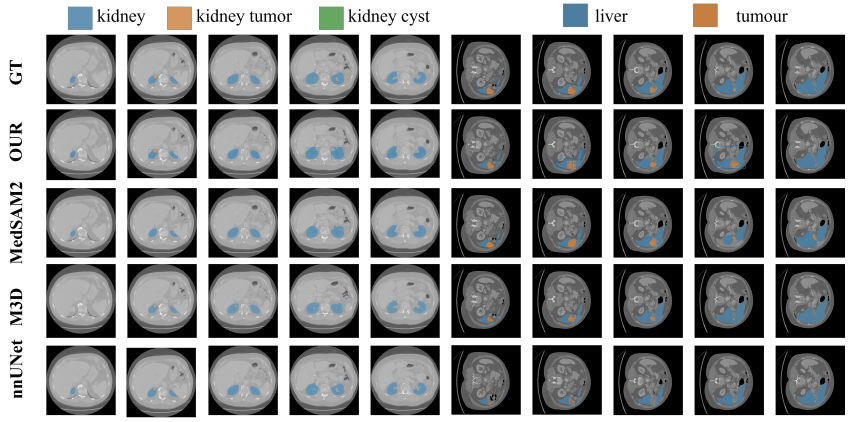


Figure 4: Qualitative comparison of 3D segmentation results on challenging cases.

load the input volume, not total end-to-end GPU memory; it excludes model weights and runtime buffers. This protocol isolates the memory pressure introduced by high-resolution 3D inputs.

R1Seg-3D and M3D-LaMed typically operate on $32 \times 256 \times 256$ crops, which restrict global context. PhysiVLM instead preserves full resolution along the Z axis with moderate input-memory cost and reaches 73.2% Dice. This input-memory advantage does not imply lower end-to-end cost: autoregressive generation increases inference latency. Together with Tables 2 and 3, these results show a clear trade-off: PhysiVLM improves overlap and long-range stability, but boundary outliers and sequential latency remain. PhysiVLM therefore prioritizes full-resolution global context over real-time speed.

Stage 1 Localization Performance.

We separately evaluate Stage 1 text-conditioned Z-interval prediction. Because medical segmentation lacks a direct baseline for this objective, we adapt the video moment-retrieval protocol of Moment-DETR Lei et al. (2021) from time to physical depth. PhysiVLM achieves 72.3% R1@0.3 on the full validation set; this metric diagnoses coarse range retrieval and does not replace mask-level evaluation.

4.3 Ablation Study

MSD ablations quantify incremental component gains (Table 5) and isolate input, depth encoding, and localization choices (Table 6).

Global Locator and 2.5D Grid. The decoupled baseline with default MedSAM2 propagation reaches 65.4% Dice. Stage 1 predicts $[\hat{z}_{\text{start}}, \hat{z}_{\text{end}}]$, removes irrelevant background, and raises Dice to 67.3%. The 2.5D grid further summarizes cross-slice context at low token cost and outperforms long single-slice input; supplementary ablations provide additional variants.

Spatial Prompting. Ordinal slice indices cannot represent physical depth. PhysiVLM remains stable across 1–3 mm resampling (Fig. 5). Adding the physical token $\langle z \rangle$ and anatomical priors increases Dice to 70.6% and lowers HD95 from 167.3 to 154.9 (Table 5), confirming more reliable spatial perception under variable slice thickness.

Table 5: **Ablation Study of PhysiVLM Core Components.** The reported VRAM values correspond to the peak memory of input data loading for one case, rather than the total end-to-end GPU memory consumption of the full model during inference.

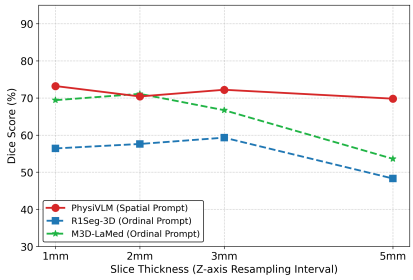
Decoupled	Global Loc.	Spatial Prompt	Guidance	Dice ↑	HD95 ↓
✓	–	–	M2-P	65.4	132.4
✓	✓	–	M2-P	67.3	167.3
✓	✓	✓	M2-P	70.6	154.9
✓	✓	✓	V-G (Ours)	**73.2**	**143.1**

VLM-Driven Guidance and Oracle Localization Analysis. Default MedSAM2 propagation (M2-P) relies on memory matching and can drift as appearance changes. VLM-guided prompting (V-G) instead performs semantic slice localization, reaching 73.2% Dice and 143.1 HD95. Replacing both Stage 1 intervals and Stage 2 boxes with ground-truth cues raises Dice to 85.2% and lowers HD95 to 66.8 (Table 6). This Oracle measures the upper bound of the complete localization pipeline, not Stage 1 alone; Stage 2 box and anchor errors also contribute.

Table 6: **Additional MSD ablations.** * marks the default. *Oracle* replaces Stage 1 intervals and Stage 2 boxes with ground-truth cues.

Component	Variant	Dice ↑	HD95 ↓
Input Format	Plain 2D	65.5	182.4
	Slice Stacking	67.3	167.3
	2.5D Grid*	**73.2**	**143.1**
Depth Enc.	Ordinal Index	69.4	160.2
	Physical $\langle z \rangle$*	**73.2**	**143.1**
Anchor & Loc.	Center Base	68.8	175.6
	M2-P	70.6	154.9
	V-G*	**73.2**	**143.1**
	Oracle (GT)	*85.2*	*66.8*

Figure 5: Z-axis resampling analysis. Denser sampling preserves context at small spacing, while performance remains stable from 1 to 3 mm, indicating reduced sensitivity to spacing variation.



4.4 Discussion

PhysiVLM separates physical-space reasoning from pixel-level mask generation in 3D-as-Video segmentation.

Physical Spatial Reasoning. PhysiVLM encodes absolute Z coordinates with $\langle z \rangle$ tokens instead of treating slices as uniformly sampled frames, improving robustness to resampling and unseen thickness.

Efficiency and Practicality. The two-stage pipeline localizes the relevant physical interval before mask generation, reducing the input-memory burden of full-resolution processing. This design complements the segmentation gains in Table 4, while the sequential VLM decoding remains a computational trade-off.

Limitations. MedSAM2 limits boundary accuracy, as reflected by HD95, and autoregressive reasoning increases inference latency. Faster decoding and end-to-end physical-coordinate integration remain future work.

AI Use Statement

In this work, we did not use generative AI tools for any research tasks requiring disclosure under the ICLR AI policy, including developing theoretical or conceptual frameworks, formulating or refining hypotheses, designing research methodology or experiments, implementing methods, processing or analyzing data, or interpreting experimental results. These aspects of the work were carried out by the authors.

We used generative AI tools solely for language editing of the manuscript, including improving grammar, clarity, readability, and academic phrasing. Generative AI tools were not used to generate, modify, or analyze experimental data, quantitative results, scientific figures, or images. All AI-assisted textual edits were reviewed by the authors to ensure that they did not alter the scientific meaning or claims of the manuscript. We take full responsibility for the final content of this work, including all text, results, figures, and conclusions.

Reproducibility statement

We have made substantial efforts to ensure the reproducibility of this work. The complete implementation of PhysiVLM, including the source code, inference pipeline, and evaluation scripts, is anonymously available at `https://anonymous.4open.science/r/PhysiVLM-1E33`. The main paper describes the overall model architecture, training and inference procedures, and experimental setup, while additional implementation details, hyperparameter settings, dataset processing procedures, and supplementary experiments are provided in the appendix. We use publicly available datasets where applicable and report the corresponding data sources and evaluation protocols in the paper. Together, these materials are intended to facilitate independent reproduction and verification of the reported results.

References

Fan Bai, Yuxin Du, Tiejun Huang, Max Q. H. Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models, 2024. URL https://arxiv.org/abs/2404.00578.

Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, et al. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, pp. 103280, 2024.

Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis, 2019. URL https://arxiv.org/abs/1904.00625.

An-Chieh Cheng, Yang Fu, Yukang Chen, Zhijian Liu, Xiaolong Li, Subhashree Radhakrishnan, Song Han, Yao Lu, Jan Kautz, Pavlo Molchanov, Hongxu Yin, Xiaolong Wang, and Sifei Liu. 3d aware region prompted vision language model, 2025. URL https://arxiv.org/abs/2509.13317.

Yuxin Du, Fan Bai, Tiejun Huang, and Bo Zhao. Segvol: Universal and interactive volumetric medical image segmentation. *arXiv preprint arXiv:2311.13385*, 2023.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, Weicheng Dai, Murong Xu, Hadrien Reynaud, Muhammed Furkan Dasdelen, Bastian Wittmann, Tamaz Amiranashvili, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, Abdullah Alanbay, Anjany Sekuboyina, Berkan Lafci, Ahmet Kaplan, Zhiyong Lu, Malgorzata Polacin, Bernhard Kainz, Christian Bluethgen, Kayhan Batmanghelich, Mehmet Kemal Ozdemir, and Bjoern Menze. Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature Biomedical Engineering*, February 2026. ISSN 2157-846X. doi: 10.1038/s41551-025-01599-y. URL http://dx.doi.org/10.1038/s41551-025-01599-y.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation, 2021. URL https://arxiv.org/abs/2103.10504.

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, January 2022.

Nicholas Heller, Fabian Isensee, Dasha Trofimova, Resha Tejpaul, Zhongchen Zhao, Huai Chen, Lisheng Wang, Alex Golts, Daniel Khapun, Daniel Shats, Yoel Shoshan, Flora Gilboa-Solomon,

Yasmeen George, Xi Yang, Jianpeng Zhang, Jing Zhang, Yong Xia, Mengran Wu, Zhiyang Liu, Ed Walczak, Sean McSweeney, Ranveer Vasdev, Chris Hornung, Rafat Solaiman, Jamee Schoephoerster, Bailey Abernathy, David Wu, Safa Abdulkadir, Ben Byun, Justice Spriggs, Griffin Struyk, Alexandra Austin, Ben Simpson, Michael Hagstrom, Sierra Virnig, John French, Nitin Venkatesh, Sarah Chan, Keenan Moore, Anna Jacobsen, Susan Austin, Mark Austin, Subodh Regmi, Nikolaos Papanikolopoulos, and Christopher Weight. The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct, 2023.

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

Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F Jaeger. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 488–498. Springer, 2024.

Yuanfeng Ji, Haotian Bai, Jie Yang, Chongjian Ge, 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. *arXiv preprint arXiv:2206.08023*, 2022.

Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, Oriane Siméoni, Huy V. Vo, Patrick Labatut, and Piotr Bojanowski. Dinov2 meets text: A unified framework for image- and pixel-level vision-language alignment, 2024.

A. Emre Kavur, N. Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, Bora Baydar, Dmitry Lachinov, Shuo Han, Josef Pauli, Fabian Isensee, Matthias Perkonigg, Rachana Sathish, Ronnie Rajan, Debdoot Sheet, Gurbandurdy Dovletov, Oliver Speck, Andreas Nürnberger, Klaus H. Maier-Hein, Gözde Bozdağı Akar, Gözde Ünal, Oğuz Dicle, and M. Alper Selver. CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, April 2021. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2020.101950. URL http://www.sciencedirect.com/science/article/pii/S1361841520303145.

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

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model, 2024. URL https://arxiv.org/abs/2308.00692.

Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, pp. 12, 2015.

Jie Lei, Tamara L. Berg, and Mohit Bansal. Qvhighlights: Detecting moments and highlights in videos via natural language queries, 2021. URL https://arxiv.org/abs/2107.09609.

Chenxin Li, Xinyu Liu, Wuyang Li, Cheng Wang, Hengyu Liu, and Yixuan Yuan. U-kan makes strong backbone for medical image segmentation and generation. *arXiv preprint arXiv:2406.02918*, 2024a.

Hao Li, Han Liu, Dewei Hu, Jiacheng Wang, and Ipek Oguz. Promise: Prompt-driven 3d medical image segmentation using pretrained image foundation models. In *2024 IEEE international symposium on biomedical imaging (ISBI)*, pp. 1–5. IEEE, 2024b.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42: 60–88, 2017. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2017.07.005. URL https://www.sciencedirect.com/science/article/pii/S1361841517301135.

Chenyang Ma, Kai Lu, Ta-Ying Cheng, Niki Trigoni, and Andrew Markham. Spatialpin: Enhancing spatial reasoning capabilities of vision-language models through prompting and interacting 3d priors. *Advances in neural information processing systems*, 37:68803–68832, 2024a.

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

Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang. Medsam2: Segment anything in 3d medical images and videos. *arXiv preprint arXiv:2504.03600*, 2025.

Qinhao, Long Yu, Shengwei Tian, Xujiong Ye, and Lei Zhang. R1seg-3d: Rethinking reasoning segmentation for medical 3d cts, 2025.

Blaine Rister, Darvin Yi, Kaushik Shivakumar, Tomomi Nobashi, and Daniel L Rubin. Ct-org, a new dataset for multiple organ segmentation in computed tomography. *Scientific Data*, 7(1):381, 2020.

Maximilian Rokuss, Moritz Langenberg, Yannick Kirchhoff, Fabian Isensee, Benjamin Hamm, Constantin Ulrich, Sebastian Regnery, Lukas Bauer, Efthimios Katsigiannopulos, Tobias Norajitra, et al. Voxtell: Free-text promptable universal 3d medical image segmentation. *arXiv preprint arXiv:2511.11450*, 2025.

Pengcheng Shi, Jiawei Chen, Jiaqi Liu, Xinglin Zhang, Tao Chen, and Lei Li. Medal s: Spatio-textual prompt model for medical segmentation. *arXiv preprint arXiv:2511.13001*, 2025.

Amber L. Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram van Ginneken, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc Gollub, Jennifer Golia-Pernicka, Stephan H. Heckers, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Eugene Vorontsov, Lena Maier-Hein, and M. Jorge Cardoso. A large annotated medical image dataset for the development and evaluation of segmentation algorithms, 2019. URL https://arxiv.org/abs/1902.09063.

Jirui Tian, Jinrong Zhang, Shenglan Liu, Luhao Xu, Zhixiong Huang, and Gao Huang. Dtos: Dynamic time object sensing with large multimodal model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13810–13820, 2025. doi: 10.1109/CVPR52734.2025.01289.

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

Xiaofan Zhang Kang Li Shaoting Zhang Wenhui Lei, Xu Wei. Medlsam: Localize and segment anything model for 3d medical images. *arXiv preprint arXiv:*, 2023.

Yu Xin, Gorkem Can Ates, Kuang Gong, and Wei Shao. Med3dvlm: An efficient vision-language model for 3d medical image analysis. *IEEE Journal of Biomedical and Health Informatics*, 2025.

Yang Xing, Jiong Wu, Savas Ozdemir, Ying Zhang, Yang Yang, Wei Shao, and Kuang Gong. Medvlsam2: A unified 3d medical vision-language model for multimodal reasoning and prompt-driven segmentation, 2026. URL https://arxiv.org/abs/2601.09879.

Hongbo Ye, Fenghe Tang, Peiang Zhao, Zhen Huang, Dexin Zhao, Minghao Bian, and S. Kevin Zhou. U-rwkv: Lightweight medical image segmentation with direction-adaptive rwkv, 2025. URL https://arxiv.org/abs/2507.11415.

Feng Yu, Jiacheng Cao, Li Liu, and Minghua Jiang. Superlightnet: Lightweight parameter aggregation network for multimodal brain tumor segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5197–5206, 2025. doi: 10.1109/CVPR52734.2025.00490.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

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

Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation, 2022. URL https://arxiv.org/abs/2109.03201.

Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE Transactions on Image Processing*, 32:4036–4045, 2023. doi: 10.1109/TIP.2023.3293771.

A Appendix A: Implementation Details and Clarifications

Our entire project is open-sourced at https://anonymous.4open.science/r/PhysiVLM-1E33

A.1 Gradient Flow of the Mask Loss

In Stage 2, we use MedSAM2 as a frozen segmentation expert and generate segmentation masks from the bounding boxes predicted by the upstream model. A key point is that, although MedSAM2 is frozen at the parameter level, that is, `requires_grad=False`, the gradient of the mask loss with respect to the box coordinates still exists and can flow back through the computation graph to the upstream box decoder and the LoRA parameters in the VLM. In other words, freezing parameters does not mean cutting off the gradient path. As long as the input prompts are not explicitly `detach`ed during the forward pass, and the intermediate mappings are differentiable, the loss can still produce gradients with respect to the input prompts. This point is illustrated in Figure 6. :contentReference[oaicite:0]index=0

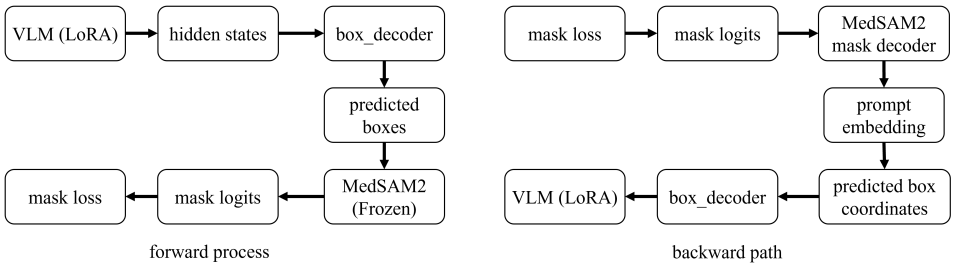


Figure 6: Forward and backward paths in Stage 2. MedSAM2 remains frozen, but the mask loss is still differentiable with respect to the predicted box coordinates and can therefore supervise the upstream prompt generation modules.

The corresponding gradient path can be written as

$$\frac{\partial \mathcal{L}_{\text{mask}}}{\partial \theta_{\text{VLM}}} = \frac{\partial \mathcal{L}_{\text{mask}}}{\partial M} \cdot \frac{\partial M}{\partial B} \cdot \frac{\partial B}{\partial \theta_{\text{VLM}}}, \tag{12}$$

where M denotes the mask logits, B denotes the predicted box coordinates, and θ_{VLM} denotes the trainable upstream parameters, including the LoRA modules and the box decoder. Since $B = (x_1, y_1, x_2, y_2)$ is continuous, and both the prompt construction from B and the subsequent mask decoding in MedSAM2 are differentiable, we have

$$\frac{\partial \mathcal{L}_{\text{mask}}}{\partial B} \neq 0. \tag{13}$$

As a result, the mask-level supervision does not only affect the segmentation output itself. It also constrains box prediction, which encourages the upstream modules to generate prompts that are more useful for accurate segmentation. :contentReference[oaicite:1]index=1

In our implementation, the parameters of MedSAM2 are frozen and are therefore not updated during training. At the same time, we do not apply `detach` to the input box prompts, so the prompt embeddings generated from the box coordinates remain inside the computation graph. The gradient can thus pass through this frozen module to the upstream input side, while the MedSAM2 parameters remain unchanged. This design allows the model to use high-quality mask-level supervision from a strong segmentation expert, while avoiding the computational cost and optimization instability of fully joint end-to-end training. :contentReference[oaicite:2]index=2

The full system is not fully differentiable. It is more accurate to describe it as a partially differentiable system, because several discrete operations do not participate in gradient propagation. These operations include Hungarian matching for prediction-to-ground-truth assignment, non-maximum suppression, IoU-based filtering, anchor slice selection, and invalid-box filtering. For example,

```
correct_pred_box = pred_box[~error_mask]
```

is used only to remove invalid boxes and does not introduce a differentiable relaxation. Therefore, the differentiability we refer to in this paper specifically concerns the supervision path from the mask loss to the continuous box coordinates and then to the trainable upstream modules, rather than every component in the full inference and post-processing pipeline. :contentReference[oaicite:3]index=3

This design differs from a conventional cascaded pipeline. In a standard setting, the segmentation module is often treated as an independent post-processing stage,

```
box → segmentation (detach)
```

so segmentation errors do not constrain box prediction. In our method, the pipeline instead follows

```
box → segmentation → loss → backpropagation to box
```

which means that the segmentation model acts as a differentiable supervision source rather than a pure post-processing module. In this sense, MedSAM2 serves as a frozen expert supervisor: its parameters remain fixed, but the loss from its output can still shape the upstream prompt generation process without fine-tuning the segmentation model itself. Overall, the key advantage of Stage 2 is that MedSAM2 provides high-quality mask supervision, and this supervision can directly optimize prompt generation, that is, box prediction, through a differentiable path, without introducing the cost of large-scale joint training. :contentReference[oaicite:4]index=4

A.2 Stage 1 Modeling and Hungarian Matching

Stage 1 follows a DETR-style set prediction framework. Given the hidden representation associated with the query, the model predicts a fixed number of candidate intervals through the output `<rec>` tokens. In our implementation, the number of candidates is

$$K = 10. \tag{14}$$

Although most training samples contain only one target interval, that is,

$$M_{\mathrm{gt}} = 1, \tag{15}$$

we still use the set prediction formulation so that the model learns to produce the best candidate among multiple predictions while naturally using the unmatched candidates as background negatives. :contentReference[oaicite:5]index=5

The assignment between predictions and ground-truth targets is computed with Hungarian matching. For a candidate prediction p and a ground-truth interval g, the matching cost is defined as

$$\mathrm{cost} = \lambda_1 \|p - g\|_1 + \lambda_2 \big(1 - \mathrm{IoU}(p, g)\big) + \lambda_3 \big(1 - p_{\mathrm{fg}}\big), \tag{16}$$

where $\|p - g\|_1$ measures the regression error, $\mathrm{IoU}(p, g)$ measures the overlap between the predicted and ground-truth intervals, and p_{fg} denotes the foreground confidence of the candidate. After matching, one prediction is assigned as foreground, and the remaining $K - 1$ predictions are treated as background for classification supervision. The Stage 1 training objective is therefore

$$\mathcal{L}_{\mathrm{stage1}} = \lambda_1 \mathcal{L}_{\mathrm{L1}} + \lambda_2 \mathcal{L}_{\mathrm{IoU}} + \lambda_3 \mathcal{L}_{\mathrm{cls}}. \tag{17}$$

We use this design for three reasons. First, Hungarian matching avoids heuristic selection based on `argmax` during training and provides a unified one-to-one assignment between predictions and targets. Second, unmatched candidates naturally provide negative supervision, which improves both the classification head and the overall training stability. Third, although the current task mainly involves single-target interval prediction, the same formulation can be extended to future multi-target settings without changing the overall architecture. :contentReference[oaicite:6]index=6

A.3 Experimental Settings and Reproducibility

All experiments are conducted on 8 NVIDIA H800 GPUs with bf16 mixed precision, using PyTorch and DeepSpeed. The optimization is divided into two stages. For Stage 1, the learning rate is set to 5×10^{-5}, the batch size is 1, and the total number of training epochs is about 15000. For Stage 2, the learning rate is 2×10^{-5}, the batch size is also 1, and the total number of training epochs is 15000. :contentReference[oaicite:7]index=7

For parameter updates, we train only lightweight task-specific modules, including LoRA with rank 64, the box/<rec> decoders, and the task-specific token embeddings. The visual encoder and MedSAM2 remain frozen throughout training. This design keeps the general representation ability of the pretrained visual backbone while adapting the model through parameter-efficient modules. For MedSAM2, we use it only as a fixed segmentation expert that provides reliable mask-level supervision and is not jointly optimized. :contentReference[oaicite:8]index=8

For data preprocessing, we first convert each MRI volume into slices and organize each sample as a three-channel input corresponding to the previous, current, and next slices, which preserves local 3D context. All inputs are normalized with per-sample min-max normalization to reduce intensity variation across cases. For evaluation, Stage 1 uses R1@0.3 to measure interval localization, while Stage 2 uses Dice, HD95, and IoU to measure segmentation quality. During model selection, we do not use validation-driven early stopping. Instead, we save checkpoints periodically and select the final model through offline evaluation. To reduce randomness, we fix the random seed in all experiments. :contentReference[oaicite:9]index=9

Table 7: Basic training configuration.

Parameter	Value
per_device_train_batch_size	1
gradient_accumulation_steps	1
lr_scheduler_type	cosine
warmup_ratio	0.03
weight_decay	1×10^{-4}
bf16	True
tf32	True
gradient_checkpointing	True
evaluation_strategy	no validation
dataloader_num_workers	4 (1 for Stage 2)

Table 8: Stage 1 training hyperparameters.

Parameter	Value
learning_rate	5×10^{-5}
num_train_epochs	15000
n_frms	10 (uniform sampling)

Table 9: Stage 2 training hyperparameters.

Parameter	Value
learning_rate	2×10^{-5}
num_train_epochs	1500
save_steps	2500
n_frms	10

We also note that the dynamic sampling in Stage 2 depends on the interval predicted by Stage 1, and all experiments use the same prompt template for fair comparison. :contentReference[oaicite:10]index=10

A.4 Prompt Templates and Training Configuration

A.4.1 Prompt Templates

We use fixed instruction templates in both Stage 1 and Stage 2 to keep the training and testing settings consistent.

Stage 1 (Slice Interval Localization). The user input template is

```
Can you locate the slice ranges corresponding to this clinical description
in the MRI scan?
{caption}
```

where $\{caption\}$ denotes the original medical description from the dataset, and each lesion corresponds to an independent dialogue. The model output template is

```
Considered, I would pinpoint the target MRI slice ranges as:
<rec>...<rec>.
My analysis yields detailed answers above.
```

Each sample always generates 10 <rec> tokens. This number is independent of the true number of intervals and corresponds to the candidate slots in the set prediction design, that is, $K = 10$. :contentReference[oaicite:11]index=11

Stage 2 (Slice-Wise Segmentation). For slices that contain the target lesion, we use the following output template:

```
I would show the following targeted locations in the <z_k>
slice from this MRI scan:
<box> ... <box>. The above marks pathological regions.
```

Here, <z_k> denotes the spatial index token of the current slice, with values ranging from <z_1> to <z_10>. The number of <box> tokens is set to match the number of ground-truth (GT) boxes in the current slice. Each <box> token is then decoded by the box regression head into a candidate bounding box.

For negative slices that do not contain the target, we use the following output template:

```
I would show the following targeted locations in the <z_k> slice
from this MRI scan:
No match for the target. The above marks pathological regions.
```

A.5 Quantitative Analysis of Mask Drift

In the main paper, we argue that single-anchor sequential propagation tends to accumulate errors over a long axial range, which gradually shifts the predicted mask away from the true target. To test this mechanism directly, rather than only through volume-level segmentation scores, we design a distance-stratified analysis to measure mask drift more explicitly.

A.5.1 A.6.1 Motivation

Standard segmentation metrics, such as Dice and IoU, are usually averaged over the full volume and therefore cannot reveal how errors accumulate along the propagation path. We therefore define mask drift as the systematic degradation of segmentation quality as the distance between a test slice and its prompt anchor increases.

If this effect exists, slices farther from the prompt anchor should show larger geometric deviation and lower segmentation quality. If the multi-anchor design is effective, this distance-dependent degradation should be substantially weaker.

A.5.2 Compared Methods

We compare two settings. The first is **MedSAM2-Single**, which uses a single bounding box on the center slice and applies the default bidirectional propagation. The second is **PhysiVLM (Full)**, which uses the full multi-anchor initialization and aggregation strategy proposed in this paper. The first setting follows a single-anchor propagation scheme, whereas the second shortens the effective propagation distance through multiple anchors.

A.5.3 DISTANCE DEFINITION

For any test slice z, we define its distance to the nearest prompt anchor as

$$d(z) = \min_{a \in \mathcal{A}} |z - a|, \tag{18}$$

where $\mathcal{A}$ denotes the set of anchor slices used by the method. For the single-anchor setting, $\mathcal{A}$ contains only one element. For our method, $\mathcal{A}$ contains multiple anchors.

A.5.4 EVALUATION METRICS

We report the following slice-level metrics: Dice@d, which is the Dice score grouped by distance; IoU@d, which is the IoU grouped by distance; Centroid Shift@d, which is the Euclidean distance between the centroids of the predicted mask and the ground-truth mask; and Area Error@d, which is the relative error in mask area.

The distance bins are defined as

$$\{0,\ 1\text{–}2,\ 3\text{–}5,\ 6\text{–}10,\ >10\}.$$

To further quantify how performance changes with distance, we fit a linear model between Dice and propagation distance,

$$\mathrm{Dice}(z) = \alpha - \beta\, d(z), \tag{19}$$

and use $|\beta|$ as the **drift slope**. A larger $|\beta|$ indicates a stronger decline in segmentation quality as propagation distance increases, and therefore more severe mask drift.

A.5.5 EXPERIMENTAL RESULTS

We conduct this analysis on the MSD dataset. For each test volume, we first compute the distance from each slice to its nearest prompt anchor according to the anchors used during inference. We then group the slices by distance and average the segmentation results over the full test set. The results are shown in Table 10.

Table 10: Distance-stratified analysis of mask drift. We group test slices by their distance to the nearest prompt anchor. Higher Dice and IoU indicate better performance, while lower Centroid Shift, Area Error, and Drift Slope indicate weaker drift. Drift Slope ($|\beta|$) is obtained by fitting slice-level Dice against propagation distance.

Method	Metric	0	1–2	3–5	6–10	>10	Drift Slope ($\lvert\beta\rvert$) ↓
Single-Anchor	Dice (%)	72.5	60.3	53.3	54.2	69.9	0.00626
	IoU (%)	62.3	51.5	46.9	50.6	66.3	
	Centroid Shift ↓	2.98	11.21	14.26	19.56	33.51	
	Area Error ↓	0.249	1.04	1.66	3.25	10.62	
PhysiVLM (Full)	Dice (%)	75.8	74.2	76.0	73.5	72.0	0.00249
	IoU (%)	65.9	69.6	71.3	70.2	68.4	
	Centroid Shift ↓	1.38	1.78	2.80	3.30	4.61	
	Area Error ↓	0.305	0.252	0.222	0.247	0.283	

A.5.6 ANALYSIS

Table 10 shows a clear difference in cross-slice stability between single-anchor propagation and our multi-anchor strategy.

First, **single-anchor propagation shows clear mask drift**. As the distance between the test slice and the prompt anchor increases, both Centroid Shift and Area Error rise steadily in the Single-Anchor setting. This pattern indicates that the predicted mask gradually deviates from the true target in both position and size, which is consistent with cumulative propagation error. In contrast, our full model

keeps both Centroid Shift and Area Error much lower across all distance bins, which suggests that the multi-anchor design effectively constrains geometric drift during propagation.

Second, **the multi-anchor strategy weakens the distance-related performance decay**. PhysiVLM (Full) has a much smaller drift slope than Single-Anchor, which means that its performance changes less as propagation distance increases. This result supports the main intuition of our method: shortening the effective propagation path through multi-anchor initialization reduces error accumulation along the Z axis.

The table also shows that Dice and IoU for Single-Anchor decrease at intermediate distances and then partially recover in the > 10 bin. This pattern does not mean that long-range propagation becomes more stable. A more plausible explanation is the interaction between anchor position and lesion morphology along the axial direction. In MRI volumes, lesions often have larger cross-sectional area in the middle slices and smaller extent near the two ends. Slices near the lesion boundary are therefore usually smaller, sparser, and harder to segment. When the single anchor is located closer to the lesion boundary, the slices that fall into the > 10 bin may include more slices near the lesion body, which can raise the mean Dice value in that bin. However, this local recovery does not change the overall drift pattern. The geometric metrics, especially Centroid Shift and Area Error, still increase substantially in the same bin, which indicates that the predicted masks continue to deviate from the true target. In other words, the non-monotonic change in Dice for a few distance bins is more likely caused by the coupling between axial lesion morphology and anchor placement than by genuine long-range stability in the single-anchor setting.

By contrast, PhysiVLM (Full) remains much more stable across distance bins in Dice, IoU, Centroid Shift, and Area Error. This result suggests that the multi-anchor strategy improves not only overall segmentation accuracy but also cross-slice consistency and robustness.

A.5.7 SUMMARY

This experiment formalizes mask drift as a distance-dependent degradation in segmentation quality and geometric accuracy, and verifies this effect through slice-level stratified analysis. The results show that single-anchor propagation is more prone to positional drift and area distortion over long distances, whereas the proposed multi-anchor strategy substantially reduces this error accumulation and improves the stability and robustness of cross-slice segmentation.