Where to Look Matters: Rethinking Sub-Volume Sampling in 3D Medical Self-supervised Learning

Junkai Liu **Le Zhang**†
School of Engineering, University of Birmingham
jxl1920@student.bham.ac.uk l.zhang.16@bham.ac.uk

Website **Code**

Abstract

Self-supervised learning on 3D medical images commonly relies on random sub-volume sampling during pretraining. However, random cropping implicitly treats all spatial regions as equally worth observing, despite anatomical priors and spatial redundancy that make regions inherently unequal in their value for representation learning. To this end, we first show through controlled crop-quality manipulations that better sub-volumes lead to better downstream transfer, revealing the observation policy as an overlooked bottleneck in 3D medical SSL. Motivated by this finding, we revisit sub-volume sampling as a *where-to-look* problem. We propose `VolumeProbe`, a plug-and-play framework that replaces random cropping with an offline coarse-to-fine probe bank guided by *anatomy-awareness*, *informativeness*, and *diversity*. We instantiate `VolumeProbe` as a lightweight `Lite` variant using geometric and image-statistical cues, and a diagnostic `Feat` variant using frozen visual features, to examine whether lightweight volume-intrinsic priors suffice or richer semantic cues reliably provide additional value. Extensive experiments across multiple 3D medical SSL backbones and downstream tasks demonstrate that `VolumeProbe-Lite` consistently improves transfer performance over random sampling. Mechanistic analyses further link these gains to more anatomically plausible and informative observations, reduced spatial redundancy, more efficient learning dynamics, and more structured representations.

1 Introduction

Self-supervised learning (SSL) has emerged as a powerful paradigm for pretraining on large-scale 3D medical images without dense annotations [68, 69, 60]. Existing 3D medical SSL pipelines commonly draw fixed-size sub-volumes from each scan via random cropping, which has become the default observation scheme [17, 65, 82, 81]. Representation learning is then driven by repeatedly sampling such sub-volumes over each training iteration. Yet this practice implicitly assumes that all spatial regions are equally valuable for representation learning. In 3D volumes, even after standard foreground cropping, random sub-volume sampling can still allocate the fixed crop budget to low-yield, boundary-dominated, or repetitive regions, while anatomically salient regions provide disproportionately richer observations for representation learning (Fig. 1(a)) [33, 55, 79]. These observations raise a fundamental question that has been largely underexplored in 3D medical SSL:

Given a fixed crop budget, where should a 3D medical SSL model look?

Our **core insight** is that, under the sub-crop setting, the quality of sampled sub-volumes can directly shape representation learning in 3D medical SSL. Pretraining is ultimately driven by the evidence exposed through repeated observations [29, 56, 24]. Intuitively, crops covering anatomically plausible

†: Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

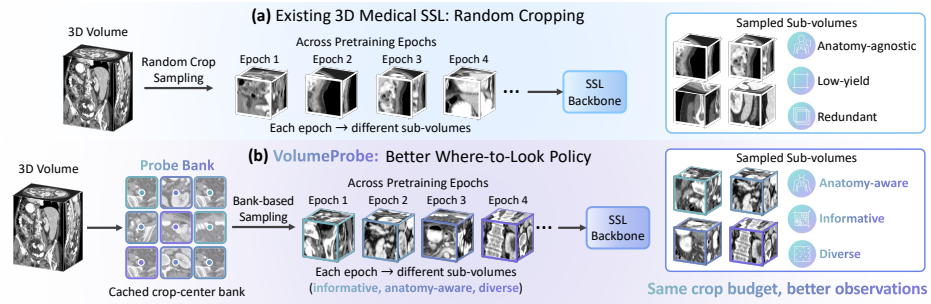


Figure 1: `VolumeProbe` **reallocates the fixed crop budget toward better observations.** **(a)** Random cropping in 3D medical SSL often samples low-yield, anatomy-agnostic, and redundant sub-volumes. **(b)** `VolumeProbe` builds a cached probe bank guided by anatomy-awareness, informativeness, and diversity, and samples higher-quality sub-volumes across epochs.

and information-rich regions provide meaningful supervisory signals, whereas redundant or low-yield crops squander the limited observation budget [54, 49, 23, 61]. As we illustrate in Sec. 2, these differences in crop quality already translate into measurable differences in downstream performance. This suggests that random cropping should not be viewed merely as a preprocessing convenience, but rather as a critical observation policy. Building on this insight, we reformulate sub-volume sampling as a structured *where-to-look* problem rather than an unconstrained random process, enabling the sub-volume cropping budget to be reallocated toward better observations (Fig. 1(b)).

These observations motivate us to propose `VolumeProbe`, a training-free, plug-and-play framework for structured sub-volume sampling in 3D medical SSL. `VolumeProbe` replaces random cropping with a probe bank built on three simple and intuitive principles: (i) *anatomy-awareness*, to favor anatomically plausible regions; (ii) *informativeness*, to prioritize sub-volumes with richer structural content; and (iii) *diversity*, to reduce redundancy among observations. Concretely, `VolumeProbe` scores candidate locations, i.e., 3D crop-center coordinates in the normalized volume space, under these principles to build a cached probe bank offline, from which sub-volumes are drawn during pretraining. To understand what makes an effective observation policy under this setting, we instantiate `VolumeProbe` in two variants: a lightweight `Lite` version driven by low-cost geometric priors, and a `Feat` version built on frozen dense features. This lets us ask whether lightweight priors already suffice, or whether richer semantic cues offer additional value. Across both variants, `VolumeProbe` reallocates a fixed crop budget toward higher-quality observations without modifying the SSL backbone or training objective. Our main contributions can be summarized as follows:

- **Finding.** Through a controlled motivation study, we unveil sub-volume observation quality as a previously underexplored factor that influences representation learning in 3D medical SSL.
- **Framework.** We propose `VolumeProbe`, a training-free, plug-and-play module that instantiates the *where-to-look* formulation through a probe bank in both `Lite` and `Feat` variants.
- **Understanding.** Beyond benchmark gains, principle-driven mechanistic analyses and the `Lite` vs `Feat` comparison reveal how observation quality translates into learning and clarify when lightweight priors suffice versus when richer semantic cues yield additional value.
- **Broad Applicability.** Extensive experiments on 3 SSL frameworks across 3 downstream tasks show that `VolumeProbe` consistently improves transfer performance over random cropping, demonstrating its plug-and-play compatibility with diverse SSL backbones.

2 Motivation Analysis: Why Observation Quality Matters

In this section, we investigate whether sub-volume observation quality is relevant to representation learning in 3D medical SSL through a small-scale experiment using VoCo [68] as the SSL backbone.

Observation waste under a fixed crop budget. As shown in Fig. 2(a), even after standard foreground cropping, random crops can still cover peripheral or structure-sparse regions. Quantitatively, they contain 55.83% foreground voxels on average, yet only 19.79% overlap annotated organ masks on the mask-available subset [32, 53, 67, 44]. This gap indicates that random sampling often observes non-background but anatomy-agnostic regions, leaving organ-rich evidence underutilized.

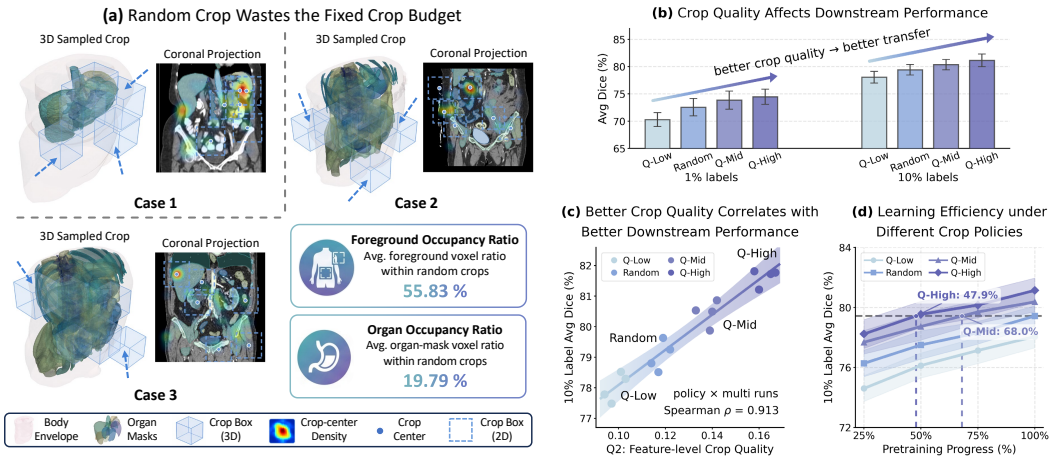


Figure 2: **Crop quality matters for 3D medical representation learning.** **(a)** Random cropping frequently spends the fixed crop budget on structure-sparse regions. **(b)** Policies defined by a semantic quality axis Q_1 yield a monotonic change in downstream Dice. **(c)** This trend remains under an alternative feature-level proxy Q_2. **(d)** Better crop quality further improves learning efficiency.

Sub-volume quality shifts downstream transfer. To test whether this waste affects representation learning, we construct four static probe policies along an oracle semantic quality axis Q_1, defined as the number of distinct organ classes contained in each crop. For each mask-available volume, we sample candidate crop centers, rank them by Q_1, and form static banks from the bottom (*Q-Low*), middle (*Q-Mid*), and top (*Q-High*) quartiles, while *Random* samples uniformly from all candidates. Pretraining follows the VoCo-10k protocol [68] for 50% of the full training steps, with Q-manipulation applied only to volumes with available organ masks [32, 53, 67, 44]. Fig. 2(b) shows that downstream average Dice on BTCV [32] and AMOS [28] increases monotonically from *Q-Low* to *Q-High* under both 1% and 10% label regimes, with *Random* lying in between. This suggests that crop quality is a controllable axis that random sampling does not fully exploit.

The quality-transfer link holds under an alternative proxy. Fig. 2(c) replaces Q_1 with a feature-level quality Q_2 computed as the variance of frozen MedicalNet [7] features on the sampled crops. The correlation with downstream average Dice under 10% labels remains positive (Spearman $\rho = 0.913$), so the quality-transfer relationship is not bound to a specific oracle.

Crop quality improves learning efficiency. Fig. 2(d) reveals a clear ordering throughout pretraining, with *Q-High* and *Q-Mid* consistently above *Random* and *Q-Low* lowest. Notably, *Q-High* reaches the final Dice of Random@100% at only 47.9% pretraining progress, and *Q-Mid* at 68.0%, showing that higher-quality observations yield comparable transfer with less pretraining.

However, neither Q_1 nor Q_2 is available during pretraining, as the former requires organ labels and the latter an external feature extractor. These findings thus motivate `VolumeProbe` as a training-free observation policy that approximates the quality axis using only volume-intrinsic signals.

3 Methodology

3.1 Overview

Let $\mathbf{x}$ denote a preprocessed 3D volume with normalized coordinate space $\Omega = [0, 1]^3$. As shown in Fig. 3, `VolumeProbe` first builds a coarse-to-fine candidate set $\mathcal{V}(\mathbf{x}) \subset \Omega$, where Level 1 partitions Ω into a coarse grid of $G_x \times G_y \times G_z$ bins and Level 2 refines only the selected high-value bins into fine candidate centers (i.e., 3D position coordinates). From this pool, `VolumeProbe` selects a probe bank $\mathcal{B}(\mathbf{x}) = \{v_1, \ldots, v_K\}$ of K sub-volume centers offline, from which one center is sampled during pretraining to crop a fixed-size sub-volume. In what follows, we establish three principles for candidate scoring and probe bank construction (Sec. 3.2), instantiate them in a `Lite` variant (Sec. 3.3) and a `Feat` variant (Sec. 3.4), and describe pretraining integration (Sec. 3.5).

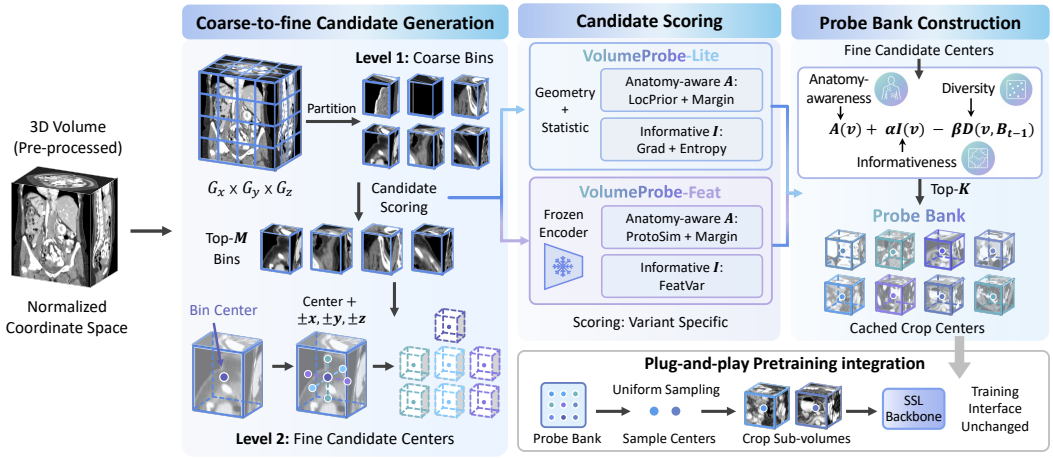


Figure 3: **Overview of `VolumeProbe`.** `VolumeProbe` constructs an offline probe bank of normalized 3D crop centers guided by anatomy-awareness, informativeness, and diversity. Pretraining then samples sub-volumes from this bank without changing the SSL backbone or objective.

3.2 Principles for Probe Bank Construction

An effective probe bank should concentrate the fixed crop budget on candidates that are worth observing. We characterize each candidate v through three principles. The *anatomy-aware* term $A(v)$ favors locations in meaningful body regions rather than weak or boundary-dominated areas. The *informativeness* term $I(v)$ rewards candidates whose local region contains richer structural content. The *diversity* term, expressed as a redundancy penalty $D(v, \mathcal{B})$, discourages candidates that would make the bank locally overcrowded. We combine them into a unified objective as

$$v_t = \arg \max_{v \in \mathcal{V}(x) \setminus \mathcal{B}_{t-1}} \Big[A(v) + \alpha I(v) - \beta D(v, \mathcal{B}_{t-1}) \Big], \qquad t = 1, \ldots, K. \tag{1}$$

where α and β are weight parameters. The resulting greedy construction is tractable and deterministic, and `VolumeProbe-Lite` and `VolumeProbe-Feat` share this formulation, differing only in how the three terms are quantified.

3.3 `VolumeProbe-Lite`

`VolumeProbe-Lite` instantiates probe quality using geometric priors and image statistics. We rank Level-1 bins by $A(v) + \alpha I(v)$ in Eq. 1 and keep the top-M for refinement. Within each retained bin, Level-2 candidates consist of the bin center and up to six axis-aligned neighbors offset by a fixed fraction of the bin size along $\pm x, \pm y, \pm z$, clamped to stay inside the bin. We then rescore these fine candidates and construct the top-K probe bank by greedy redundancy-aware selection.

Anatomy-aware. We model anatomical plausibility using a simple geometric prior:

$$A(v) = \mathrm{LocPrior}(v) + \mathrm{Margin}(v). \tag{2}$$

Here, $\mathrm{LocPrior}(v) = 1 - \mathrm{dist}$, where dist is the normalized distance from v to the volume center with a smaller weight on the z axis to reflect volumetric anisotropy. $\mathrm{Margin}(v)$ is the normalized minimum distance from v to the volume boundary, penalizing candidates near the border. This geometric prior is not a center-cropping rule, but an anatomy-correlated plausibility cue that is jointly combined with informativeness, diversity, and coarse-to-fine candidate selection.

Informative. We quantify local informativeness by combining gradient and entropy cues:

$$I(v) = \widetilde{\mathrm{Grad}}(v) + \widetilde{\mathrm{Ent}}(v). \tag{3}$$

where $\widetilde{\mathrm{Grad}}(v)$ is the normalized mean gradient magnitude and $\widetilde{\mathrm{Ent}}(v)$ is the normalized intensity entropy computed in a local evaluation patch around v. Coarse-bin scoring evaluates the two terms over the entire bin, whereas fine-candidate scoring uses a local patch of the pretraining sub-crop size centered at v. The gradient term favors stronger structural transitions, while the entropy term avoids overly homogeneous regions.

Diversity. To prevent the bank from collapsing to a few nearby high-score locations, we use an axis-wise redundancy penalty:

$$D(v, B) = \max_{u \in B} \prod_{a \in \{x,y,z\}} \left[1 - \frac{|v_a - u_a|}{\delta_a}\right]_+ . \tag{4}$$

where $\delta_a = \eta \cdot \mathrm{roi}_a / \mathrm{shape}_a$ matches the anisotropic footprint of the pretraining crop and η is a fixed factor. The product form makes the penalty large only when v is close to an existing bank member along all three axes simultaneously, matching the overlap pattern of axis-aligned 3D crops.

Relying on geometric properties and statistics alone, `VolumeProbe-Lite` serves as a controlled reference point against which the benefit of richer semantic scoring (Sec. 3.4) can be isolated.

3.4 `VolumeProbe-Feat`

`VolumeProbe-Feat` investigates whether frozen semantic features provide additional value beyond the statistical cues in `VolumeProbe-Lite`. It shares the same candidate generation and bank-selection pipeline, but replaces the scoring cues with probe-level features extracted by a frozen vision encoder.

Anatomy-aware. We replace the geometric anatomy prior with a feature-consistency prior. Let $\bar{\mathbf{f}}_v$ be the pooled frozen feature within the local region centered at v, and let $b(v)$ denote its coarse bin. For each bin b, we compute a prototype $\boldsymbol{\mu}_b$ by averaging probe-level features assigned to it across the training set. The anatomy-aware score is

$$A_{\mathrm{feat}}(v) = \widetilde{\mathrm{ProtoSim}}(v) + \widetilde{\mathrm{Margin}}(v), \quad \mathrm{ProtoSim}(v) = \frac{1 + \cos(\bar{\mathbf{f}}_v, \boldsymbol{\mu}_{b(v)})}{2}. \tag{5}$$

where $\mathrm{ProtoSim}(v)$ measures whether a probe is feature-consistent with its coarse anatomical region, while $\mathrm{Margin}(v)$ reuses the boundary suppression from `VolumeProbe-Lite`. Both terms are normalized to $[0, 1]$ per volume, yielding $\widetilde{\mathrm{ProtoSim}}(v)$ and $\widetilde{\mathrm{Margin}}(v)$, respectively.

Informative. We quantify informativeness as local heterogeneity in the frozen feature space:

$$I_{\mathrm{feat}}(v) = \widetilde{\mathrm{FeatVar}}(v), \quad \mathrm{FeatVar}(v) = \frac{1}{|R(v)|} \sum_{p \in R(v)} \left\| \mathbf{f}_p - \bar{\mathbf{f}}_v \right\|_2^2 . \tag{6}$$

where $R(v)$ is the local probe region and $\mathbf{f}_p$ is the frozen feature at position p.

Diversity. We keep the axis-wise redundancy penalty unchanged from `VolumeProbe-Lite`, so the `Lite`-vs-`Feat` comparison isolates the effect of the scoring source while using the same candidate construction and bank-selection interface.

3.5 Pretraining Integration

Once the per-volume probe bank $\mathcal{B}(\mathbf{x})$ is constructed offline, `VolumeProbe` integrates into SSL pretraining without learnable parameters, online reweighting, or substantial recurring overhead. During pretraining, we sample crop centers either from the cached bank or via standard random cropping, then pass the resulting fixed-size sub-volumes to the unchanged SSL backbone.

4 Experiments

We organize our experiments around five research questions, covering both downstream performance and the mechanisms behind the observed gains.

- **RQ1**: Does `VolumeProbe` improve medical SSL across downstream datasets and SSL backbones?
- **RQ2**: Do lightweight geometric priors (`Lite`) already suffice for effective sub-volume probing, or does richer semantic information (`Feat`) offer additional value?
- **RQ3**: Are the three design principles of `VolumeProbe` (anatomy-awareness, informativeness, and diversity) necessary for its downstream gains?
- **RQ4**: How do the anatomy-aware, informative, and diversity principles shape the resulting probe bank and probing behavior?
- **RQ5**: Do these improved probing behaviors translate into measurable changes in how the pretraining model learns?

Table 1: **Results on segmentation and classification tasks.** For segmentation, we report Dice Score (%) under different training data proportions. For classification, we report Accuracy (%). The best results are **bolded**, and the second best results are underlined.

Method	Dice Score (%) ↑												Acc. (%) ↑
	BTCV			AMOS			WORD			Atlas-MRI			CC-CCII
	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%	
From Scratch													
UNETR [18]	28.05	42.85	79.82	23.67	60.06	77.02	30.89	73.76	78.73	18.65	40.83	49.97	88.65
Swin-UNETR [17]	27.71	51.33	80.53	28.94	63.45	82.51	43.68	74.03	80.39	17.19	39.46	48.50	88.32
General SSL													
MAE3D [8]	31.86	55.79	81.33	37.82	62.83	82.26	39.23	72.53	78.97	28.46	42.01	52.85	90.09
SimMIM [72]	34.73	56.34	76.03	30.98	63.70	80.83	37.56	73.25	79.60	26.22	41.53	52.47	91.27
MoCo v3 [19]	38.27	77.91	79.54	34.67	65.76	81.95	41.43	72.89	80.73	30.57	43.28	54.31	87.76
Medical Image SSL													
Swin-UNETR [17]	28.52	48.31	81.54	24.63	65.49	80.97	48.16	75.07	81.79	24.09	41.96	53.38	89.42
GVSL [21]	24.86	41.79	81.87	34.39	63.45	81.38	37.87	74.10	82.36	28.63	42.83	54.86	93.26
GL-MAE [81]	52.46	65.83	82.01	48.14	72.43	84.20	62.60	75.42	83.17	32.54	46.02	57.19	92.13
MIM [82]	60.53	74.20	81.97	50.40	73.59	84.72	60.58	77.03	83.98	33.19	45.20	57.03	92.75
VoCo [68]	63.33	77.85	83.85	55.81	73.34	84.44	64.24	77.28	84.06	34.85	47.56	58.83	93.45
+ VolumeProbe-Lite	67.21$_{\uparrow 3.88}$	79.39$_{\uparrow 1.54}$	83.97$_{\uparrow 0.12}$	58.64$_{\uparrow 2.83}$	74.95$_{\uparrow 1.61}$	**87.52**$_{\uparrow 3.08}$	68.93$_{\uparrow 4.69}$	**80.12**$_{\uparrow 2.84}$	**85.56**$_{\uparrow 1.50}$	38.68$_{\uparrow 3.83}$	49.27$_{\uparrow 1.71}$	59.98$_{\uparrow 1.15}$	93.98$_{\uparrow 0.53}$
+ VolumeProbe-Feat	68.07$_{\uparrow 4.74}$	**79.51**$_{\uparrow 1.66}$	83.89$_{\uparrow 0.04}$	57.86$_{\uparrow 2.05}$	74.70$_{\uparrow 1.36}$	87.13$_{\uparrow 2.69}$	68.42$_{\uparrow 4.18}$	79.74$_{\uparrow 2.46}$	84.99$_{\uparrow 0.93}$	37.93$_{\uparrow 3.08}$	49.06$_{\uparrow 1.50}$	59.42$_{\uparrow 0.59}$	94.25$_{\uparrow 0.80}$
S2DC [48]	66.77	77.93	84.09	55.19	74.28	84.68	69.14	78.96	84.46	33.79	46.18	58.53	94.19
+ VolumeProbe-Lite	**69.47**$_{\uparrow 2.70}$	78.74$_{\uparrow 0.81}$	**84.43**$_{\uparrow 0.34}$	58.83$_{\uparrow 3.64}$	75.31$_{\uparrow 1.03}$	87.17$_{\uparrow 2.49}$	69.73$_{\uparrow 0.59}$	80.04$_{\uparrow 1.08}$	85.22$_{\uparrow 0.76}$	37.16$_{\uparrow 3.37}$	48.08$_{\uparrow 1.90}$	59.44$_{\uparrow 0.91}$	**94.87**$_{\uparrow 0.68}$
+ VolumeProbe-Feat	68.76$_{\uparrow 1.99}$	78.92$_{\uparrow 0.99}$	84.19$_{\uparrow 0.10}$	59.47$_{\uparrow 4.28}$	75.17$_{\uparrow 0.89}$	86.87$_{\uparrow 2.19}$	70.28$_{\uparrow 1.14}$	79.82$_{\uparrow 0.86}$	85.43$_{\uparrow 0.97}$	38.19$_{\uparrow 4.40}$	48.83$_{\uparrow 2.65}$	59.81$_{\uparrow 1.28}$	94.34$_{\uparrow 0.15}$
MedGMAE [13]	66.19	77.11	83.22	58.79	75.65	84.90	68.09	77.73	84.19	36.24	48.16	59.42	93.08
+ VolumeProbe-Lite	66.85$_{\uparrow 0.66}$	78.44$_{\uparrow 1.33}$	83.48$_{\uparrow 0.26}$	**60.26**$_{\uparrow 1.47}$	**77.14**$_{\uparrow 1.49}$	85.62$_{\uparrow 0.72}$	70.67$_{\uparrow 2.58}$	79.42$_{\uparrow 1.69}$	84.36$_{\uparrow 0.17}$	**39.71**$_{\uparrow 3.47}$	**50.23**$_{\uparrow 2.07}$	60.19$_{\uparrow 0.77}$	93.58$_{\uparrow 0.50}$
+ VolumeProbe-Feat	66.57$_{\uparrow 0.38}$	79.09$_{\uparrow 1.98}$	83.29$_{\uparrow 0.07}$	59.57$_{\uparrow 0.78}$	76.80$_{\uparrow 1.15}$	85.93$_{\uparrow 1.03}$	**71.15**$_{\uparrow 3.06}$	79.08$_{\uparrow 1.35}$	84.25$_{\uparrow 0.06}$	39.48$_{\uparrow 3.24}$	49.83$_{\uparrow 1.67}$	**60.40**$_{\uparrow 0.98}$	93.81$_{\uparrow 0.73}$

4.1 Experimental Setup

Datasets. We integrate `VolumeProbe` into three 3D medical SSL frameworks, VoCo [68], S2DC [48], and MedGMAE [13], following their original pretraining protocols [32, 10, 53, 51, 44, 15, 67, 37], and replacing only the random cropping with `VolumeProbe` while keeping the number of sampled sub-volumes unchanged. Downstream evaluation spans three tasks. For segmentation, we use BTCV [32], AMOS [28], WORD [42], and Atlas-MRI [50] under 1%, 10%, and 100% data proportions with official splits. For classification, we adopt CC-CCII [74]. For registration, we evaluate on IXI [31] and OASIS [45] following [68, 69]. Additional details are deferred to Appendix A.

Implementation Details. For our 3D medical SSL baselines [68, 48, 13], we strictly follow their original default settings for pretraining and downstream evaluation, including preprocessing, crop size, number of sampled sub-volumes, backbone, and task-specific protocols. For `VolumeProbe`, we use the coarse grid $(G_x, G_y, G_z) = (4, 4, 2)$, with $M = 24$ and $K = 40$. The weights α and β are both set to 1.0 and the redundancy scale factor η to 0.35. For `VolumeProbe-Feat`, we use 3DINO-ViT [73] as the feature extractor. Further details are provided in Appendix B.

4.2 Experiments on Downstream Tasks (RQ1)

We compare `VolumeProbe` against both General [8, 72, 19] and Medical SSL methods [18, 17, 21, 81, 82, 68, 48, 13], using official implementations and pretrained weights where available, or results drawn from prior work [68, 13]. As shown in Table 1, substituting random cropping with `VolumeProbe` yields gains across all three SSL backbones on four datasets, with the most pronounced improvements under low-label regimes. Beyond segmentation, `VolumeProbe` further delivers steady accuracy improvements on CC-CCII [74], demonstrating that the representational benefits generalize to classification. Table 2 further corroborates these findings on registration, where gains on IXI [31] and OASIS [45] are modest yet consistently positive across all three backbones, confirming the plug-and-play effectiveness of `VolumeProbe` across diverse downstream tasks.

4.3 `Lite` vs. `Feat`: Are Lightweight Priors Sufficient? (RQ2)

Tables 1 and 2 show that `VolumeProbe-Lite`, relying only on training-free geometric and statistical cues, already achieves strong and consistent gains. Meanwhile, incorporating frozen semantic features in `VolumeProbe-Feat` does not reliably improve over `VolumeProbe-Lite`. This observation indicates that **volume-intrinsic priors are sufficient to construct effective probe banks**. We therefore treat `VolumeProbe-Feat` as a diagnostic external-feature variant rather than the main method. Unless otherwise specified, `VolumeProbe` in the remainder of this paper refers to `VolumeProbe-Lite`.

4.4 Ablation Study (RQ3)

Component Ablation. In Table 3, `VolumeProbe` delivers consistent gains over the random-cropping baseline on both VoCo [68] and S2DC [48], while ablating any single principle degrades performance.

Table 2: **Results on the registration task on IXI and OASIS datasets**.

Method	Dice Score (%) ↑	
	IXI	OASIS
From Scratch		
VoxelMorph [1]	71.54	78.55
TransMorph [6]	74.46	81.61
Swin-UNETR [17]	72.63	81.81
Medical Image SSL		
SUP [58]	67.67	81.54
SuPreM [38]	72.90	81.18
VoCo [68]	73.66	84.38
+ VolumeProbe-Lite	74.39$_{\uparrow 0.73}$	85.55$_{\uparrow 1.17}$
+ VolumeProbe-Feat	74.17$_{\uparrow 0.51}$	85.19$_{\uparrow 0.81}$
S2DC [48]	72.93	82.97
+ VolumeProbe-Lite	73.70$_{\uparrow 0.77}$	84.09$_{\uparrow 1.12}$
+ VolumeProbe-Feat	73.98$_{\uparrow 1.05}$	83.86$_{\uparrow 0.89}$
MedGMAE [13]	73.69	85.71
+ VolumeProbe-Lite	**74.88**$_{\uparrow 1.19}$	85.96$_{\uparrow 0.25}$
+ VolumeProbe-Feat	74.28$_{\uparrow 0.59}$	**86.17**$_{\uparrow 0.46}$

Table 3: **Ablation study of our principles for probe selection**.

Method	Dice Score (%) ↑		
	BTCV	AMOS	WORD
VoCo [68]	83.85	84.44	84.06
+ VolumeProbe	83.97$_{\uparrow 0.12}$	87.52$_{\uparrow 3.08}$	85.56$_{\uparrow 1.50}$
w/o A	83.87$_{\uparrow 0.02}$	84.96$_{\uparrow 0.52}$	84.47$_{\uparrow 0.41}$
w/o I	83.86$_{\uparrow 0.01}$	85.89$_{\uparrow 1.45}$	85.04$_{\uparrow 0.98}$
w/o D	83.93$_{\uparrow 0.08}$	87.10$_{\uparrow 2.66}$	85.29$_{\uparrow 1.23}$
S2DC [48]	84.09	84.68	84.46
+ VolumeProbe	84.43$_{\uparrow 0.34}$	87.17$_{\uparrow 2.49}$	85.22$_{\uparrow 0.76}$
w/o A	84.20$_{\uparrow 0.11}$	86.20$_{\uparrow 1.52}$	84.73$_{\uparrow 0.27}$
w/o I	84.27$_{\uparrow 0.18}$	85.45$_{\uparrow 0.77}$	84.81$_{\uparrow 0.35}$
w/o D	84.35$_{\uparrow 0.26}$	86.93$_{\uparrow 2.25}$	85.10$_{\uparrow 0.64}$

Notably, dropping anatomy-awareness or informativeness incurs the largest drops, underscoring that anatomically plausible and content-rich observations are particularly essential for robust transfer. Please refer to Appendices C.7 and C.10 for more fine-grained ablation studies and efficiency analysis.

Comparison with Simple Non-random Samplers. We further examine whether the gains can be explained by simply replacing random cropping with a non-random heuristic. As reported in Table 4, center-biased, uniform-grid, gradient-only, and entropy-only sampling yield limited or inconsistent improvements across datasets. In contrast, `VolumeProbe` provides more consistent gains, suggesting that effective sub-volume selection benefits from jointly considering anatomical plausibility, informativeness, and diversity rather than any single heuristic. Detailed experimental settings and sampler definitions are provided in Appendix C.5.

Table 4: **Comparison of different non-random sampling strategies for VoCo**.

Sampling Method	Dice Score (%) ↑		
	BTCV	AMOS	WORD
VoCo [68]	83.85	84.44	84.06
+ Center-Biased	83.79$_{\downarrow 0.06}$	85.32$_{\uparrow 0.88}$	83.95$_{\downarrow 0.11}$
+ Uniform-Grid	83.88$_{\uparrow 0.03}$	85.79$_{\uparrow 1.35}$	85.26$_{\uparrow 1.20}$
+ Grad-Only	83.74$_{\downarrow 0.11}$	84.53$_{\uparrow 0.09}$	83.87$_{\downarrow 0.19}$
+ Entropy-Only	83.80$_{\downarrow 0.05}$	84.31$_{\downarrow 0.13}$	84.00$_{\downarrow 0.06}$
+ VolumeProbe-Lite	83.97$_{\uparrow 0.12}$	87.52$_{\uparrow 3.08}$	85.56$_{\uparrow 1.50}$
+ VolumeProbe-Feat	83.89$_{\uparrow 0.04}$	87.13$_{\uparrow 2.69}$	84.99$_{\uparrow 0.93}$

4.5 Understanding: How Do the Three Principles Shape Where to Look? (RQ4)

4.5.1 Anatomy-aware Preference

Dataset-level Center Spatial Distribution.

Fig. 4 compares dataset-level normalized crop center distributions. *Random* spreads centers diffusely with limited overlap onto organ-dense regions, whereas *w/o A* forms sharper hotspots driven mainly by non-anatomical cues. `VolumeProbe` instead concentrates sub-crop centers within the organ-union contour and aligns more closely with the anatomical reference density, raising foreground occupancy from 55.83% to 75.63% and organ occupancy from 19.79% to 45.71% over *Random*. The anatomy-aware term thus goes beyond a simple center bias, actively steering the probe bank toward **anatomically plausible and content-rich observations**.

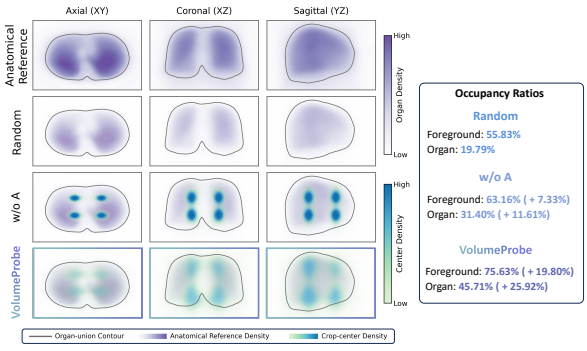


Figure 4: **Anatomy-aware analysis.** We aggregate organ masks to construct an anatomical reference density (purple), and overlay crop-center densities. `VolumeProbe` better aligns crop centers with organ-dense regions and achieves higher foreground and organ occupancy.

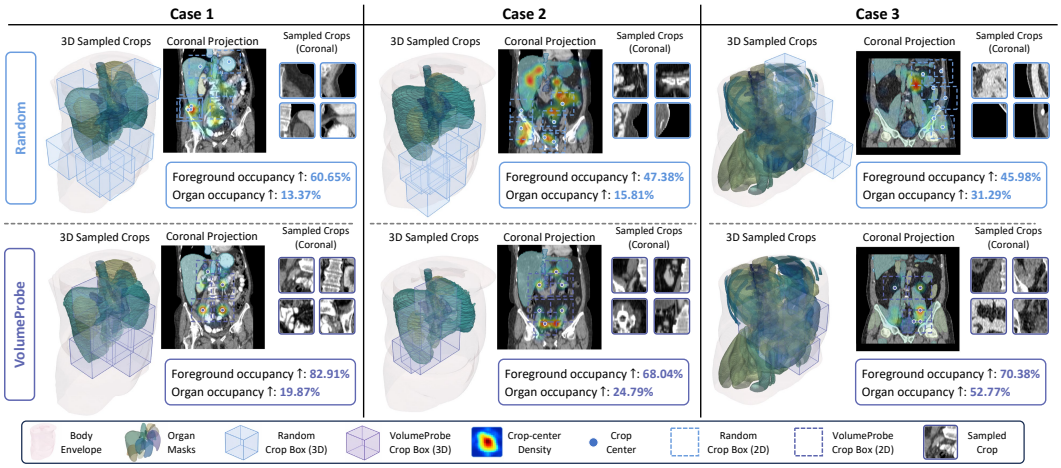


Figure 5: **Representative cases for anatomy-aware sampling.** We plot 3D crop boxes overlaid on organ masks, the coronal projection with sampled centers, 2D crop boxes, and example cropped views. Compared with *Random*, `VolumeProbe` places crops more consistently inside the body and around anatomically meaningful regions, raising foreground/organ occupancy ratios.

Representative Case Studies for Anatomy-aware Sampling. Fig. 5 provides case-level evidence across representative volumes from BTCV [32] and TotalSegmentator [67]. *Random* frequently places crops near body boundaries or partially outside the anatomical region, whereas `VolumeProbe` concentrates them within the body interior, yielding higher foreground and organ occupancy ratios.

4.5.2 Informative Observations

We examine whether the informativeness term facilitates content-richer observations using both intrinsic image statistics and an external anatomical signal from TotalSegmentator [67]. As shown in Fig. 6, `VolumeProbe` consistently accumulates higher gradient/entropy utility than *Random* and *w/o I*, while also observing more distinct organs and more any-organ hits despite not optimizing these anatomical labels. This highlights that **informativeness not only increases local structural content, but also improves the anatomical usefulness of the selected observations**, providing a stronger basis for the learning gains analyzed in Sec. 4.6.

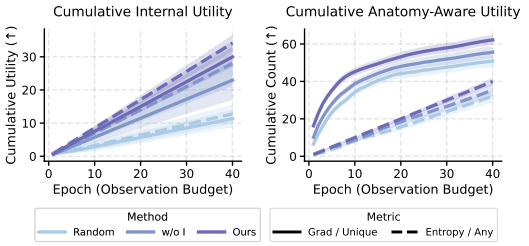


Figure 6: **Informativeness analysis**. *Left*: cumulative internal utility, measured by gradient- and entropy-based metrics. *Right*: cumulative anatomy-aware utility measured by the number of distinct organs observed and any-organ hits.

4.5.3 Static Diversity and Redundancy

We analyze whether the diversity term prevents the cached bank from concentrating around a few high-score regions. Fig. 7 showcases that removing diversity (*w/o D*) reduces support coverage and increases same-bin redundancy, even falling below *Random* in coverage, which suggests that informative scoring alone tends to collapse candidates into locally salient areas. In contrast, `VolumeProbe` maintains broader spatial support with lower redundancy, indicating that diversity is essential for **preventing bank collapse** rather than merely spreading candidates apart.

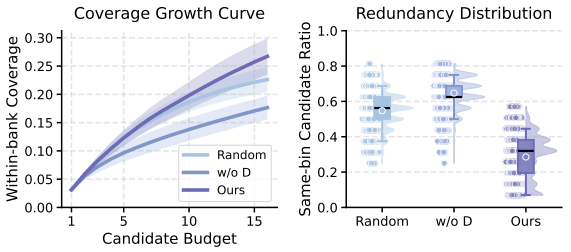


Figure 7: **Diversity analysis.** *Left:* within-bank coverage as candidate budget increases. `VolumeProbe` consistently achieves broader support. *Right:* same-bin candidate ratio measuring local redundancy. Our approach yields the lowest redundancy.

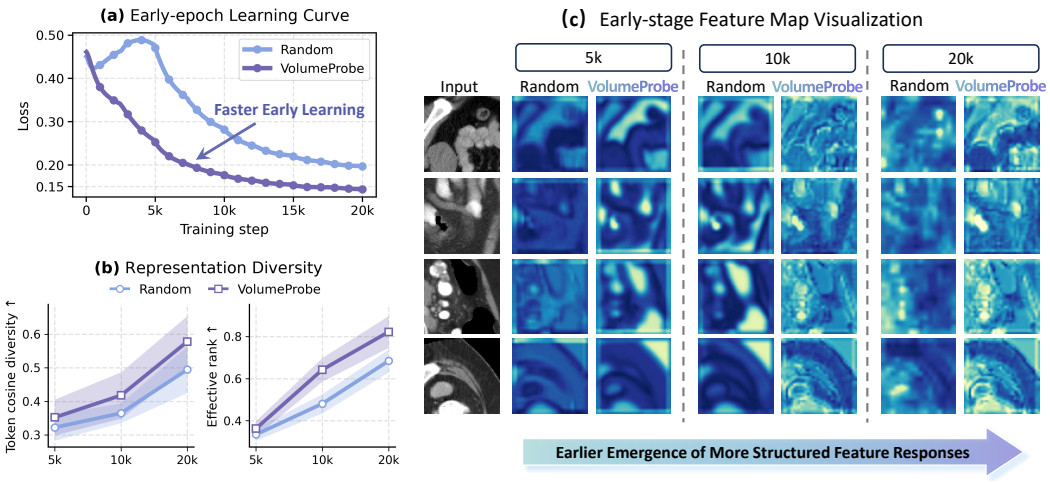


Figure 8: **Learning relevance of `VolumeProbe`.** **(a)** Early-stage learning curves, where `VolumeProbe` shows faster and more stable loss reduction. **(b)** Early representation diversity measured by token-level cosine diversity and effective rank. **(c)** Feature-map visualization, where `VolumeProbe` produces more structure-aligned representations.

4.6 Learning Relevance: From Better Crops to Better Representations (RQ5)

Sec. 4.5 has shown that `VolumeProbe` reshapes *where* the model looks. We now explore whether these improved observations affect *how* the model learns. By focusing on early pretraining stages (up to 20k steps, i.e., 20% of pretraining progress) of VoCo [68], we isolate the effect of observation quality before convergence narrows inter-method differences, thus closing the loop with Sec. 2.

Faster and More Stable Optimization. As shown in Fig. 8(a), `VolumeProbe` enters a stable loss descent earlier than *Random* and maintains a lower pretraining loss throughout early training. It reaches the loss of *Random*@20% pretraining progress at only 7.9% progress, indicating that the cached probe bank provides more efficient optimization signals under the same crop budget. This advantage persists under matched extended pretraining while retaining better compute efficiency (Appendix C.8).

Less Collapsed Representations. Beyond optimization, Fig. 8(b) indicates that `VolumeProbe` also improves early representation geometry, yielding higher token-level cosine diversity and effective rank across checkpoints. The former measures pairwise non-redundancy among token features, while the latter captures the global spread of the representation covariance. Their consistent gains suggest that `VolumeProbe` induces less redundant spatial features and a broader embedding spectrum.

	Organ Occ. (%)	Internal Utility@40	Coverage @15	Joint Rank	Mean Dice (%)
Random	19.79	11.36	0.226	0.233	84.12
Center-Biased	34.61	19.69	0.186	0.400	84.35
Uniform-Grid	28.57	17.68	0.292	0.467	84.98
Grad-Only	32.90	26.14	0.127	0.367	84.05
Entropy-Only	36.08	25.72	0.134	0.433	84.04
w/o A	31.40	31.27	0.243	0.600	84.43
w/o I	41.93	22.95	0.266	0.600	84.93
w/o D	44.76	31.76	0.176	0.700	85.44
VolumeProbe-Feat	44.92	28.74	0.269	0.833	85.34
VolumeProbe-Lite	45.71	29.96	0.267	0.867	85.68
Spearman ρ with Mean Dice	**0.600**	**0.467**	**0.648**	**0.863**	—

Figure 9: **Cross-strategy association between mechanistic diagnostics and downstream transfer.**

Structure-aligned Responses. We further probe the trained encoders in Fig. 8(c). As shown, `VolumeProbe`-trained models produce less diffuse and more anatomy-aligned feature responses than *Random*. Additional quantitative metrics in Appendix D.3 further confirm this trend across mask-available volumes. Together, these results verify that **better probing behavior not only accelerates early optimization, but also promotes less collapsed and more structure-aware representations**.

Linking Mechanistic Diagnostics to Downstream Transfer. We finally examine whether the mechanistic diagnostics above track downstream utility across ten samplers from Tables 3 and 4, without additional pretraining or fine-tuning. We consider the following diagnostics:

- **Organ Occupancy** (Sec. 4.5.1) measures the fraction of sampled crop voxels overlapping annotated organ masks, reflecting anatomical validity.
- **Internal Utility@40** (Sec. 4.5.2) measures cumulative local structural information over 40 observations, defined as $\mathrm{Utility@40} = \sum_{t=1}^{40} (\widetilde{\mathrm{Grad}}_t + \widetilde{\mathrm{Ent}}_t)/2$, where gradient magnitude and entropy are normalized within each volume.
- **Coverage@15** (Sec. 4.5.3) measures the spatial support covered by 15 selected centers, normalized to $[0, 1]$, reflecting sampling coverage.
- **Joint Rank** equally combines the percentile ranks of the three diagnostics: $\mathrm{JointRank}_i = (R_{\mathrm{Organ},i} + R_{\mathrm{Utility},i} + R_{\mathrm{Coverage},i})/3$.
- **Mean Dice** averages the matched 100%-label Dice on BTCV, AMOS, and WORD, serving as the downstream transfer measure.

Fig. 9 summarizes the cross-strategy comparison across the ten sampling methods, with original values annotated in each cell. Individual diagnostics show only moderate correlations with Mean Dice ($\rho = 0.600$, 0.467, and 0.648); for example, high coverage or utility alone does not consistently translate into better transfer. In contrast, Joint Rank correlates more strongly with downstream performance ($\rho = 0.863$, exact $p = 0.0022$), remaining stable under leave-one-strategy-out analysis ($\rho = 0.812$–0.912). **Downstream transfer is therefore most consistently associated with a balanced combination of anatomical validity, local informativeness, and spatial coverage rather than any single diagnostic alone.**

5 Related Work

3D Medical Image SSL. Self-supervised learning has emerged as a major paradigm for 3D medical image analysis, largely to alleviate the scarcity of annotations [68, 52, 4, 77, 43]. Existing methods are predominantly built on contrastive learning [68, 48, 20], masked image modeling [56, 13, 8, 82, 76], or their hybrids, refining representation learning through multi-view augmentation, contextual priors or reconstruction [14, 41, 47], which have advanced label-efficient transfer across downstream tasks [71, 69, 11, 64]. Yet most efforts concentrate on learning from sampled sub-volumes while taking the sampling process itself as given. The implicit assumption that uniformly random sub-volumes constitute adequate observations remains largely unquestioned, despite well-known heterogeneity of medical volumes in tissue composition, anatomical structure, and foreground density [34, 33, 25, 78, 59], leaving the observation policy underlying sub-crop pretraining underexplored.

Spatial Observation Allocation in SSL. Prior work on data selection in SSL mainly studies *which* training samples deserve more attention, through curriculum learning, self-paced learning, or importance sampling [3, 12], operating at the dataset level and leaving intra-sample spatial selection unexplored. In 2D SSL, non-uniform observation has been explored via object-aware or foreground/saliency-sensitive view construction [46, 39, 40], attention-based region selection [9, 80, 30], video keyframe selection [57, 66, 16, 75, 36, 5], and optimized masking in masked image modeling [63, 2, 70, 27, 22, 62, 35]. In 3D medical imaging, anatomical priors, spatial redundancy, anisotropic spacing, and sub-crop pretraining together make *where-to-look* within a volume a distinct problem [68, 56, 55, 33]. While some medical SSL methods encode spatial structure in the pretext task [68, 69], the observation mechanism itself remains random cropping.

6 Conclusion

In this paper, we shed light on sub-volume sampling, treating it not merely as a preprocessing step but as a structured *where-to-look* problem in a fixed-crop setting. We propose `VolumeProbe`, a simple yet effective module that replaces unconstrained random cropping with a probe bank guided by anatomy-awareness, informativeness, and diversity. Experiments across multiple SSL frameworks show that `VolumeProbe` redirects sub-volume sampling toward higher-quality observations, yielding consistent gains in transfer performance and more efficient learning. We hope this work helps illuminate sub-volume observation as an underexplored design dimension in 3D medical SSL.

Limitations. `VolumeProbe` builds a static probe bank offline and therefore cannot adapt its probing to the model's evolving learning dynamics. Future work will explore training-aware probe selection and more adaptive spatial priors beyond the non-canonical fields of view evaluated here.

References

[1] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca. Voxelmorph: A learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, 2019.

[2] W. G. C. Bandara, N. Patel, A. Gholami, M. Nikkhah, M. Agrawal, and V. M. Patel. Adamae: Adaptive masking for efficient spatiotemporal learning with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14507–14517, 2023.

[3] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, 2009.

[4] Q. Bi, J. Yi, H. Zheng, W. Ji, Y. Huang, Y. Li, and Y. Zheng. Learning generalized medical image representation by decoupled feature queries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12):11252–11269, 2025.

[5] X. Cao, H. Guo, J. Qian, G. Nan, C. Wang, Y. Pan, T. Hou, X. Wang, and Y. Gao. Videominer: Iteratively grounding key frames of hour-long videos via tree-based group relative policy optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23773–23783, 2025.

[6] J. Chen, E. C. Frey, Y. He, W. P. Segars, Y. Li, and Y. Du. Transmorph: Transformer for unsupervised medical image registration. *Medical Image Analysis*, 82:102615, 2022.

[7] S. Chen, K. Ma, and Y. Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.

[8] Z. Chen, D. Agarwal, K. Aggarwal, W. Safta, M. M. Balan, and K. Brown. Masked image modeling advances 3d medical image analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1970–1980, 2023.

[9] Z. Chen, G. Huang, W. Li, J. Teng, K. Wang, J. Shao, C. C. Loy, and L. Sheng. Siamese detr. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15722–15731, 2023.

[10] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26(6):1045–1057, 2013.

[11] G. Dai, C. Zhou, Y. Zhou, R. Zhang, Y. Zhang, C. Hou, T. Huang, J. Chen, and S. Zhang. Orochi: Versatile biomedical image processor. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[12] T. Evans, N. Parthasarathy, H. Merzić, and O. J. Hénaff. Data curation via joint example selection further accelerates multimodal learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 141240–141260. Curran Associates, Inc., 2024.

[13] X. Fu, F. Tang, R. Wang, Y. Li, L. Han, J. Lu, Z. Jiang, and S. K. Zhou. MedGMAE: Gaussian masked autoencoders for medical volumetric representation learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

[14] Y. Gao, Y. Zhang, C. Wang, J. Liu, M. Varma, J.-B. Delbrouck, A. Chaudhari, and C. Langlotz. Learning generalizable 3d medical image representations from mask-guided self-supervision. *arXiv preprint arXiv:2603.13660*, 2026.

[15] A. J. Grossberg, A. S. Mohamed, H. Elhalawani, W. C. Bennett, K. E. Smith, T. S. Nolan, B. Williams, S. Chamchod, J. Heukelom, M. E. Kantor, et al. Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy. *Scientific data*, 5(1):1–10, 2018.

[16] W. Guo, Z. Chen, S. Wang, J. He, Y. Xu, J. Ye, Y. Sun, and H. Xiong. Logic-in-frames: Dynamic keyframe search via visual semantic-logical verification for long video understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[17] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pages 272–284, 2021.

[18] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 574–584, 2022.

[19] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[20] Y. He and S. Li. Vector contrastive learning for pixel-wise pretraining in medical vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19827–19837, 2025.

[21] Y. He, G. Yang, R. Ge, Y. Chen, J.-L. Coatrieux, B. Wang, and S. Li. Geometric visual similarity learning in 3d medical image self-supervised pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9538–9547, 2023.

[22] V. Hondru, F. A. Croitoru, S. Minaee, R. T. Ionescu, and N. Sebe. Masked image modeling: A survey. *International Journal of Computer Vision*, 133(10):7154–7200, 2025.

[23] M. R. Hosseinzadeh Taher, J. Hong, R. Soni, and G. Avinash. Learning foundation models from multi-organ medical images by capturing consistency and diversity of anatomical structures. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 294–304, 2025.

[24] Y. Hou, S. Gould, and L. Zheng. Learning to select views for efficient multi-view understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20135–20144, 2024.

[25] D.-X. Huang, X.-H. Zhou, M.-J. Gui, X.-L. Xie, S.-Q. Liu, S.-Y. Wang, T.-Y. Xiang, R.-Z. Ma, N.-F. Xiao, and Z.-G. Hou. Vasomim: Vascular anatomy-aware masked image modeling for vessel segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 4985–4993, 2026.

[26] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[27] Y. Ishikawa, M. Kondo, and Y. Aoki. Data collection-free masked video modeling. In *European Conference on Computer Vision*, pages 37–56, 2024.

[28] Y. Ji, H. Bai, C. GE, J. Yang, Y. Zhu, R. Zhang, Z. Li, L. Zhanng, W. Ma, X. Wan, and P. Luo. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 36722–36732, 2022.

[29] Y. Jiang, M. Sun, H. Guo, X. Bai, K. Yan, L. Lu, and M. Xu. Anatomical invariance modeling and semantic alignment for self-supervised learning in 3d medical image analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15859–15869, 2023.

[30] I. Kakogeorgiou, S. Gidaris, B. Psomas, Y. Avrithis, A. Bursuc, K. Karantzalos, and N. Komodakis. What to hide from your students: Attention-guided masked image modeling. In *European Conference on Computer Vision*, pages 300–318, 2022.

[31] B. Kim, D. H. Kim, S. H. Park, J. Kim, J.-G. Lee, and J. C. Ye. Cyclemorph: Cycle consistent unsupervised deformable image registration. *Medical Image Analysis*, 71:102036, 2021.

[32] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. 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, page 12, 2015.

[33] J. Lee, V. Dang, G.-H. Yu, A. Le, Z. Rahman, J.-H. Jang, H. Lee, K.-Y. Kim, J.-S. Kim, and J.-Y. Kim. Hu-based foreground masking for 3d medical masked image modeling. In *International Workshop on Applications of Medical AI*, pages 122–131, 2025.

[34] W. Lei, W. Xu, K. Li, X. Zhang, and S. Zhang. Medlsam: Localize and segment anything model for 3d ct images. *Medical Image Analysis*, 99:103370, 2025.

[35] G. Li, H. Zheng, D. Liu, C. Wang, B. Su, and C. Zheng. Semmae: Semantic-guided masking for learning masked autoencoders. In *Advances in Neural Information Processing Systems*, volume 35, pages 14290–14302, 2022.

[36] P. Li, I. Abdullaeva, A. Gambashidze, A. Kuznetsov, and I. Oseledets. Maxinfo: A training-free key-frame selection method using maximum volume for enhanced video understanding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7198–7207, 2026.

[37] W. Li, C. Qu, X. Chen, P. R. Bassi, Y. Shi, Y. Lai, Q. Yu, H. Xue, Y. Chen, X. Lin, Y. Tang, Y. Cao, H. Han, Z. Zhang, J. Liu, T. Zhang, Y. Ma, J. Wang, G. Zhang, A. Yuille, and Z. Zhou. Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. *Medical Image Analysis*, 97:103285, 2024.

[38] W. Li, A. Yuille, and Z. Zhou. How well do supervised models transfer to 3d image segmentation. In *The Twelfth International Conference on Learning Representations*, volume 1, 2024.

[39] X. Li, Y. Yang, X. Li, J. Wu, Y. Yu, B. Ghanem, and M. Zhang. Genview: Enhancing view quality with pretrained generative model for self-supervised learning. In *European Conference on Computer Vision*, pages 306–325, 2024.

[40] Z. Li, Y. Zhu, F. Yang, W. Li, C. Zhao, Y. Chen, Z. Chen, J. Xie, L. Wu, R. Zhao, M. Tang, and J. Wang. Univip: A unified framework for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14627–14636, 2022.

[41] J. Liu, L. Shao, and L. Zhang. Meduet: Disentangled unified pretraining for 3d medical image synthesis and analysis. *arXiv preprint arXiv:2602.17901*, 2026.

[42] X. Luo, W. Liao, J. Xiao, J. Chen, T. Song, X. Zhang, K. Li, D. N. Metaxas, G. Wang, and S. Zhang. Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. *Medical Image Analysis*, 82:102642, 2022.

[43] Z. Luo, S. Rui, Z. Tang, G. Wang, and X. Wang. Invcoss: Inversion-driven continual self-supervised learning in medical multi-modal image pre-training. *arXiv preprint arXiv:2512.19213*, 2025.

[44] J. Ma, Y. Zhang, S. Gu, C. Ge, E. Wang, Q. Zhou, Z. Huang, P. Lyu, J. He, and B. Wang. Automatic organ and pan-cancer segmentation in abdomen ct: the flare 2023 challenge. *arXiv preprint arXiv:2408.12534*, 2024.

[45] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner. Open access series of imaging studies (oasis): Cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 19(9):1498–1507, 2007.

[46] S. Mo, H. Kang, K. Sohn, C.-L. Li, and J. Shin. Object-aware contrastive learning for debiased scene representation. In *Advances in Neural Information Processing Systems*, volume 34, pages 12251–12264, 2021.

[47] X. Ni, L. Wu, J. Zhuang, Q. Wang, M. Wu, V. Vardhanabhuti, L. Zhang, H. Gao, and H. Chen. Mg-3d: Multi-grained knowledge-enhanced vision-language pre-training for 3d medical image analysis. *Medical Image Analysis*, 111:104027, 2026.

[48] T. Pan, Z. Tan, K. Guo, D. Xu, W. Xu, C. Jiang, X. Guo, Y. Qi, and Y. Cheng. Structure-aware semantic discrepancy and consistency for 3d medical image self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20257–20267, 2025.

[49] X. Qi, G. Zhang, J. Wu, S. Zhang, X. Hou, and G. Wang. Fdas: Foundation model distillation and anatomic structure-aware multi-task learning for self-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 192–202, 2025.

[50] F. Quinton, R. Popoff, B. Presles, S. Leclerc, F. Meriaudeau, G. Nodari, O. Lopez, J. Pellegrinelli, O. Chevallier, D. Ginhac, J.-M. Vrigneaud, and J.-L. Alberini. A tumour and liver automatic segmentation (atlas) dataset on contrast-enhanced magnetic resonance imaging for hepatocellular carcinoma. *Data*, 8(5), 2023.

[51] M.-P. Revel, S. Boussouar, C. de Margerie-Mellon, I. Saab, T. Lapotre, D. Mompoint, G. Chassagnon, A. Milon, M. Lederlin, S. Bennani, et al. Study of thoracic ct in covid-19: The stoic project. *Radiology*, 301(1):E361–E370, 2021.

[52] S. Rui, L. Chen, Z. Tang, L. Wang, M. Liu, S. Zhang, and X. Wang. Multi-modal vision pre-training for medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5164–5174, 2025.

[53] A. A. A. Setio, A. Traverso, T. de Bel, M. S. Berens, C. van den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. Geurts, R. van der Gugten, P. A. Heng, B. Jansen, M. M. de Kaste, V. Kotov, J. Y.-H. Lin, J. T. Manders, A. Sóñora-Mengana, J. C. García-Naranjo, E. Papavasileiou, M. Prokop, M. Saletta, C. M. Schaefer-Prokop, E. T. Scholten, L. Scholten, M. M. Snoeren, E. L. Torres, J. Vandemeulebroucke, N. Walasek, G. C. Zuidhof, B. van Ginneken, and C. Jacobs. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The luna16 challenge. *Medical Image Analysis*, 42:1–13, 2017.

[54] M. R. H. Taher, M. B. Gotway, and J. Liang. Representing part-whole hierarchies in foundation models by learning localizability composability and decomposability from anatomy via self supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11269–11281, 2024.

[55] F. Tang, R. Xu, Q. Yao, X. Fu, Q. Quan, H. Zhu, Z. Liu, and S. K. Zhou. Hyspark: Hybrid sparse masking for large scale medical image pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 330–340, 2024.

[56] F. Tang, Q. Yao, W. Ma, C. Wu, Z. Jiang, and S. K. Zhou. Hi-end-mae: Hierarchical encoder-driven masked autoencoders are stronger vision learners for medical image segmentation. *Medical Image Analysis*, 107:103770, 2026.

[57] X. Tang, J. Qiu, L. Xie, Y. Tian, J. Jiao, and Q. Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29118–29128, 2025.

[58] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20730–20740, 2022.

[59] T. Tao, Z. Wang, H. Zhang, T. N. Arvanitis, and L. Zhang. Diffkan-inpainting: Kan-based diffusion model for brain tumor inpainting. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–4, 2025.

[60] T. Wald, C. Ulrich, S. Lukyanenko, A. Goncharov, A. Paderno, M. Miller, L. Maerkisch, P. Jaeger, and K. Maier-Hein. Revisiting mae pre-training for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5186–5196, 2025.

[61] T. Wald, C. Ulrich, J. Suprijadi, S. Ziegler, M. Nohel, R. Peretzke, G. Kohler, and K. Maier-Hein. An openmind for 3d medical vision self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23839–23879, 2025.

[62] H. Wang, J. Fan, Y. Wang, K. Song, T. Wang, X. Zhang, and Z. Zhang. Bootstrap masked visual modeling via hard patch mining. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6200–6214, 2025.

[63] H. Wang, K. Song, J. Fan, Y. Wang, J. Xie, and Z. Zhang. Hard patches mining for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10375–10385, 2023.

[64] S. Wang, M. Safari, Q. Li, C.-W. Chang, R. LJ Qiu, J. Roper, D. S. Yu, and X. Yang. Vision foundation model for 3d magnetic resonance imaging segmentation, classification, and registration. *Medical Image Analysis*, 110:103992, 2026.

[65] Y. Wang, Z. Li, J. Mei, Z. Wei, L. Liu, C. Wang, S. Sang, A. L. Yuille, C. Xie, and Y. Zhou. Swinmm: masked multi-view with swin transformers for 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 486–496, 2023.

[66] Z. Wang, S. Yu, E. Stengel-Eskin, J. Yoon, F. Cheng, G. Bertasius, and M. Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3272–3283, 2025.

[67] J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. W. Sauter, T. Heye, D. T. Boll, J. Cyriac, S. Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.

[68] L. Wu, J. Zhuang, and H. Chen. Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22873–22882, 2024.

[69] L. Wu, J. Zhuang, and H. Chen. Large-scale 3d medical image pre-training with geometric context priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3801–3818, 2026.

[70] W. Xiang, C. Liu, H. Su, and H. Yu. Aemim: Adversarial examples meet masked image modeling. *International Journal of Computer Vision*, 134(4):159, 2026.

[71] Y. Xie, J. Zhang, Y. Xia, and Q. Wu. Unimiss+: Universal medical self-supervised learning from cross-dimensional unpaired data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10021–10035, 2024.

[72] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9653–9663, 2022.

[73] T. Xu, S. Hosseini, C. Anderson, A. Rinaldi, R. G. Krishnan, A. L. Martel, and M. Goubran. A generalizable 3d framework and model for self-supervised learning in medical imaging. *npj Digital Medicine*, 8(1):639, 2025.

[74] K. Zhang, X. Liu, J. Shen, Z. Li, Y. Sang, X. Wu, Y. Zha, W. Liang, C. Wang, K. Wang, et al. Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography. *Cell*, 181(6):1423–1433, 2020.

[75] S. Zhang, J. Yang, J. Yin, Z. Luo, and J. Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22056–22065, 2025.

[76] Y. Zhang, H. Zhai, J. Guo, Z. Li, J. Liu, and H. Han. Masked image modeling for generalizable organelle segmentation in volume em. *IEEE Transactions on Medical Imaging*, pages 1–1, 2026.

[77] C. Zhao, Y. Lyu, A. Z. Chowdury, E. S. Harake, A. Kondepudi, A. T. Rao, X. Hou, H. Lee, and T. C. Hollon. Towards scalable language-image pre-training for 3d medical imaging. *Transactions on Machine Learning Research*, 2026.

[78] Z. Zhao, Y. Zhang, C. Wu, X. Zhang, X. Zhou, Y. Zhang, Y. Wang, and W. Xie. Large-vocabulary segmentation for medical images with text prompts. *NPJ Digital Medicine*, 8(1):566, 2025.

[79] H.-Y. Zhou, C. Lu, S. Yang, X. Han, and Y. Yu. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3499–3509, 2021.

[80] K. Zhu, Y.-Y. He, and J. Wu. Coarse is better? a new pipeline towards self-supervised learning with uncurated images. *Pattern Recognition*, 161:111324, 2025.

[81] J. Zhuang, L. Luo, Q. Wang, M. Wu, L. Luo, and H. Chen. Advancing volumetric medical image segmentation via global-local masked autoencoders. *IEEE Transactions on Medical Imaging*, 44(11):4226–4238, 2025.

[82] J. Zhuang, L. Wu, Q. Wang, P. Fei, V. Vardhanabhuti, L. Luo, and H. Chen. Mim: Mask in mask self-supervised pre-training for 3d medical image analysis. *IEEE Transactions on Medical Imaging*, 44(9):3727–3740, 2025.

Appendix

A Datasets

A.1 Pretraining Datasets

For each SSL backbone, we use the same pretraining dataset as in its original work and follow the corresponding official implementation to ensure a fair and consistent comparison.

VoCo [68]. Following the default VoCo setup, we pretrain VoCo on VoCo-10k [68], a large-scale CT collection comprising BTCV [32], TCIA Covid19 [10], LUNA16 [53], STOIC 2021 [51], TotalSegmentator [67], FLARE23 [44], and HNSCC [15]. In total, VoCo-10k contains approximately 9.9k CT volumes. We use the released open-source data splits from prior work [68, 81, 82]. A detailed summary of the pretraining datasets is provided in Table 5.

S2DC [48]. Following the original S2DC setting, we also use VoCo-10k [68] as the pretraining dataset, resulting in approximately 9.9k CT volumes.

MedGMAE [13]. For MedGMAE, we follow its original pretraining setup and use AbdomenAtlas-Mini 1.0 [37], which contains about 5.2k CT volumes, as summarized in Table 5.

Table 5: Overview of pretraining datasets.

Dataset	Modality	Region of Interest	#Volumes
BTCV [32]	CT	Abdomen	24
TCIA Covid19 [10]	CT	Chest	722
LUNA16 [53]	CT	Chest	843
STOIC 2021 [51]	CT	Chest	2,000
FLARE23 [44]	CT	Abdomen	4,000
HNSCC [15]	CT	Head/Neck	1,071
TotalSegmentator [67]	CT	Head/Neck/Chest/Leg/ Abdomen/Pelvis/Feet	1,203
AbdomenAtlasMini 1.0 [37]	CT	Abdomen	5,195

A.2 Downstream Datasets

Details of the downstream datasets and splits are provided in Table 6. We evaluate the pretrained representations on three downstream task families. For segmentation, we use BTCV [32], AMOS [28], WORD [42], and Atlas-MRI [50]. For classification, we use CC-CCII [74]. For registration, we use IXI [31] and OASIS [45]. Following prior work [68, 69, 13, 56], we adopt the established data splits for each benchmark whenever available to ensure comparable evaluation.

B Implementation Details

B.1 Pretraining Setup

Table 7 presents the baseline-specific preprocessing and pretraining settings. We strictly follow the official setup of VoCo [68], S2DC [48], and MedGMAE [13], including the preprocessing pipeline, augmentation strategy, optimizer, learning-rate schedule, and pretraining length. This ensures that the comparison isolates the effect of the sub-volume sampling policy rather than changes in backbone-specific training strategies. The key differences are their sub-volume configurations, namely, VoCo uses one 64^3 crop, S2DC uses one 96^3 crop, and MedGMAE uses three 96^3 crops per volume. For MedGMAE, we independently sample three crop centers either from the cached probe bank or via standard random cropping at each iteration, without imposing an additional diversity constraint. This configuration keeps MedGMAE's original three-crop training scheme unchanged, while replacing only the source of crop centers.

During pretraining, we sample 70% of crop centers from the cached probe bank and the remaining 30% via standard random cropping. For bank-sampled centers, we apply a small local jitter independently along each spatial axis. Specifically, the center is shifted by a uniformly sampled integer offset within

Table 6: Overview of downstream datasets.

Task	Dataset	Modality	#Train				#Valid
			Full	1%	10%	100%	
Segmentation	BTCV [32]	CT	24	1	2	24	6
	AMOS [28]	CT	200	2	20	200	100
	WORD [42]	CT	100	1	10	100	20
	Atlas-MRI [50]	MRI	45	1	5	45	15
Classification	CC-CCII [74]	CT	2,514	–	–	–	1,664
Registration	IXI [31]	MRI	403	–	–	–	115
	OASIS [45]	MRI	394	–	–	–	19

Table 7: Overview of preprocessing and pretraining settings used for pretraining.

Setting	VoCo [68]	S2DC [48]	MedGMAE [13]
Preprocessing Settings			
Spacing	$1.5 \times 1.5 \times 1.5$	$1.5 \times 1.5 \times 1.5$	$1.5 \times 1.5 \times 1.5$
Intensity	$[-175, 250]$	$[-175, 250]$	$[-175, 250]$
Norm	$[0, 1]$	$[0, 1]$	$[0, 1]$
Sub-volume size	$64 \times 64 \times 64$	$96 \times 96 \times 96$	$96 \times 96 \times 96$
Sub-volume number	1	1	3
Augmentation	Random Rotate and Flip	Random Rotate and Flip	Random Rotate, Flip, Scale, Shift
Pretraining Settings			
Pretraining steps	100k	100k	100k
Optimizer	AdamW	AdamW	AdamW
Weight decay	1e-2	1e-1	0
Optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$	$\beta_1, \beta_2 = 0.9, 0.999$	$\beta_1, \beta_2 = 0.9, 0.95$
Optimizer LR	1e-3	4e-4	1e-5
Batch size	4	4	8
LR schedule	Warmup cosine	Warmup cosine	Warmup cosine
Warmup steps	2k	0.5k	2k

$\pm\text{round}(\gamma r_a)$, where r_a is the sub-volume size on axis a and $\gamma = 0.05$ by default, and the jittered center is clipped to the valid image range. All methods are pretrained for 100k steps with AdamW and a warmup cosine schedule. `VolumeProbe` only replaces random sampling with cached probe-bank sampling, leaving all other settings unchanged.

B.2 Downstream Settings

For segmentation and classification, we follow the same preprocessing and fine-tuning protocols as prior work [68, 69], as summarized in Table 8. Specifically, VoCo [68] and S2DC [48] use Swin-UNETR [17] as the downstream backbone, while MedGMAE [13] uses UNETR [18]. Following previous work [56, 13], we fine-tune segmentation models for 5000, 2000, and 1000 epochs under 1%, 10%, and 100% label regimes, respectively. For registration, we use TransMorph [6] as the downstream baseline, with preprocessing and training protocols kept consistent with VoCo [68] for fair comparison.

B.3 Heterogeneity of Evaluated SSL Frameworks

Although all evaluated medical SSL frameworks [68, 48, 13] leverage sub-volume cropping, this is exactly the interface targeted by our `VolumeProbe`. Notably, the three frameworks cover different SSL objectives, downstream architectures, and crop configurations. Concretely, VoCo [68] and S2DC [48] are contrastive-style methods using Swin-UNETR [17], whereas MedGMAE [13] follows masked modeling and uses UNETR [18]. They also differ in crop size and number, with VoCo using one 64^3 crop, S2DC using one 96^3 crop, and MedGMAE using three 96^3 crops per volume. These differences highlight that `VolumeProbe` is evaluated across distinct objectives, architectures, and single- versus multi-crop interfaces, supporting its plug-and-play applicability to representative 3D medical SSL pipelines.

Table 8: Overview of preprocessing and fine-tuning settings for downstream evaluation.

Preprocessing Settings	
Spacing	$1.5 \times 1.5 \times 1.5$
Intensity	$[-175, 250]$
Norm	$[0, 1]$
Sub-volume size	$96 \times 96 \times 96$
Sub-volume number	4
Augmentation	Random Rotate, Flip, Scale, Shift
Fine-tuning Settings	
Optimizer	AdamW
Weight decay	1e-5
Optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$
Optimizer LR	3e-4
Batch size	$1 \times 4 = 4$
Inference	Sliding Window Inference
Inference overlap	0.75
Sw batch size	4

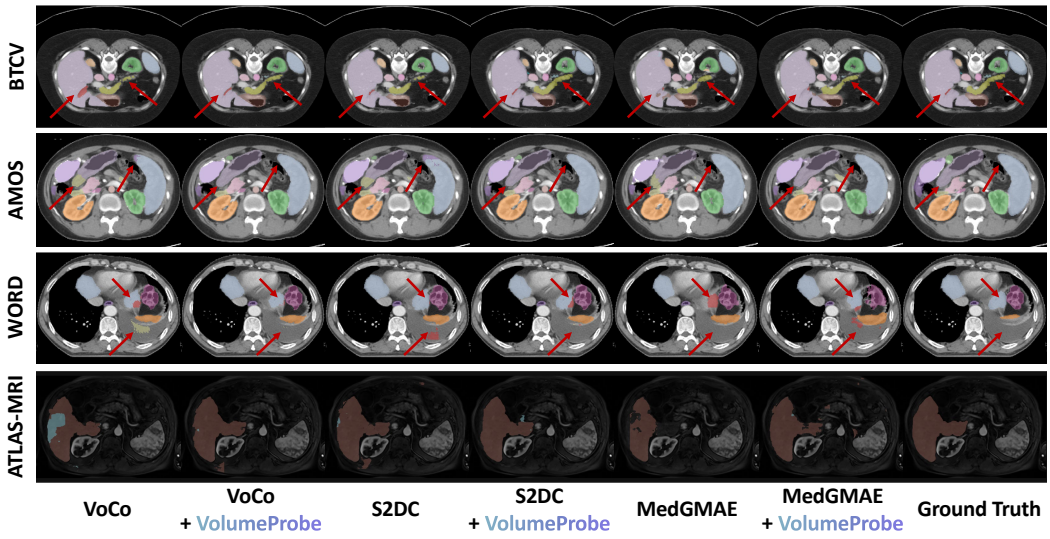


Figure 10: Qualitative visualization of segmentation results under the 10% label setting.

B.4 Frozen Feature Encoder for `VolumeProbe-Feat`

For `VolumeProbe-Feat`, we use 3DINO-ViT [73] as a frozen feature encoder only for offline probe-bank construction. 3DINO-ViT is pretrained on a large-scale multimodal 3D medical imaging corpus of about 100k 3D scans, covering CT, MRI, and PET scans across diverse anatomical regions, rather than on natural images. We regard `VolumeProbe-Feat` as an external-feature variant and use it mainly to study whether richer semantic features further improve probe selection. It is worth noting that `VolumeProbe-Lite` does not use any external encoder and already shows consistent gains, so our main conclusion does not rely on 3DINO-ViT pretraining data. Moreover, rather than introducing new cues, we highlight observation policy as an underexplored aspect and show that simple lightweight cues of `VolumeProbe-Lite` already suffice under this *where-to-look* formulation.

C More Experimental Results

C.1 Multi-run Robustness Analysis

We further evaluate whether the main downstream gains are sensitive to random seeds. For the 100% label setting and registration, we keep the official train/test splits fixed and repeat downstream

Table 9: **Multi-run robustness on segmentation tasks.** Dice Score (%) is reported as mean$_{\pm \mathrm{std}}$ over three independent runs.

Method	Dice Score (%) ↑											
	BTCV			AMOS			WORD			Atlas-MRI		
	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%
VoCo [68]	63.33$_{\pm 1.5}$	77.85$_{\pm 0.7}$	83.85$_{\pm 0.2}$	55.81$_{\pm 2.3}$	73.34$_{\pm 0.6}$	84.44$_{\pm 0.1}$	64.24$_{\pm 1.2}$	77.28$_{\pm 0.4}$	84.06$_{\pm 0.2}$	34.85$_{\pm 1.9}$	47.56$_{\pm 0.9}$	58.83$_{\pm 0.3}$
+ VolumeProbe-Lite	67.21$_{\pm 1.3}$	79.39$_{\pm 0.5}$	83.97$_{\pm 0.1}$	58.64$_{\pm 1.5}$	74.95$_{\pm 0.6}$	**87.52**$_{\pm 0.1}$	68.93$_{\pm 1.6}$	**80.12**$_{\pm 0.5}$	**85.56**$_{\pm 0.2}$	38.68$_{\pm 2.2}$	49.27$_{\pm 1.0}$	59.98$_{\pm 0.2}$
+ VolumeProbe-Feat	68.07$_{\pm 1.5}$	**79.51**$_{\pm 0.9}$	83.89$_{\pm 0.1}$	57.86$_{\pm 1.9}$	74.70$_{\pm 0.4}$	87.13$_{\pm 0.1}$	68.42$_{\pm 1.7}$	79.74$_{\pm 0.4}$	84.99$_{\pm 0.2}$	37.93$_{\pm 2.4}$	49.06$_{\pm 1.2}$	59.42$_{\pm 0.2}$
S2DC [48]	66.77$_{\pm 1.8}$	77.93$_{\pm 0.8}$	84.09$_{\pm 0.1}$	55.19$_{\pm 1.6}$	74.28$_{\pm 0.9}$	84.68$_{\pm 0.2}$	69.14$_{\pm 2.0}$	78.96$_{\pm 0.6}$	84.46$_{\pm 0.1}$	33.79$_{\pm 2.1}$	46.18$_{\pm 0.8}$	58.53$_{\pm 0.2}$
+ VolumeProbe-Lite	**69.47**$_{\pm 1.5}$	78.74$_{\pm 0.9}$	**84.43**$_{\pm 0.1}$	58.83$_{\pm 2.1}$	75.31$_{\pm 0.7}$	87.17$_{\pm 0.1}$	69.73$_{\pm 1.5}$	80.04$_{\pm 0.3}$	85.22$_{\pm 0.1}$	37.16$_{\pm 2.7}$	48.08$_{\pm 0.9}$	59.44$_{\pm 0.3}$
+ VolumeProbe-Feat	68.76$_{\pm 1.3}$	78.92$_{\pm 1.0}$	84.19$_{\pm 0.1}$	59.47$_{\pm 1.4}$	75.17$_{\pm 0.3}$	86.87$_{\pm 0.1}$	70.28$_{\pm 1.9}$	79.82$_{\pm 0.5}$	85.43$_{\pm 0.2}$	38.19$_{\pm 2.2}$	48.83$_{\pm 1.0}$	59.81$_{\pm 0.2}$
MedGMAE [13]	66.19$_{\pm 2.0}$	77.11$_{\pm 1.1}$	83.22$_{\pm 0.2}$	58.79$_{\pm 2.4}$	75.65$_{\pm 0.8}$	84.90$_{\pm 0.1}$	68.09$_{\pm 1.7}$	77.73$_{\pm 0.6}$	84.19$_{\pm 0.1}$	36.24$_{\pm 2.9}$	48.16$_{\pm 1.2}$	59.42$_{\pm 0.3}$
+ VolumeProbe-Lite	66.85$_{\pm 1.7}$	78.44$_{\pm 1.0}$	83.48$_{\pm 0.1}$	**60.26**$_{\pm 1.9}$	**77.14**$_{\pm 0.9}$	85.62$_{\pm 0.1}$	70.67$_{\pm 1.9}$	79.42$_{\pm 0.7}$	84.36$_{\pm 0.1}$	**39.71**$_{\pm 2.5}$	**50.23**$_{\pm 1.2}$	60.19$_{\pm 0.3}$
+ VolumeProbe-Feat	66.57$_{\pm 2.1}$	79.09$_{\pm 0.7}$	83.29$_{\pm 0.1}$	59.57$_{\pm 1.5}$	76.80$_{\pm 0.6}$	85.93$_{\pm 0.2}$	**71.15**$_{\pm 1.4}$	79.08$_{\pm 0.9}$	84.25$_{\pm 0.1}$	39.48$_{\pm 2.2}$	49.83$_{\pm 1.0}$	**60.40**$_{\pm 0.2}$

Table 10: **Multi-run robustness on classification tasks.**

Method	Accuracy (%) ↑
	CC-CCII
VoCo [68]	93.45$_{\pm 0.4}$
+ VolumeProbe-Lite	93.98$_{\pm 0.9}$
+ VolumeProbe-Feat	94.25$_{\pm 0.5}$
S2DC [48]	94.19$_{\pm 1.1}$
+ VolumeProbe-Lite	**94.87**$_{\pm 0.8}$
+ VolumeProbe-Feat	94.34$_{\pm 1.2}$
MedGMAE [13]	93.08$_{\pm 0.6}$
+ VolumeProbe-Lite	93.58$_{\pm 0.5}$
+ VolumeProbe-Feat	93.81$_{\pm 0.6}$

Table 11: **Multi-run robustness on registration tasks.**

Method	Dice Score (%) ↑	
	IXI	OASIS
VoCo [68]	73.66$_{\pm 0.4}$	84.38$_{\pm 0.3}$
+ VolumeProbe-Lite	74.39$_{\pm 0.3}$	85.55$_{\pm 0.5}$
+ VolumeProbe-Feat	74.17$_{\pm 0.4}$	85.19$_{\pm 0.3}$
S2DC [48]	72.93$_{\pm 0.2}$	82.97$_{\pm 0.4}$
+ VolumeProbe-Lite	73.70$_{\pm 0.4}$	84.09$_{\pm 0.6}$
+ VolumeProbe-Feat	73.98$_{\pm 0.3}$	83.86$_{\pm 0.6}$
MedGMAE [13]	73.69$_{\pm 0.5}$	85.71$_{\pm 0.3}$
+ VolumeProbe-Lite	**74.88**$_{\pm 0.2}$	85.96$_{\pm 0.4}$
+ VolumeProbe-Feat	74.28$_{\pm 0.4}$	**86.17**$_{\pm 0.2}$

training with three different optimization seeds. For the 1% and 10% label settings, we use three labeled-subset seeds, where each seed defines the same labeled subset for the random-cropping baseline, `VolumeProbe-Lite` and `VolumeProbe-Feat`, together with the corresponding downstream optimization seed. This paired protocol isolates the effect of the sampling policy while avoiding changes to the official evaluation split.

As reported in Tables 9, 10, and 11, `VolumeProbe` consistently improves the random-cropping baselines across segmentation, classification, and registration under multi-run evaluation. The variance is larger in low-label segmentation and classification, while the 100% label and registration settings remain relatively stable. The observed gains remain consistent across runs, indicating that `VolumeProbe` is robust to seed-level variation across downstream tasks and SSL backbones.

C.2 Subject-level Variability and Paired Significance

To complement the run-level robustness analysis, we further conduct a subject-level analysis over the same three matched downstream runs. For each test subject, we first average its Dice scores across the three runs, and then compute the mean and standard deviation across subjects. Statistical significance is evaluated using a two-sided paired permutation test between `VolumeProbe-Lite` and its corresponding random-cropping baseline. This protocol evaluates whether the observed improvements remain consistent at the subject level rather than only across optimization runs. As shown in Table 12, `VolumeProbe` reaches $p \leq 0.05$ in 33 of the 36 paired comparisons, while all mean improvements remain positive. The results therefore indicate that the gains are consistent at the subject level across the matched three-run aggregate.

C.3 Visualization Results

Fig. 10 provides qualitative segmentation results under the 10% label setting across four downstream datasets. Compared with the corresponding SSL baselines, adding `VolumeProbe` produces more complete and anatomically aligned predictions. These visual results further support that improved sub-volume sampling leads to more transferable representations across different SSL backbones.

Table 12: **Subject-level variability and paired significance on segmentation tasks.** Dice Score (%) is reported with subject-level SD in subscript and significance in superscript. Superscripts denote two-sided paired permutation tests against the corresponding random-cropping baseline: n.s., $p > 0.05$; *, $p \leq 0.05$; **, $p \leq 0.01$; ***, $p \leq 0.001$; ****, $p \leq 0.0001$. For BTCV ($N = 6$), the minimum exact two-sided p-value is 0.03125, so significance is capped at *.

Method	Dice Score (%) ↑											
	BTCV			AMOS			WORD			Atlas-MRI		
	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%
VoCo [68]	63.33 $_{\pm 5.9}$	77.85 $_{\pm 7.2}$	83.85 $_{\pm 1.8}$	55.81 $_{\pm 3.9}$	73.34 $_{\pm 4.5}$	84.44 $_{\pm 1.6}$	64.24 $_{\pm 3.2}$	77.28 $_{\pm 2.9}$	84.06 $_{\pm 1.2}$	34.85 $_{\pm 5.5}$	47.56 $_{\pm 4.3}$	58.83 $_{\pm 1.2}$
+ `VolumeProbe`	67.21 $^{*}_{\pm 6.8}$	79.39 $^{*}_{\pm 8.6}$	83.97 $^{\text{n.s.}}_{\pm 1.3}$	58.64 $^{***}_{\pm 4.0}$	74.95 $^{****}_{\pm 4.2}$	87.52 $^{****}_{\pm 1.4}$	68.93 $^{****}_{\pm 3.7}$	80.12 $^{****}_{\pm 3.5}$	85.56 $^{****}_{\pm 1.5}$	38.68 $^{**}_{\pm 4.2}$	49.27 $^{\text{n.s.}}_{\pm 3.5}$	59.98 $^{**}_{\pm 1.0}$
S2DC [48]	66.77 $_{\pm 8.2}$	77.93 $_{\pm 5.7}$	84.09 $_{\pm 2.3}$	55.19 $_{\pm 4.2}$	74.28 $_{\pm 3.9}$	84.68 $_{\pm 1.5}$	69.14 $_{\pm 4.0}$	78.96 $_{\pm 4.3}$	84.46 $_{\pm 1.3}$	33.79 $_{\pm 4.6}$	46.18 $_{\pm 5.7}$	58.53 $_{\pm 1.0}$
+ `VolumeProbe`	69.47 $^{*}_{\pm 7.4}$	78.74 $^{*}_{\pm 5.6}$	84.43 $^{*}_{\pm 1.5}$	58.83 $^{****}_{\pm 4.0}$	75.31 $^{****}_{\pm 4.1}$	87.17 $^{****}_{\pm 1.9}$	69.73 $^{***}_{\pm 3.5}$	80.04 $^{***}_{\pm 3.9}$	85.22 $^{****}_{\pm 1.0}$	37.16 $^{**}_{\pm 5.0}$	48.08 $^{**}_{\pm 3.8}$	59.44 $^{**}_{\pm 0.9}$
MedGMAE [13]	66.19 $_{\pm 5.9}$	77.11 $_{\pm 6.9}$	83.22 $_{\pm 1.9}$	58.79 $_{\pm 4.2}$	75.65 $_{\pm 5.0}$	84.90 $_{\pm 2.1}$	68.09 $_{\pm 3.4}$	77.73 $_{\pm 2.9}$	84.19 $_{\pm 1.2}$	36.24 $_{\pm 4.3}$	48.16 $_{\pm 3.9}$	59.42 $_{\pm 1.2}$
+ `VolumeProbe`	66.85 $^{*}_{\pm 6.3}$	78.44 $^{\text{n.s.}}_{\pm 7.2}$	83.48 $^{*}_{\pm 2.4}$	60.26 $^{****}_{\pm 4.3}$	77.14 $^{****}_{\pm 4.4}$	85.62 $^{***}_{\pm 2.2}$	70.67 $^{***}_{\pm 3.9}$	79.42 $^{***}_{\pm 3.5}$	84.36 $^{****}_{\pm 1.5}$	39.71 $^{**}_{\pm 4.0}$	50.23 $^{**}_{\pm 4.2}$	60.19 $^{***}_{\pm 1.4}$

Table 13: **Evaluation under nnU-Net pipelines.** Subject-level Dice Score (%) is reported as mean ± SD. All variants within each dataset share the same auto-configured nnU-Net 3D full-resolution plan and training protocol.

Method	Dice Score (%) ↑			
	AMOS		WORD	
	1%	100%	1%	100%
nnU-Net (from scratch)	29.53$_{\pm 4.7}$	84.89$_{\pm 1.9}$	49.82$_{\pm 3.3}$	84.66$_{\pm 1.5}$
VoCo (nnU-Net) [68]	52.96$_{\pm 3.4}$	85.22$_{\pm 2.0}$	65.47$_{\pm 3.9}$	84.89$_{\pm 1.2}$
+ `VolumeProbe`	**55.43**$_{\pm 3.5}$	**86.90**$_{\pm 1.8}$	**67.53**$_{\pm 4.2}$	**85.37**$_{\pm 1.5}$
S2DC (nnU-Net) [48]	49.86$_{\pm 2.8}$	85.38$_{\pm 2.2}$	67.21$_{\pm 3.1}$	84.83$_{\pm 1.4}$
+ `VolumeProbe`	**54.05**$_{\pm 3.2}$	**86.82**$_{\pm 2.0}$	**70.34**$_{\pm 4.0}$	**85.46**$_{\pm 1.1}$

C.4 Evaluation under nnU-Net Pipelines

Following [69], we use the official nnU-Net pipeline [26] and perform dataset fingerprinting, planning, and preprocessing separately for AMOS [28] and WORD [42]. Within each dataset, all variants share the same auto-configured 3D full-resolution plan, data split, and training protocol, while differing only in whether the pretrained weights are obtained using Random or `VolumeProbe` sampling. Table 13 confirms that `VolumeProbe` consistently improves both VoCo [68] and S2DC [48] across the two datasets and label regimes, demonstrating that its benefits persist under a strong, systematically configured segmentation framework.

C.5 Experimental Details for Simple Non-random Samplers

We provide additional experimental details for the simple non-random samplers compared in Table 4. All samplers require no labels, learnable parameters, or online model feedback. For a fair comparison, each method follows the same pretraining setup as `VolumeProbe`. For every preprocessed volume, we construct an offline bank of K crop centers and follow the same mixed sampling policy as `VolumeProbe`, with 70% bank-based sampling and 30% standard random cropping during pretraining. The crop size, number of crops, SSL backbone, augmentation, optimizer, learning-rate schedule, and total pretraining steps are kept unchanged across all sampling strategies.

We consider four simple non-random policies:

- **Center-Biased.** Selects candidates closest to the center of the foreground-cropped volume, isolating the effect of a simple centrality prior.
- **Uniform-Grid.** Distributes crop centers approximately uniformly over the valid crop-center space, isolating spatial coverage.
- **Grad-Only.** Ranks candidates by the mean local gradient magnitude, capturing local structural variation.
- **Entropy-Only.** Ranks candidates by local intensity entropy, capturing local intensity heterogeneity.

Table 14: **Occupancy ratios of sampled sub-volumes.** Foreground occupancy measures the average foreground voxel ratio within sampled crops, while organ occupancy measures the average organ-mask voxel ratio on the mask-available subset.

Method	Foreground Occupancy (%) ↑	Organ Occupancy (%) ↑
Random	55.83	19.79
Center-Biased	69.12$_{\uparrow 13.29}$	34.61$_{\uparrow 14.82}$
Uniform-Grid	62.74$_{\uparrow 6.91}$	28.57$_{\uparrow 8.78}$
Grad-Only	71.34$_{\uparrow 15.51}$	32.90$_{\uparrow 13.11}$
Entropy-Only	68.76$_{\uparrow 12.93}$	36.08$_{\uparrow 16.29}$
`VolumeProbe-Lite`	75.63$_{\uparrow 19.80}$	45.71$_{\uparrow 25.92}$
`VolumeProbe-Feat`	77.19$_{\uparrow 21.36}$	44.92$_{\uparrow 25.13}$

Table 15: **Additional ablation of principles on classification and registration tasks.**

Method	CC-CCII Acc. (%) ↑	IXI Dice (%) ↑
VoCo [68]	93.45	73.66
+ `VolumeProbe`	**93.98**$_{\uparrow 0.53}$	**74.39**$_{\uparrow 0.73}$
w/o A	93.53$_{\uparrow 0.08}$	74.05$_{\uparrow 0.39}$
w/o I	93.61$_{\uparrow 0.16}$	73.59$_{\downarrow 0.07}$
w/o D	93.78$_{\uparrow 0.33}$	74.12$_{\uparrow 0.46}$

Table 16: **Relative gain retained after removing each criterion.**

Removed Criterion	Seg. (%)	Cls. (%)	Reg. (%)
w/o A	20.2	15.1	53.4
w/o I	51.9	30.2	-9.6
w/o D	84.5	62.3	63.0

Together, these policies isolate simple spatial and image-statistical cues while keeping the remaining pretraining pipeline unchanged.

C.6 Occupancy Analysis across Sampling Strategies

To complement the downstream comparison in Table 4, we further examine how different sampling strategies allocate the sub-volume crop budget across anatomically valid regions. We report the foreground occupancy ratio and organ occupancy ratio on VoCo-10k [68, 69] in Table 14. Foreground occupancy measures the fraction of crop voxels within the body foreground, while organ occupancy measures the fraction overlapping annotated organs on the mask-available subset, including BTCV [32], LUNA16 [53], FLARE23 [44], and TotalSegmentator [67].

Random achieves only 55.83% foreground occupancy and 19.79% organ occupancy, while simple non-random policies improve these ratios to varying degrees. Nevertheless, they remain below `VolumeProbe-Lite`, which reaches 75.63% foreground occupancy and 45.71% organ occupancy. `VolumeProbe-Feat` further increases foreground occupancy to 77.19% while achieving 44.92% organ occupancy. Taken together, simple spatial or statistical heuristics partially reduce observation waste, while the joint consideration of anatomical plausibility, informativeness, and diversity yields more anatomically valid and organ-rich observations.

C.7 Ablation Study

Task-dependent contribution of the three principles. We further examine whether anatomy-awareness (A), informativeness (I), and diversity (D) contribute beyond segmentation. Following Sec. 4.4, we separately remove each criterion using VoCo [68] and evaluate on CC-CCII [74] classification and IXI [31] registration, while keeping all other settings unchanged. As reported in Table 15, the complete objective achieves the best performance across both metrics. Removing informativeness causes the largest degradation on IXI, while removing anatomy-awareness leads to the largest reduction in classification performance.

To quantify task dependence, we further compute the fraction of the full `VolumeProbe` gain retained after removing each criterion, $(\text{Ablation} - \text{VoCo})/(\text{VolumeProbe} - \text{VoCo})$. For segmentation, we use the mean Dice gain over BTCV [32], AMOS [28], and WORD [42] under the 100% label setting; for classification and registration, we use accuracy and Dice, respectively. Table 16 indicates that anatomy-awareness contributes more strongly to segmentation and classification, whereas

Table 17: Fine-grained ablation of principle terms for VoCo [68]. We remove each term from the anatomy-aware or informativeness score while keeping the remaining pipeline unchanged.

Variant	Anatomy-aware		Informative		Dice Score (%) ↑		
	LocPrior	Margin	Grad	Entropy	BTCV	AMOS	WORD
VoCo [68]	–	–	–	–	83.85	84.44	84.06
+ VolumeProbe	✓	✓	✓	✓	$83.97_{\uparrow 0.12}$	$87.52_{\uparrow 3.08}$	$85.56_{\uparrow 1.50}$
w/o LocPrior	×	✓	✓	✓	$83.90_{\uparrow 0.05}$	$85.60_{\uparrow 1.16}$	$84.21_{\uparrow 0.15}$
w/o Margin	✓	×	✓	✓	$83.87_{\uparrow 0.02}$	$86.73_{\uparrow 2.29}$	$85.40_{\uparrow 1.34}$
w/o Grad	✓	✓	×	✓	$83.91_{\uparrow 0.06}$	$87.13_{\uparrow 2.69}$	$85.36_{\uparrow 1.30}$
w/o Entropy	✓	✓	✓	×	$83.95_{\uparrow 0.10}$	$86.97_{\uparrow 2.53}$	$85.47_{\uparrow 1.41}$

Table 18: Ablation study of coarse-bin partition strategy for VoCo [68].

Coarse Bins Partition	Dice Score (%) ↑		
	BTCV	AMOS	WORD
(G_x, G_y, G_z)			
$(3, 3, 1)$	81.90	84.46	83.80
$(3, 3, 2)$	82.57	84.51	83.71
$(4, 4, 1)$	83.40	86.27	85.09
$(4, 4, 2)$	83.97	87.52	85.56
$(5, 5, 1)$	83.26	86.75	84.92
$(5, 5, 2)$	83.89	87.04	85.37

informativeness is particularly important for registration. Diversity has a smaller but consistently beneficial contribution across the evaluated task families.

Principle terms. Table 17 provides a fine-grained ablation of the anatomy-aware and informativeness scores. Removing any individual term consistently weakens performance compared with the full VolumeProbe, confirming that LocPrior, Margin, Grad, and Entropy each contribute to probe-bank quality. Among them, removing LocPrior causes the most pronounced degradation on AMOS [28] and WORD [42], indicating that anatomical plausibility is especially important for selecting effective sub-volumes.

Coarse-bin Partition. Table 18 studies the effect of the coarse partition resolution used in candidate generation. A moderately anisotropic grid, $(G_x, G_y, G_z) = (4, 4, 2)$, achieves the best performance across three downstream datasets, reflecting that it provides a favorable balance between spatial localization and candidate coverage. Coarser partitions overlook informative regions, while overly fine partitions introduce fragmented candidate pools and slightly degrade the probe-bank quality.

Sensitivity to M and K. Tables 19 and 20 verify the sensitivity to the number of retained coarse bins M and probe-bank size K. For M, performance improves as more coarse bins are retained and peaks overall at $M = 24$, while further increasing M to 32 brings no consistent gain, confirming that overly many coarse candidates may introduce less informative regions. For K, enlarging the probe bank steadily improves performance up to $K = 40$, after which the gains saturate and slightly fluctuate. We therefore use $M = 24$ and $K = 40$ as the default setting, which provides a good balance between candidate diversity and probe quality.

Sensitivity to α and β. Tables 21 and 22 study the sensitivity to the informativeness weight α and diversity weight β. Performance remains stable across $\alpha, \beta \in \{0.5, 1.0, 1.5\}$, with $\alpha = 1.0$ and $\beta = 1.0$ achieving the best overall results across BTCV, AMOS, and WORD. The stable performance confirms that VolumeProbe is not overly sensitive to moderate changes in these weighting hyperparameters, and we use $\alpha = \beta = 1.0$ by default.

C.8 Convergence under Extended Pretraining

To examine whether the gains of VolumeProbe mainly reflect faster convergence or persist under longer pretraining, we extend VoCo [68] pretraining from the original 100k steps to 160k steps under matched continuation schedules for both *Random* sampling and VolumeProbe. Starting from the

Table 19: Ablation study of M for VoCo [68].

M	Dice Score (%) ↑		
	BTCV	AMOS	WORD
8	82.16	84.83	83.71
16	82.79	86.12	84.29
24	83.97	87.52	85.56
32	83.71	87.60	85.11

Table 20: Ablation study of K for VoCo [68].

K	Dice Score (%) ↑		
	BTCV	AMOS	WORD
10	81.89	83.76	83.81
20	82.70	85.39	84.61
30	83.61	87.10	85.24
40	83.97	87.52	85.56
50	83.88	87.29	85.62

Table 21: Ablation study of α for VoCo [68].

α	Dice Score (%) ↑		
	BTCV	AMOS	WORD
0.5	83.91	87.43	85.37
1.0	83.97	87.52	85.56
1.5	83.89	87.29	85.25

Table 22: Ablation study of β for VoCo [68].

β	Dice Score (%) ↑		
	BTCV	AMOS	WORD
0.5	83.94	87.30	85.26
1.0	83.97	87.52	85.56
1.5	83.86	87.24	85.48

100k checkpoints, we resume both methods for an additional 60k steps with their optimizer states restored, using the same 2k-step warmup followed by cosine decay, while keeping all other settings unchanged.

Fig. 11 compares downstream Dice across the four matched checkpoints. `VolumeProbe` consistently outperforms *Random* at every checkpoint in all four settings. The gap narrows on WORD 1%, while remaining largely stable in the other three settings. Notably, `VolumeProbe` @100k still outperforms *Random*@160k in all four settings, indicating that longer *Random* pretraining does not close the gap within the tested range.

The efficiency conclusion remains after accounting for total compute. Including the one-time bank construction and the 4.4% recurring overhead (Table 24), `VolumeProbe` @100k requires 284.9 GPU hours, compared with approximately 323.2 and 430.9 GPU hours for *Random*@120k and *Random*@160k, respectively. It therefore achieves better performance with 11.8% and 33.9% less total compute. Training to $1.6\times$ the original schedule cannot determine the unlimited-training limit. In summary, `VolumeProbe` achieves better performance and compute efficiency within the tested practical training budget, with longer-term convergence left for future investigation.

C.9 Robustness to Non-canonical Fields of View

To examine the spatial assumptions of `VolumeProbe-Lite`, we construct a non-canonical field-of-view (FOV) stress test by shifting each foreground-cropped pretraining volume by 15–20% in-plane or 8–12% through-plane using crop-and-pad, moving anatomy toward the boundary and introducing partial truncation. Using VoCo [68], we compare *Random*, `VolumeProbe-Lite`, and `VolumeProbe-I+D`, which removes *LocPrior* and *Margin* while retaining informativeness and diversity.

Table 23 indicates that `VolumeProbe-Lite` remains better than *Random* under shifted/local FOVs, but its downstream advantage becomes smaller when the geometric assumptions are weakened. Removing the geometry terms improves robustness, with `VolumeProbe-I+D` achieving the best occupancy and Dice under the shifted setting. The geometric prior is effective for foreground-normalized, approximately canonical volumes, but becomes less reliable under non-canonical fields of view. For strongly shifted or truncated scans, attenuating or removing the geometry terms may be preferable.

C.10 Efficiency Analysis

To assess the computational overhead introduced by `VolumeProbe`, we compare random sampling and `VolumeProbe`-based sampling under each SSL backbone in Table 24. All methods use their corresponding official preprocessing and pretraining settings, and all timing experiments are conducted on Intel(R) Data Center GPU Max 1550 GPUs. We report the one-time offline probe-bank

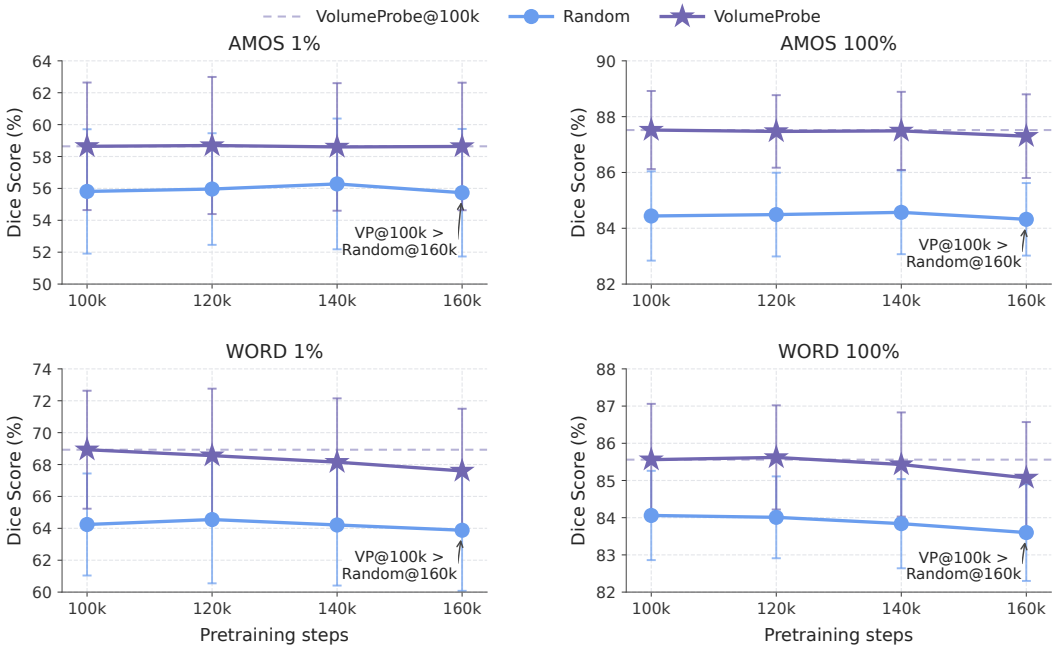


Figure 11: **Convergence under extended pretraining.** Downstream Dice of VoCo with *Random* sampling and `VolumeProbe` from 100k to 160k steps. `VolumeProbe` consistently outperforms *Random*; the gap narrows on WORD 1% but remains stable elsewhere, and `VolumeProbe` @100k still exceeds *Random*@160k in all settings.

Table 23: **Robustness to non-canonical fields of view.** Foreground occupancy, organ occupancy, and AMOS 10% Dice Score (%) are reported.

Pretraining FOV	Sampling Policy	Foreground Occ. (%) ↑	Organ Occ. (%) ↑	AMOS 10% Dice (%) ↑
Canonical	Random	55.83	19.79	73.34
	`VolumeProbe-Lite`	**75.63**	**45.71**	**74.95**
Shifted/local FOV	Random	47.62	15.87	72.08
	`VolumeProbe-Lite`	54.12	25.13	72.40
	`VolumeProbe-I+D`	**60.94**	**31.13**	**73.71**

construction time, probe-bank cache footprint, per-epoch pretraining time, total pretraining time, and the relative training overhead compared with *Random* under the same backbone and training setup. The overhead is computed from total pretraining time, while offline construction time is reported separately. Since VoCo [68] and S2DC [48] employ the same VoCo-10k pretraining dataset, they share the same probe-bank construction pipeline and cached probe banks, leading to identical offline construction time for `VolumeProbe-Lite`. Differences in sub-volume size lead to different `VolumeProbe-Feat` probe-bank construction times for VoCo and S2DC. Additionally, the cache footprint is determined only by the number of pretraining volumes and the bank size, since both `VolumeProbe-Lite` and `VolumeProbe-Feat` ultimately cache the same type of 3D crop-center coordinates, independent of the SSL backbone.

Overall, `VolumeProbe` remains lightweight and practical for large-scale pretraining. The total pretraining overhead is only 4.4% for VoCo [68], 2.1% for S2DC [48], and 2.2% for MedGMAE [13]. The cached probe banks require only 41.5 MB for VoCo-10k [68] and 21.8 MB for AbdomenAtlasMini [37], which is negligible compared with typical 3D medical pretraining datasets. Although feature extraction makes offline construction longer for `VolumeProbe-Feat` than for `VolumeProbe-Lite`, both variants have the same per-epoch and total pretraining time once the bank is cached. The above statistics confirm that `VolumeProbe` preserves the original SSL pretraining pipeline, **adds only a small recurring cost, and introduces minimal storage overhead, making it practical for larger-scale deployments**.

Table 24: **Efficiency comparison between random sampling and `VolumeProbe`.** For each SSL backbone, we report the probe-bank construction time, cache footprint, per-epoch pretraining time, total pretraining time, and relative overhead compared with *Random* under the same training settings.

Method	Offline Construction Time (h)	Cache Footprint (MB)	Per-epoch Time (min)	Total Pretraining Time (GPU h)	Overhead vs. Random
VoCo [68]	–	–	392.6	269.3	–
+ `VolumeProbe-Lite`	3.79	41.5	406.8	281.1	+4.4%
+ `VolumeProbe-Feat`	9.26	41.5	406.8	281.1	+4.4%
S2DC [48]	–	–	1089.9	753.3	–
+ `VolumeProbe-Lite`	3.79	41.5	1137.4	768.9	+2.1%
+ `VolumeProbe-Feat`	9.57	41.5	1137.4	768.9	+2.1%
MedGMAE [13]	–	–	19.89	662.9	–
+ `VolumeProbe-Lite`	1.84	21.8	20.31	677.2	+2.2%
+ `VolumeProbe-Feat`	3.93	21.8	20.31	677.2	+2.2%

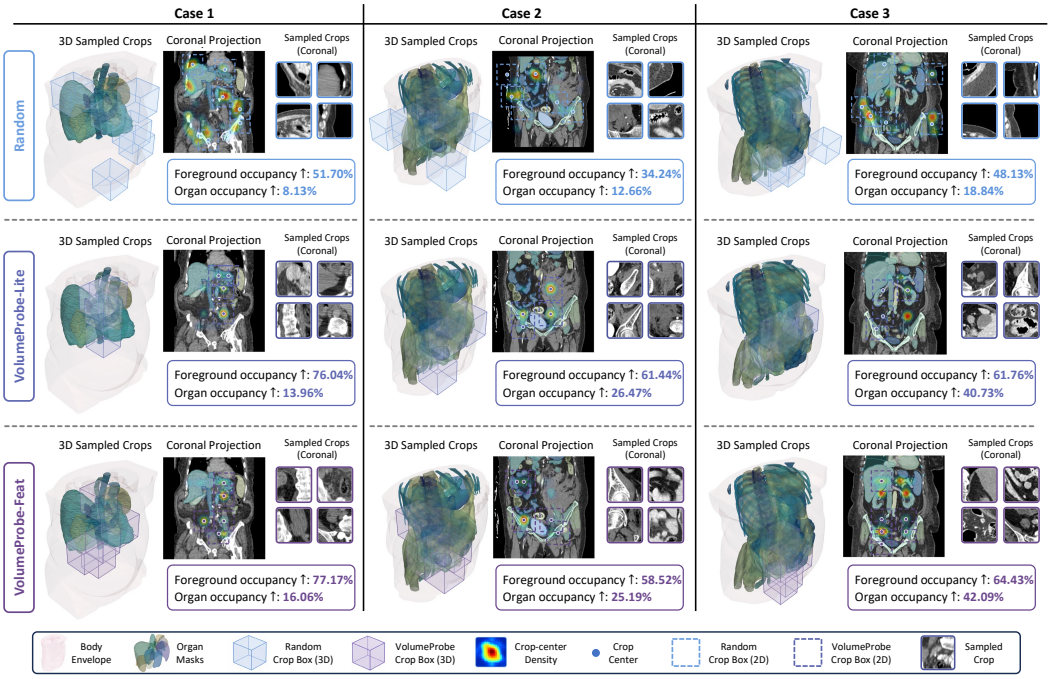


Figure 12: **Additional cases for anatomy-aware sampling.** We compare *Random*, `VolumeProbe-Lite`, and `VolumeProbe-Feat` across three volumes. For each case, we show 3D crop boxes overlaid with organ masks, coronal crop-center density, sampled crop centers, 2D crop boxes, and example cropped views. Both `VolumeProbe` variants sample more anatomically valid and organ-rich regions than *Random*, leading to higher foreground and organ occupancy ratios.

D Additional Analysis of `VolumeProbe`

D.1 Additional Case Visualizations

Fig. 12 provides additional case studies comparing *Random*, `VolumeProbe-Lite`, and `VolumeProbe-Feat`. Across different volumes, *Random* often samples boundary-dominated or weakly anatomical regions, while both `VolumeProbe` variants place crops more consistently inside the body and around organ-rich areas. The improved foreground and organ occupancy ratios further imply that both `Lite` and `Feat` redirect the sub-crop budget toward anatomically valid observations.

D.2 Cross-epoch Trajectory

Fig. 13 compares how different sampling strategies move across epochs under the same sub-volume cropping scheme. We summarize each trajectory using four metrics, i.e., path length measures the

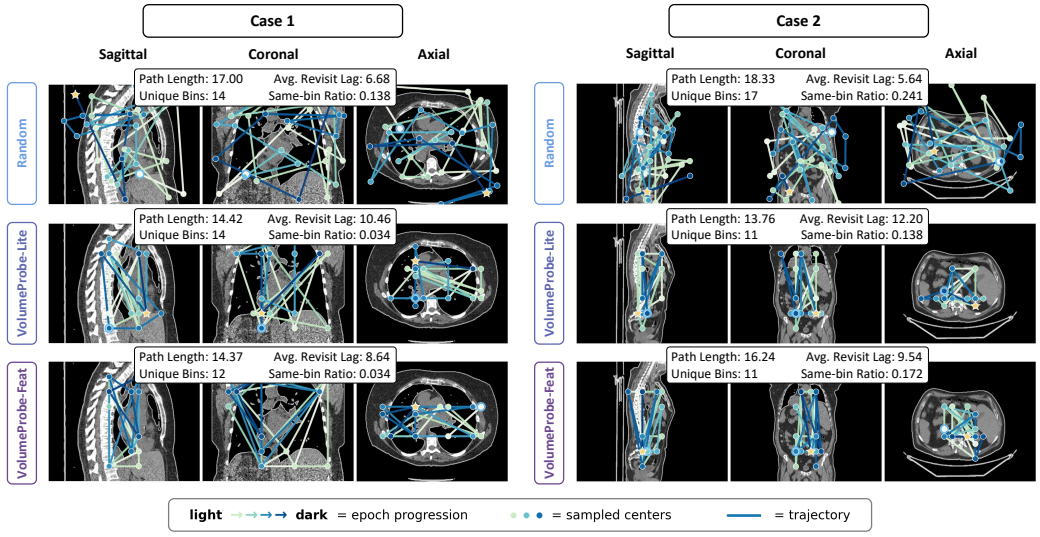


Figure 13: **Cross-epoch trajectory analysis.** We visualize sampled crop centers across epochs for two representative cases under *Random*, `VolumeProbe-Lite`, and `VolumeProbe-Feat` on sagittal, coronal, and axial views. The light-to-dark trajectory denotes epoch progression, and the inset reports path length, unique bins, average revisit lag, and same-bin ratio.

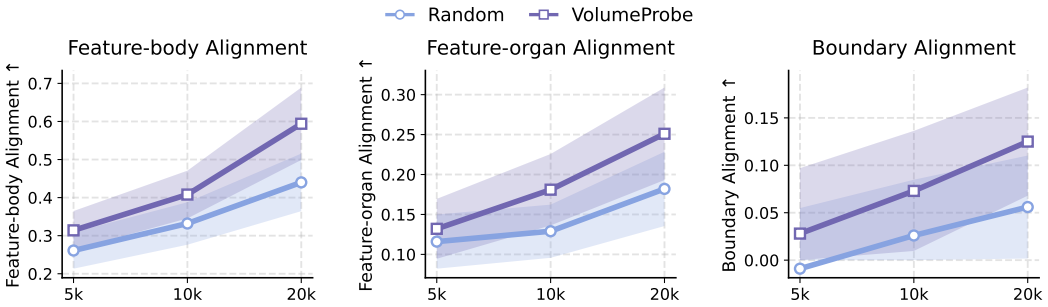


Figure 14: **Quantitative analysis of structure-aligned feature responses.** We evaluate *Random* and `VolumeProbe` at 5k, 10k, and 20k early pretraining iterations using feature responses extracted from the same encoder stage. We report feature-body alignment, feature-organ alignment, and boundary alignment, with shaded regions indicating standard deviation across sampled sub-volumes. `VolumeProbe` consistently achieves higher alignment scores than *Random*, supporting the qualitative feature-map observations in Fig. 8(c).

cumulative displacement between consecutive sampled centers, unique bins count the number of distinct coarse bins visited, average revisit lag measures the epoch interval before returning to a previously visited bin, and same-bin ratio quantifies the fraction of repeated visits within the same coarse bin. *Random* often produces long but weakly structured trajectories with frequent local revisits, whereas `VolumeProbe-Lite` and `VolumeProbe-Feat` keep the trajectory more constrained to anatomically plausible regions. This is reflected by lower same-bin ratios and larger revisit lags, suggesting that the cached probe bank reduces redundant observations while preserving structured cross-epoch exploration. Notably, `VolumeProbe-Feat` does not necessarily maximize path length, but tends to yield semantically more coherent trajectories within meaningful anatomical support.

D.3 Additional Learning Relevance Analysis

Quantitative analysis of structure-aligned responses. To complement the qualitative feature-map visualization in Fig. 8(c), we further quantify whether early feature responses align with anatomical structures. Specifically, we compute three metrics on BTCV [32] using channel-aggregated activation

maps from the same encoder stage. Feature-body alignment and feature-organ alignment measure the fraction of activation mass falling inside the foreground body mask and organ-union mask, respectively, reflecting whether responses avoid background regions and focus on anatomically meaningful structures. Boundary alignment is measured by the Spearman correlation between activation responses and image-gradient magnitude within the body region, indicating whether features follow local structural transitions.

Fig. 14 reports the three alignment metrics across 5k, 10k, and 20k iterations, where `VolumeProbe` consistently achieves higher scores than *Random*. The improvements become more pronounced as pretraining proceeds, suggesting that `VolumeProbe` encourages feature responses to concentrate more on valid body regions, organ-relevant areas, and local structural boundaries. These results provide quantitative support for the structure-aligned responses observed in Fig. 8(c), mitigating the concern that the qualitative examples could be due to case selection.

D.4 Analysis of the Feat Variant

We further analyze why `VolumeProbe-Feat` does not consistently outperform `VolumeProbe-Lite` despite using richer semantic features. We first clarify that MedicalNet [7] is used only for the diagnostic feature-level proxy Q_2 in the motivation study, whereas `VolumeProbe-Feat` uses a frozen 3DINO-ViT [73] encoder for offline probe-bank construction. As reported in the main results (Tables 1 and 2), `VolumeProbe-Feat` consistently improves over *Random* sampling, indicating that the frozen features provide useful semantic information, but it offers no consistent advantage over the lightweight `Lite` variant.

Several factors may contribute to this behavior. The frozen encoder is optimized for general semantic representation rather than local crop-utility estimation, while semantic invariance may suppress fine-grained intensity and structural variations directly captured by gradient and entropy cues. Moreover, the frozen feature space and coarse-bin prototypes are not optimized for crop utility and cannot adapt to volume-specific characteristics or evolving training stages. Consequently, the intrinsic cues in `Lite` appear more directly aligned with informative sub-volume selection under the current setting. This motivates `Lite` as our default variant, while task-adaptive or jointly learned semantic scoring remains an interesting direction for future work.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims presented in our abstract and introduction accurately reflect the contribution of our paper.

   Guidelines:
   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in Sec. 6.

   Guidelines:
   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: This work focuses on devising a plug-and-play, training-free module for sub-volume sampling in 3D medical SSL, which does not contain theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide thorough experimental and implementation details, as well as hyperparameter settings in Appendix B. The pretraining and downstream datasets (details in Appendix A) are all publicly available.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The data and code used in our paper are publicly available.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: All pretraining and downstream fine-tuning details are thoroughly described in Sec. 4.1 and Appendix B, as well as ablations for obtaining hyperparameters in Appendix C.7.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report mean and standard deviation over three independent runs for the downstream segmentation, classification, and registration results in Appendix C.1.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Given the number of experiments conducted in this work, we are unable to give an exhaustive list of compute used for each experiment. However, we provide batch sizes and hyperparameters for all pretraining and fine-tuning experiments in Appendix B, as well as pretraining time in Appendix C.10.

   Guidelines:
   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have read and understood the NeurIPS Code of Ethics and have done our best to conform.

   Guidelines:
   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no potential societal impact of our work.

    Guidelines:
    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g.,

deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our experiments do not involve this aspect.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original paper that produced the code package or dataset.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:
    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve this aspect.

    Guidelines:
    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve this aspect.

    Guidelines:
    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used

only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: This work does not incorporate any LLMs as part of the core methodology. LLMs were used only for writing and editing suggestions in this paper.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.