PATCHWISE 3D CONTEXT EMBEDDING: EQUIPPING EFFICIENT 2D FOUNDATION MODELS WITH 3D PERCEPTION FOR BRAIN HEMORRHAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Brain hemorrhage segmentation is clinically important on non-contrast computed tomography (CT), but 3D models are costly while 2D foundation models lack cross-slice context. Here, we ask whether a 2D foundation model can be efficiently equipped with the missing 3D prior for brain hemorrhage segmentation, without relying on computationally expensive dense 3D or video-style volumetric processing. We propose Patchwise 3D Context Embedding (P3CE), an efficient approach for introducing 3D prior into 2D foundation models. Instead of densely processing the full volume, P3CE represents local 3D neighborhoods as sparse point-cloud tokens, allowing information from adjacent axial slices to be captured efficiently. We pre-trained a brain-specific Point-MAE on CT volumes to capture volumetric context. Its frozen encoder generates attributed 3D patch tokens once per case for slice-wise segmentation. A slice-conditioned retrieval module selects relevant tokens for each axial slice and injects them through lightweight cross-attention and residual fusion. Segmentation remains slice-wise, with 3D context encoded once. We evaluated P3CE on Seg-CQ500 and CT-ICH. P3CE achieved the highest overlap accuracy among the compared methods (0.5686/0.4773 IoU, 0.7078/0.6343 Dice), improving Dice by 9.25 points over the SAM-Med2D baseline on Seg-CQ500 while adding only a single frozen point-encoding pass per volume, with notable improvements on small and low-contrast hemorrhages and competitive inter-slice continuity. Ablations confirmed the contributions of both the 3D point-cloud representation and point attributes. These results show that 2D foundation models can benefit from volumetric information without dense 3D computation. By separating the computation unit from the context unit, P3CE provides an efficient interface for introducing sparse 3D information into slice-wise foundation models, suggesting a potential route for extending them to other medical imaging tasks where 3D context is important, pending broader validation. This study also provides a pre-trained, brain-specific Point-MAE for head CT.

1 INTRODUCTION

Intracranial hemorrhage (ICH) is one of the most lethal forms of stroke and demands fast, accurate delineation of bleeding on non-contrast CT to guide intervention (van Asch et al., 2010; An et al., 2017; Heit et al., 2017). Small hemorrhagic lesions, including small-volume bleeds, microhemorrhages, and early subtle bleeding, are the most clinically consequential and the most frequently missed. When observed on an isolated slice, these lesions are easily confused with calcification, small vessels, the choroid plexus, or beam-hardening artifacts. Radiologists resolve this ambiguity by scrolling through adjacent slices and checking whether the structure persists and evolves like a lesion. This clinical workflow underscores a general requirement: accurate hemorrhage segmentation depends on 3D context that no single axial slice can provide.

Foundation models are rapidly scaling up (Wang et al., 2023), but they face a fundamental trade-off among computational cost, training data availability, and model performance: compared with 3D foundation models, 2D foundation models have access to broader pretraining data and require less computational cost. However, 2D foundation models lack good 3D priors. So the intuition of this work is straightforward: **can 2D foundation models be given the 3D context they lack?**

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.

Specifically, in medical foundation models, this trade-off is particularly evident for volumetric segmentation. SAM-Med3D (Wang et al., 2023) extends the Segment Anything Model (SAM) (Kirillov et al., 2023) paradigm to volumetric medical data, but inherits the dense computational cost of 3D processing. In contrast, 2D medical segmentation foundation models, such as MedSAM (Ma et al., 2024), SAM-Med2D (Cheng et al., 2023), and SAM2-UNet (Xiong et al., 2026), benefit from broader 2D pretraining, but most process CT slice by slice—either independently or in a video-style manner—and therefore lack the 3D context needed for ambiguous hemorrhagic lesions. Existing attempts to introduce volumetric context only partially address this limitation. 2.5D approaches stack a few adjacent slices as input channels, but the context window remains narrow and overlapping slices are repeatedly re-encoded. Video-based adaptations such as MedSAM2 (Ma et al., 2025) propagate information along the slice sequence through memory attention, coupling accuracy with inference order while remaining computationally redundant. This landscape therefore raises the question we will focus in this work: **can a 2D foundation model be adapted to brain hemorrhage segmentation by supplying the missing 3D prior without replacing it with a dense 3D network?**

We answer this question by changing the carrier of volumetric information from dense voxels to sparse points. Our key observation is that a local patch in 3D point space is not a collection of points restricted to one image plane: when patch centers are selected in the three-dimensional coordinate space and neighboring points are grouped around each center, a patch naturally aggregates information across adjacent axial levels. A single patch token therefore summarizes a local 3D region, including within-slice structure and cross-slice variation. Building on this observation, we propose *Patchwise 3D Context Embedding* (P3CE), an efficient approach for introducing volumetric context into 2D foundation models. For head CT, we first pre-train a brain-specific Point-MAE (Pang et al., 2022) to learn reusable 3D anatomical representations from attributed point clouds. Once pre-trained, the frozen encoder generates 3D patch tokens only once for each CT volume. A slice-conditioned retrieval module selects relevant tokens for each axial slice and injects them through lightweight cross-attention and residual fusion. Segmentation remains slice-wise, with 3D context encoded once. This design separates the computation unit (2D slice) from the context unit (3D patch): the point encoder runs once per volume, its tokens are reused by all slices, and the fusion cost depends on the number of image tokens and the small retrieved token set rather than the voxel count.

Our contributions are summarized as follows:

- We introduce a framework for supplying 3D context to 2D foundation models by changing the carrier of volumetric information from dense voxels to sparse 3D patch tokens. This allows volumetric priors to be incorporated without replacing efficient 2D processing with dense 3D computation.
- We pre-train a brain-specific Point-MAE on head-CT for reusable 3D anatomical priors. The resulting encoder provides compact volumetric representations that can be frozen and coupled with different slice-wise segmentation models, illustrating how anatomy-specific 3D priors can be learned independently of the downstream 2D foundation model.
- We propose P3CE, which combines slice-conditioned token retrieval with lightweight cross-attention and residual fusion. By separating the computation unit (2D slice) from the context unit (3D patch), volumetric information is encoded once per volume and reused across all slices.
- We validate P3CE on intracranial hemorrhage segmentation using Seg-CQ500 and CT-ICH, where it achieves the highest overlap accuracy among the compared supervised and foundation-model-based methods, with particularly pronounced improvements for small and low-contrast hemorrhages. Ablation studies further quantify the contributions of the 3D representation, point attributes, and individual framework components.

2 Related Works

2D Medical Foundation Models. The Segment Anything Model (SAM) (Kirillov et al., 2023) introduced a promptable, general-purpose segmentation framework pretrained on over one billion masks, demonstrating strong zero-shot transfer to natural images. Its medical adaptations have rapidly proliferated: MedSAM (Ma et al., 2024) fine-tunes SAM on a large-scale collection of

medical image–mask pairs across multiple modalities, while SAM-Med2D (Cheng et al., 2023) adapts the image encoder and prompt encoder to medical domains with adapter-based tuning and boundary-aware loss. SAM2-UNet (Xiong et al., 2026) integrates the SAM2 encoder into a U-Net decoder for both natural and medical segmentation. These models retain the slice-wise inference paradigm: each 2D image is processed independently, which is efficient but lacks cross-slice context. Our approach preserves this slice-wise computation while supplying sparse 3D context, directly addressing this structural limitation.

3D Medical Image Segmentation. Dense volumetric networks process the full 3D context but at substantial computational cost. nnU-Net (Isensee et al., 2021) automates architecture and training configuration for 3D U-Net variants, achieving strong performance across biomedical benchmarks, but its memory and computation scale with voxel count. SAM-Med3D (Wang et al., 2023) extends the SAM paradigm to volumetric medical data with 3D prompts and a 3D image encoder, yet inherits the same dense computation bottleneck and, although 3D pretraining is rapidly scaling up, still cannot leverage the breadth of pretraining data available in 2D. AISCT-SAM (Kuang et al., 2026) customizes SAM-Med2D with 3D context awareness and self-prompt generation for acute ischemic stroke segmentation on CT, adding volumetric modules atop a 2D backbone. In contrast, our approach avoids dense 3D feature construction entirely: the volumetric branch encodes the case once into sparse point tokens, and the fusion cost is independent of voxel count.

2.5D and Video-Based Approaches. A middle ground between 2D and 3D is to stack adjacent slices as input channels (2.5D) or to treat the slice sequence as a video. 2.5D methods enlarge the receptive field beyond a single plane but provide a narrow context window and re-encode overlapping slices redundantly. MedSAM2 (Ma et al., 2025) adapts the SAM2 video framework to 3D medical images, propagating memory attention along the slice axis. While effective, this couples prediction quality with inference order and introduces computational overhead from memory attention. Our approach differs fundamentally: instead of propagating information along a sequential axis, we extract volumetric context from a case-level point representation that is computed once and retrieved per slice, decoupling context extraction from slice ordering.

Point Cloud Representation in Medical Imaging. Point clouds offer a compact, sparse alternative to dense volumetric grids. PointNet and PointNet++ introduced permutation-invariant set processing and hierarchical local grouping for unordered 3D points. Point-MAE (Pang et al., 2022) extends masked autoencoding to point clouds, learning rich local and global representations through self-supervised reconstruction. In medical imaging, point clouds have been used for organ surface modeling, surgical scene understanding, and anatomical shape analysis, but their potential as a carrier of volumetric context for 2D segmentation has not been explored. Our work repurposes Point-MAE as a frozen case-level encoder: each CT volume is converted to an attributed point set and encoded into sparse patch tokens that summarize local 3D neighborhoods, providing a lightweight volumetric interface to the 2D segmentation model.

3 METHOD

Three-dimensional medical image segmentation requires volumetric context, because the appearance and extent of a lesion cannot always be determined from an isolated axial slice. A direct solution is to replace the image encoder with a dense 3D network. However, this substantially increases memory consumption and computation, particularly for high-resolution CT volumes. We instead preserve the efficient slice-wise formulation of the 2D foundation model and introduce a Patchwise 3D Context Embedding (P3CE) module that transfers sparse volumetric information into the 2D feature space.

The key observation is that a local patch in 3D point space is not a collection of points restricted to one image plane. Each patch is formed by grouping neighboring points in the three-dimensional coordinate space and therefore naturally aggregates information across adjacent axial levels. A single patch token consequently summarizes a local 3D region, including within-slice structure and cross-slice variation. We route these volumetric patch tokens to their corresponding axial slices and inject them into the 2D image features through lightweight cross-attention and residual fusion. The

model thus uses a 2D image encoder and a 2D mask decoder, while the information supplied to each slice remains intrinsically three-dimensional.

Let $I \in \mathbb{R}^{H\times W\times D}$ be a CT volume and $B \in \mathbb{Z}^{H\times W\times D}$ the brain-region mask. The image branch processes axial slices $I_z = I(:,:,z)$, whereas the point branch encodes the complete case only once. It outputs G sparse 3D patch tokens and their spatial centers. For each slice z, a slice-conditioned retrieval module selects a small subset of nearby patch tokens, which are then fused with the 2D image features. This design separates the computation unit from the context unit: slices remain the computation unit, while local 3D patches serve as the context unit.

3.1 Patchwise Volumetric Context Encoding

We convert each CT volume into a sparse set of N attributed points sampled from the brain region. Each point contains the three-dimensional coordinates $\mathbf{p}_i = (x_i, y_i, z_i)$, the CT intensity I_i (HU), a brain-tissue label r_i, a source-state indicator s_i, and a health-state indicator h_i, matching the attributed point representation shown in Fig. 1. The categorical attributes are encoded with one-hot mappings $\mathbf{e}_r(\cdot)$, $\mathbf{e}_s(\cdot)$, and $\mathbf{e}_h(\cdot)$, yielding

$$\mathbf{f}_i = \left[\, \mathbf{p}_i^\top,\ I_i,\ \mathbf{e}_r(r_i)^\top,\ \mathbf{e}_s(s_i)^\top,\ \mathbf{e}_h(h_i)^\top \,\right]^\top. \tag{1}$$

The resulting representation combines spatial geometry, CT appearance, brain-tissue identity, source-state information, and health-state information. In our implementation, the encoded point feature contains 18 dimensions: three coordinate values, one HU value, eight brain-tissue categories, three source-state categories, and three health-state categories. The coordinates are normalized by centering at the volume centroid $\boldsymbol{\mu}$ and scaling by the maximum radial distance,

$$\tilde{\mathbf{p}}_i = \frac{\mathbf{p}_i - \boldsymbol{\mu}}{\sigma}, \qquad \sigma = \max_i \|\mathbf{p}_i - \boldsymbol{\mu}\|_2, \tag{2}$$

and the centroid offset and scale factor along the z-axis are retained for subsequent slice routing. In our implementation, N=8,192 brain points are sampled per volume. The same attributed point-cloud construction is used throughout the segmentation pipeline.

We encode the case-level point set using a pretrained Point-MAE backbone (Pang et al., 2022). Patch centers are selected via farthest-point sampling (FPS), which iteratively appends the point farthest from the already selected centers,

$$\mathbf{c}_j = \operatorname*{arg\,max}_{\tilde{\mathbf{p}}\in\mathcal{P}\setminus\mathcal{C}_{j-1}} \ \min_{\mathbf{c}\in\mathcal{C}_{j-1}} \|\tilde{\mathbf{p}} - \mathbf{c}\|_2, \tag{3}$$

where $\mathcal{P}$ denotes the normalized point set and $\mathcal{C}_{j-1}$ the set of the first $j-1$ selected centers. Because FPS operates in the full three-dimensional coordinate space, the selected centers spread across adjacent axial levels rather than within a single plane. Local volumetric patches are then formed by grouping the k nearest neighbors of each center,

$$\mathcal{N}(\mathbf{c}_j) = \operatorname*{top\text{-}k}_i \ \|\tilde{\mathbf{p}}_i - \mathbf{c}_j\|_2, \qquad P_j = \{\mathbf{f}_i \mid i \in \mathcal{N}(\mathbf{c}_j)\}, \tag{4}$$

so the points within P_j are not confined to a single axial plane: even when the patch center is associated with slice z, its neighborhood may extend to both superior and inferior locations. This is the principal source of volumetric information in P3CE: each patch representation is derived from a local three-dimensional point set rather than an isolated image slice. In our implementation, G=64 patch centers are selected with k=16 neighbors per group.

A shared local encoder aggregates the relative geometry and complementary point attributes within each patch into a permutation-invariant token. Following the PointNet paradigm, the encoder applies shared multilayer perceptrons to each member point—using coordinates relative to the patch center together with the absolute attributes—and aggregates the group through symmetric max-pooling:

$$\mathbf{t}_j = \operatorname*{MaxPool}_{i\in\mathcal{N}(\mathbf{c}_j)} \ \mathrm{MLP}_\theta\Big(\left[\,(\tilde{\mathbf{p}}_i - \mathbf{c}_j)^\top,\ \mathbf{a}_i^\top\,\right]^\top\Big) \in \mathbb{R}^{C_P}, \tag{5}$$

where $\mathbf{a}_i$ collects the HU intensity and the one-hot brain-tissue, source-state, and health-state attributes of point i. The resulting descriptor summarizes local shape, tissue appearance, anatomical identity, state information, and cross-slice variation of the neighborhood, while remaining invariant to the ordering of points within the group.

The spatial position of each patch center is embedded via a two-layer MLP and combined with its local token. A pretrained Transformer (12 layers, 384-dimensional embeddings, 6 attention heads) then contextualizes the patch representations across the complete case, producing $\mathbf{T}_p = [\mathbf{t}_1, \ldots, \mathbf{t}_G] \in \mathbb{R}^{G \times C_p}$ together with their spatial centers. The point encoder is frozen during segmentation training, avoiding end-to-end optimization of a dense 3D backbone.

3.2 Slice-Conditioned Retrieval of 3D Patches

The point encoder operates once on the complete volume, while the segmentation network operates on axial slices. We connect the two representations by assigning each patch to a recipient slice according to its center. Let $c_j = (c_j^x, c_j^y, c_j^z)$ denote the center of patch j in the normalized coordinate system. Its recipient slice is recovered by undoing the normalization along the z-axis:

$$\hat{z}_j = \mathrm{round}(c_j^z \cdot \sigma_z + \mu_z), \tag{6}$$

where μ_z and σ_z are the saved centroid offset and scale factor.

For a target slice z, the tokens assigned to that axial location are collected as its slice-conditioned volumetric context:

$$\mathbf{F}_p(z) = \mathrm{Select}\left(\{\mathbf{t}_j \mid \hat{z}_j = z\}\right) \in \mathbb{R}^{M_z \times C_p}. \tag{7}$$

At most $M_{\max}$ tokens are retained per slice ($M_{\max}$=8 in our implementation; zero-padded when fewer patches map to a given slice).

The center coordinate is used only to route a patch token to a recipient slice; it does not restrict the token content to the center plane. Since every selected token summarizes a neighborhood formed in three-dimensional space, $\mathbf{F}_p(z)$ carries structural evidence from the regions above and below slice z. The routing operation is therefore slice-specific, whereas the routed representation remains volumetric. This enables cross-slice context without introducing an additional adjacent-slice image encoder.

The point encoder processes each volume once, and its patch tokens are reused across slices. Only slice-relevant tokens are passed to the fusion module.

3.3 Lightweight 2D–3D Context Embedding

Image Encoding. The 2D foundation model's image encoder maps a batch of axial slices to $\mathbf{F}_i \in \mathbb{R}^{B \times C \times h \times w}$. We adopt the SAM-Med2D ViT-B encoder, in which each axial slice is patch-embedded, processed through 12 windowed/global Transformer blocks, and projected to a C-dimensional feature map. To adapt the pretrained encoder to CT data while preserving its representations, we insert low-rank adaptation (LoRA) modules into the query and value projections of each attention layer and lightweight adapter modules (channel-wise recalibration followed by spatial convolution) parallel to each MLP block. For a frozen pretrained projection $\mathbf{W}_0 \in \mathbb{R}^{d_{\mathrm{out}} \times d_{\mathrm{in}}}$, LoRA reparameterizes the adapted weight as

$$\mathbf{W} = \mathbf{W}_0 + \frac{\alpha}{r}\mathbf{B}\mathbf{A}, \qquad \mathbf{B} \in \mathbb{R}^{d_{\mathrm{out}} \times r},\ \mathbf{A} \in \mathbb{R}^{r \times d_{\mathrm{in}}}, \tag{8}$$

where $r \ll \min(d_{\mathrm{in}}, d_{\mathrm{out}})$ is the low-rank factor and α a scaling constant, so that only $\mathbf{A}$ and $\mathbf{B}$ receive gradients. Both LoRA and adapter parameters are trainable; all other encoder weights are frozen.

Cross-Attention Fusion. The slice-conditioned patch tokens are projected into the image feature dimension via a linear layer and normalized to form the keys and values $\mathbf{K}_p, \mathbf{V}_p \in \mathbb{R}^{B \times M_z \times C}$. Positional information is added to the image feature map, which is then flattened and normalized to form the query tokens $\mathbf{Q}_i \in \mathbb{R}^{B \times hw \times C}$.

We use image tokens as queries and three-dimensional patch tokens as keys and values. With H_a parallel attention heads (H_a=8 in our implementation), each head h projects the queries, keys, and values into a C/H_a-dimensional subspace and computes

$$\mathrm{head}_h = \mathrm{softmax}\left(\frac{\mathbf{Q}_h \mathbf{K}_h^\top}{\sqrt{C/H_a}}\right)\mathbf{V}_h, \qquad \mathbf{A}_z = \mathrm{Concat}\left(\mathrm{head}_1, \ldots, \mathrm{head}_{H_a}\right)\mathbf{W}_O, \tag{9}$$

where $\mathbf{W}_O$ is the output projection.

Each image location retrieves volumetric context relevant to its 2D appearance, providing spatially adaptive context rather than a single global descriptor. Because the keys and values summarize 3D neighborhoods, the retrieved representation captures superior–inferior continuity.

The attended sequence $\mathbf{A}_z$ is restored to the spatial image layout and passed through a lightweight projection block (two convolutional layers with a GELU activation) to obtain a dense volumetric prompt $\mathbf{P}_z$. We inject this prompt into the pretrained image feature through residual addition:

$$\tilde{\mathbf{F}}_i(z) = \mathbf{F}_i(z) + \mathbf{P}_z. \tag{10}$$

This residual formulation preserves the original two-dimensional feature pathway and learns only a complementary volumetric correction. The cross-attention cost depends on the number of image tokens ($h \times w$) and the small set of retrieved patch tokens (M_z), rather than on the number of voxels in the complete volume. The overall per-case cost of P3CE therefore decomposes as

$$C_{\text{case}} = D \cdot (C_{\text{slice}} + C_{\text{fuse}}) + C_{\text{point}}, \qquad C_{\text{fuse}} = \mathcal{O}(hw \cdot M_z \cdot C)\,, \tag{11}$$

where C_{slice} is the cost of the 2D image encoder, C_{point} the cost of the single case-level point encoding, and C_{fuse} the cost of the fusion module. Because the point branch runs once per case and $M_z \ll H \cdot W \cdot D$, the total overhead over a purely slice-wise 2D model remains marginal, in contrast to dense volumetric encoders whose cost scales as $\mathcal{O}(H \cdot W \cdot D \cdot C)$.

3.4 Mask Decoding

The fused image features $\tilde{\mathbf{F}}_i(z)$, together with sparse point prompts, are passed to a lightweight mask decoder based on a two-way Transformer (Kirillov et al., 2023). Following the prompt-based segmentation paradigm of SAM (Kirillov et al., 2023), each slice is provided with one positive point and one negative point randomly sampled from the ground-truth (GT) mask: the positive point is sampled from the hemorrhage region, while the negative point is sampled from the brain region outside the hemorrhage. The sampled points are encoded as positional embeddings and serve as query tokens in the decoder. The sparse segmentation prompts are distinct from the attributed 3D point cloud: the former provide spatial guidance to the mask decoder, whereas the latter provide volumetric context through Point-MAE and cross-attention. The same GT-based random prompt generation is used during both training and testing.

The two-way Transformer performs bidirectional cross-attention between sparse prompts and dense fused features, integrating volumetric context with spatial guidance. The decoded image features are upscaled via two transposed convolution stages, and a hypernetwork MLP generates the final binary mask. A parallel IoU prediction head estimates mask confidence, which is used at inference time to suppress low-confidence predictions.

Overall, P3CE provides an efficient mechanism for embedding three-dimensional context into a slice-wise segmentation framework. Instead of constructing a dense volumetric feature tensor or repeatedly encoding neighboring images, it represents the complete case as a sparse set of local 3D patch tokens and routes only the relevant tokens to each axial slice. Because every patch aggregates points in three-dimensional space, its token naturally captures within-slice structure together with superior–inferior variation. By separating the computation unit from the context unit, P3CE resolves the principal limitation of purely slice-wise segmentation—its lack of cross-slice spatial context—while retaining the computational efficiency, pretrained representations, and scalable slice-wise inference of a 2D foundation model.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on Seg-CQ500 (Chilamkurthy et al., 2018) and CT-ICH (Hssayeni et al., 2020), two CT datasets containing intracranial hemorrhage annotations. Seg-CQ500 is used as the primary benchmark for comparison with existing medical image segmentation methods, while CT-ICH is used to assess generalization to a second data source. Each case is represented as a three-dimensional CT volume, and the target is a binary voxel-wise hemorrhage mask. Patient-level splitting prevents slices from the same subject appearing in different partitions.

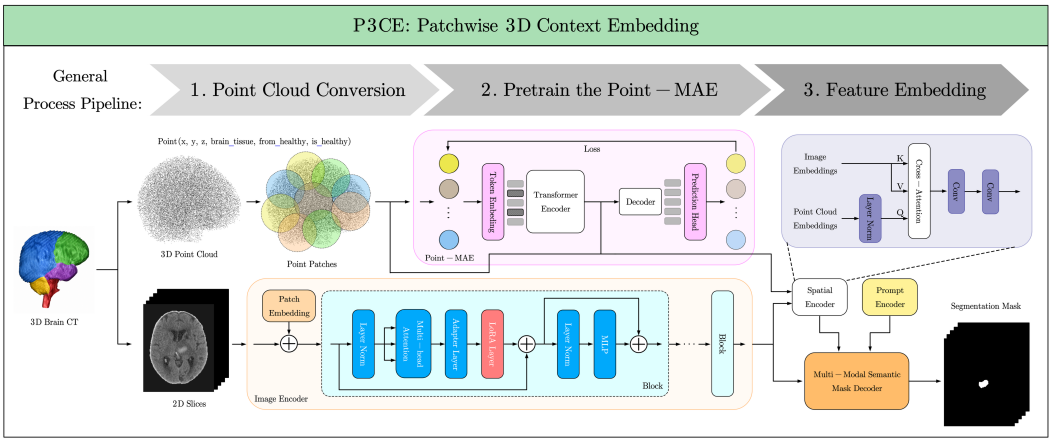

Figure 1: Overview of the P3CE framework. A frozen Point-MAE encoder processes the complete CT volume once, producing sparse 3D patch tokens. A slice-conditioned retrieval module routes relevant tokens to each axial slice. Lightweight cross-attention injects volumetric context into 2D image features via residual fusion before mask decoding.

Evaluation Metrics. We report intersection over union (IoU) and the Dice similarity coefficient to measure region overlap between the prediction P and the ground truth G. We additionally report the Topological Intersection over Slice Overlap (TISO) to measure the continuity of predicted masks between adjacent axial slices. Let M_z and M_{z+1} denote the predicted masks on two neighboring slices, and let $D_r(\cdot)$ denote morphological dilation with radius r. TISO is defined as

$$\mathrm{TISO}_z = \frac{|M_z \cap D_r(M_{z+1})| + |M_{z+1} \cap D_r(M_z)|}{|M_z| + |M_{z+1}|}. \tag{12}$$

In our implementation, r=1 pixel is used by default.

Implementation Details. Our framework consists of a slice-wise SAM-Med2D image branch and a frozen Point-MAE point branch. The point branch is initialized from our pre-trained, brain-specific Point-MAE checkpoint (Pang et al., 2022), which was pre-trained on head CT volumes with a masked-autoencoding objective to learn volumetric anatomical representations. The resulting Point-MAE encoder is kept frozen during segmentation training; the downstream projection and fusion modules are optimized together with the image branch. The point-cloud input follows the full attributed representation defined in Sec. 3.1. The image branch is initialized with the SAM-Med2D ViT-B checkpoint. Following the prompt-based segmentation paradigm of SAM, each slice uses one positive and one negative point randomly sampled from the GT mask and the surrounding brain region, respectively, during both training and testing. During training, only the LoRA modules, the adapter modules, the prompt encoder (point-feature projection, normalization, cross-attention, and convolutional projection), and the mask decoder are optimized. We use the Adam optimizer with a learning rate of 1×10^{-4}, train for 50 epochs with mixed-precision training, and use four workers for data loading. Input slices are replicated to three channels, ImageNet-normalized, and padded or resized to 224×224. The training DataLoader uses one volume sample per batch, with 16 slices sampled per item. During inference, each CT volume is encoded once by the point branch, and its tokens are reused across slices. Predictions are resized to the original image resolution, restricted to the brain region, thresholded at 0.5, and suppressed when the mask confidence is below 0.5.

Compared Methods. We compare our method with supervised and foundation-model baselines: nnU-Net (Isensee et al., 2021), SAM-Med2D (Cheng et al., 2023), SAM-Med3D (Wang et al., 2023), MedSAM2 (Ma et al., 2025), SAM2-UNet (Xiong et al., 2026), and AISCT-SAM (Kuang et al., 2026). All methods use official implementations or public checkpoints when available, with the same test sets and metrics.

Table 1: Quantitative comparison on Seg-CQ500 and CT-ICH. Higher is better for all segmentation metrics; GFLOPs/case reports the computational overhead per volume (32 slices assumed). *, $^\#$, and † denote native 3D, native 2D, and 2.5D models, respectively. Our 2D backbone with sparse 3D point context achieves the best overlap accuracy on both datasets with only $+0.9\%$ overhead over the SAM-Med2D baseline. The best results are in **bold**.

Method	Year	Seg-CQ500			CT-ICH			GFLOPs /case
		IoU	Dice	TISO	IoU	Dice	TISO	
nnU-Net*	2018	0.4715	0.5799	0.9054	0.3112	0.4061	0.8852	22,137.6
SAM-Med2D$^\#$	2023	0.4841	0.6153	0.6298	0.4098	0.5712	0.6273	848.3
SAM-Med3D*	2023	0.4937	0.6398	0.8927	0.2928	0.4197	0.8115	98.7
MedSAM2†	2025	0.5423	0.6816	0.7714	0.4597	0.6112	0.8323	1,927.4
SAM2-UNet$^\#$	2026	0.3293	0.4127	0.4511	0.1419	0.1957	0.4727	4,286.8
AISCT-SAM†	2026	0.3129	0.4278	0.8552	0.2455	0.3326	0.8816	1,087.3
Ours	–	**0.5686**	**0.7078**	0.8230	**0.4773**	**0.6343**	0.8323	855.6

4.2 Main Results

Table 1 summarizes the comparison on Seg-CQ500 and CT-ICH. Our method obtains the best IoU and Dice on both datasets. On Seg-CQ500, it reaches 0.5686 IoU and 0.7078 Dice, improving over the strongest competing method, MedSAM2, by 0.0263 and 0.0262, respectively. Compared with the slice-wise SAM-Med2D baseline, the improvements are 0.0845 IoU and 0.0925 Dice. On CT-ICH, our method achieves 0.4773 IoU and 0.6343 Dice, which are 0.0176 and 0.0231 higher than MedSAM2. These results support the effectiveness of injecting sparse volumetric context into a 2D foundation model.

Our TISO values are 0.8230 on Seg-CQ500 and 0.8323 on CT-ICH, competitive but below nnU-Net on both datasets. Thus, our method has the strongest overlap performance while maintaining competitive TISO.

Computational Overhead. Table 1 additionally reports GFLOPs per case, assuming 32 slices per volume. All GFLOPs values in the table are measured with `fvcore` using the implemented architectures and their corresponding inference configurations. For nnU-Net, the measurement uses dense sliding-window inference with 128^3 patches at 50% overlap over a $512 \times 512 \times 32$ volume; SAM-Med3D processes the volume at a native 128^3 resolution in a single pass; MedSAM2 runs the SAM2.1-Tiny encoder at 512^2 per slice together with its memory-attention module; SAM2-UNet runs the Hiera-L encoder at 352^2 per slice with a U-shaped decoder; and AISCT-SAM augments SAM-Med2D with additional adapter and self-prompting modules. For our method, the image encoder, prompt encoder, and mask decoder run once per slice, whereas the frozen point branch runs once per case. Despite the extra volumetric branch, our overhead over the purely slice-wise SAM-Med2D baseline is only $+0.9\%$, since the frozen point encoder costs 1.9 GFLOPs per case and the fusion module attends over at most $M_{\max}$ patch tokens per slice.

4.3 Ablation Studies

We perform ablations on Seg-CQ500 to isolate point-cloud prompting and point features. The first variant removes point-cloud prompting, reducing the model to the original SAM-Med2D pathway. The second retains the point-cloud branch but removes the additional point features. The full model uses the complete attributed point representation (coordinates, HU intensity, brain-tissue label, source-state indicator, and health-state indicator), patchwise volumetric encoding, slice-conditioned retrieval, and cross-attention fusion.

Removing point-cloud prompting decreases IoU from 0.5686 to 0.4841 and Dice from 0.7078 to 0.6153, showing that the volumetric branch provides a substantial gain over purely slice-wise processing. Removing the point features while retaining the point pathway yields intermediate performance (0.5116 IoU and 0.6529 Dice), indicating that the gain is not explained solely by an additional geometric token stream. The full point representation further improves TISO from 0.7144 to 0.8230.

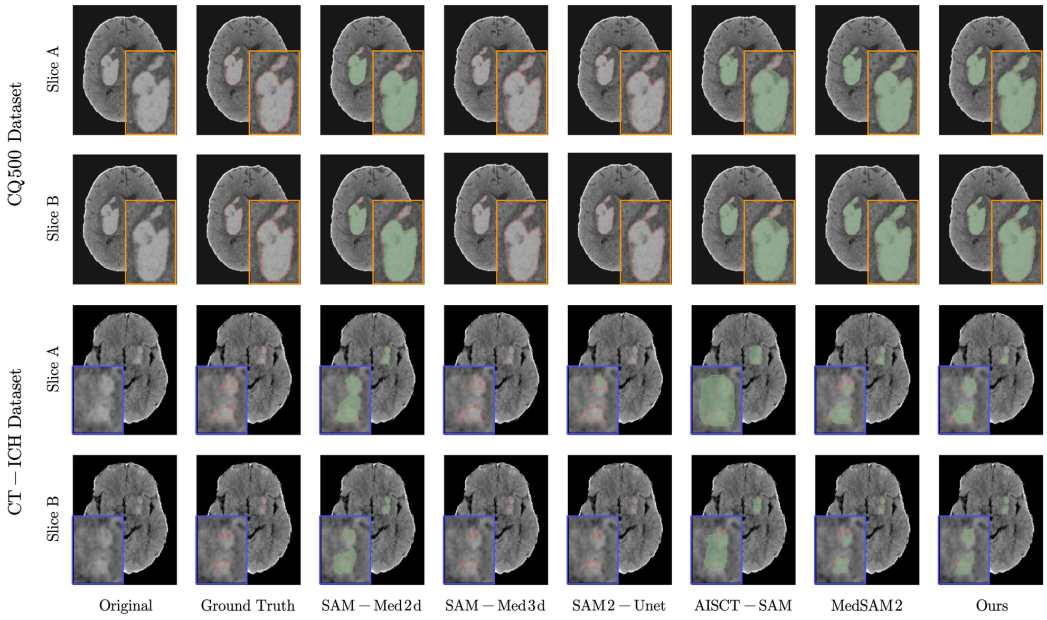

Figure 2: Qualitative comparison. Our method recovers small, low-contrast hemorrhagic lesions that slice-wise baselines miss, while maintaining smoother inter-slice consistency.

Table 2: Ablation study of point-cloud prompting and point features on Seg-CQ500.

Variant	IoU	Dice	TISO
w/o point cloud prompting (SAM-Med2D)	0.4841	0.6153	0.6298
w/o point features	0.5116	0.6529	0.7144
Ours (full model)	**0.5686**	**0.7078**	**0.8230**

5 Conclusion

We presented P3CE, a framework that introduces volumetric context into 2D foundation models for brain hemorrhage segmentation by encoding each CT volume as sparse attributed 3D patch tokens and routing them to individual slices through lightweight cross-attention.

On Seg-CQ500 and CT-ICH, P3CE achieved the highest IoU and Dice among the evaluated methods, with notable improvements on small and low-contrast hemorrhages, while the ablation study confirmed that both the 3D point-cloud representation and the point attributes contribute to the gain. Notably, inter-slice continuity (TISO) improved substantially over the 2D baseline but remained below dense 3D models such as nnU-Net, indicating that sparse point-level context captures much but not all of the volumetric coherence offered by full 3D architectures.

These findings suggest that the separation of computation unit (2D slice) from context unit (3D patch) is a practical design principle for extending 2D foundation models to volumetric medical imaging tasks, pending broader validation across institutions, scanners, and pathologies.

Future work will investigate adaptive patch selection, explicit cross-slice consistency objectives, comparison with 2.5D neighboring-slice models, uncertainty-aware post-processing, and broader validation across institutions and scanners.

REFERENCES

Sang Joon An, Tae Jung Kim, and Byung-Woo Yoon. Epidemiology, risk factors, and clinical features of intracerebral hemorrhage: An update. *Journal of Stroke*, 19(1):3–10, 2017.

Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, Hui Sun, Junjun He, Shaoting Zhang, Wentao Zhu, and Yu Qiao. SAM-Med2D. *arXiv preprint arXiv:2308.16184*, 2023.

Sasank Chilamkurthy, Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G. Campeau, Vasantha Kumar Venugopal, Vidur Mahajan, Pooja Rao, and Prashant Warier. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *The Lancet*, 392(10162): 2388–2396, 2018.

Jeremy J. Heit, Michael Iv, and Max Wintermark. Imaging of intracranial hemorrhage. *Journal of Stroke*, 19(1):11–27, 2017.

Mohammad D. Hssayeni, Muayad S. Croock, Aymar D. Salman, Hassan F. Al-khafaji, Zeena A. Yahya, and Behnaz Ghoraani. Intracranial hemorrhage segmentation using a deep convolutional model. *Data*, 5(1):14, 2020.

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

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

Hulin Kuang, Xianzhen Tan, Li S, Shichao Kan, Jin Liu, Jiarui Sun, Jingyang Zhang, Chunfeng Yang, Wu Qiu, Jiulou Zhang, Yang Chen, and Yi-Xiang Wang. AISCT-SAM: Customized SAM-Med2D with 3D context awareness and self-prompt generation for fully automatic acute ischemic stroke lesion segmentation on non-contrast CT scans. *IEEE Journal of Biomedical and Health Informatics*, 2026. doi: 10.1109/JBHI.2026.3693262.

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

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.

Yatian Pang, Wenxiao Wang, Francis E. H. Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *European Conference on Computer Vision (ECCV)*, pp. 604–621. Springer, 2022.

Charlotte J. van Asch, Merel J. A. Luitse, Gabriel J. E. Rinkel, Ingeborg van der Tweel, Ale Algra, and Catharina J. M. Klijn. Incidence, case fatality, and functional outcome of intracerebral haemorrhage over time, according to age, sex, and ethnic origin: a systematic review and meta-analysis. *The Lancet Neurology*, 9(2):167–176, 2010.

Haoyu Wang, Sizheng Guo, Jin Ye, Zhongying Deng, Junlong Cheng, Tianbin Li, Jianpin Chen, Yanzhou Su, Ziyan Huang, Yiqing Shen, Bin Fu, Shaoting Zhang, Junjun He, and Yu Qiao. SAM-Med3D. *arXiv preprint arXiv:2310.15161*, 2023.

Xinyu Xiong, Zihuang Wu, Shuangyi Tan, Wenxue Li, Feilong Tang, Ying Chen, Siying Li, Jie Ma, and Guanbin Li. SAM2-UNet: Segment anything 2 makes strong encoder for natural and medical image segmentation. *Visual Intelligence*, 4(1):2, 2026.