TRIPOD: TRIADIC 3D CT REPRESENTATION LEARNING VIA PHYSICALLY ANCHORED PRETRAINING AND POLARITY-DEBIASED TEXT ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

General-purpose foundation models for 3D computed tomography (CT) should transfer across global recognition, dense spatial prediction, and language-guided inference. Yet standardized resampling obscures scan-specific acquisition geometry, while report-derived supervision can admit textual shortcuts when concepts are associated with a dominant polarity. We present **TriPoD**, a triadic framework combining physically anchored volumetric pretraining with polarity-debiased text alignment. Pretrained on more than 470,000 quality-controlled CT volumes from multiple medical centers, TriPoD preserves native voxel lattices and voxel-to-world transforms instead of resampling entire scans isotropically. Affine-aware 3D rotary positional encoding maps patch centers into view-centered physical space, representing scale and orientation across acquisitions. For locked-image CT–report alignment, a patient-disjoint text-only audit reveals phrase-conditioned polarity leakage. Polarity-Debiased Opposite Sentence Learning (PD-OSL) addresses this shortcut by combining evidence-constrained atomic findings with polarity-debiased sampling. We evaluate the shared visual encoder across classification, segmentation, and internal and external zero-shot recognition. It transfers strongly across all three regimes. Controlled ablations show small, task-dependent effects of physical coordinates on global recognition and dataset-dependent effects on dense transfer, with a clear gain on one large multi-organ benchmark but negligible change on another. Under matched checkpoint and prompt conditions, PD-OSL reduces phrase-conditioned leakage and improves zero-shot recognition on both cohorts. These results highlight acquisition-aware geometry and shortcut-resistant language supervision as complementary design principles for transferable 3D CT representation learning.

1 INTRODUCTION

Computed tomography (CT) plays a central role in disease diagnosis, treatment planning, and longitudinal assessment. Yet most CT-based AI systems remain developed for individual tasks with task-specific supervision and adaptation (Moor et al., 2023). Recent CT foundation models have begun to reduce this fragmentation. Vision-centric approaches learn transferable volumetric representations from large-scale CT cohorts and adapt them to classification and segmentation (Pai et al., 2025; Veenboer et al., 2026; Zhu et al., 2026), whereas multimodal approaches align CT volumes with radiology reports or structured clinical semantics to support zero-shot recognition and report-related applications (Hamamci et al., 2026; Blankemeier et al., 2026; Wald et al., 2026). Despite this progress, three limitations remain. First, existing CT foundation models are typically pretrained either on anatomically restricted datasets, including head (Zhu et al., 2026), chest (Hamamci et al., 2026), and abdominal CTs (Blankemeier et al., 2026), or on multi-region but oncology-centered collections (Pai et al., 2025; Veenboer et al., 2026). Such corpora do not fully capture the joint diversity of anatomical regions, contrast phases, diseases, and acquisition protocols encountered in routine clinical practice. Second, common preprocessing pipelines resample scans to fixed voxel spacing (Pai et al., 2025; Blankemeier et al., 2026), preventing models from explicitly learning across native acquisition geometries. Third, report-based supervision can admit lexical shortcuts that bypass visual content, particularly when affirmative and negated descriptions differ system-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

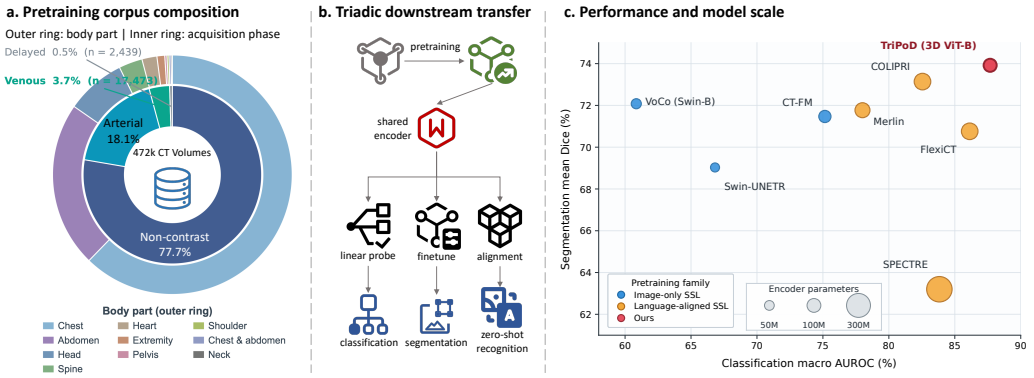

Figure 1: **Overview and empirical positioning of TriPoD. (a)** Distribution of 472,084 pretraining volumes by anatomical region (outer ring) and acquisition phase (inner ring). **(b)** A shared 3D ViT-B supports classification, segmentation, and polarity-debiased alignment for zero-shot recognition. **(c)** Transfer performance of representative pretrained CT foundation models: mean macro AUROC over eight shared classification benchmarks versus mean Dice over eleven segmentation benchmarks. Bubble area denotes encoder size; color denotes the pretraining family.

atically. We introduce TriPoD, which addresses these limitations through large-scale clinical CT curation, physically anchored volumetric pretraining, and polarity-debiased text alignment. The resulting visual backbone supports global recognition, dense spatial prediction, and language-aligned zero-shot recognition.

General-purpose CT pretraining requires broad clinical coverage. Real-world scans provide diversity in anatomy, contrast phase, disease, and acquisition protocol, but also vary substantially in voxel spacing, slice thickness, and orientation. Existing foundation models commonly manage this heterogeneity by resampling scans to a fixed voxel spacing (Pai et al., 2025; Zhu et al., 2026; Blankemeier et al., 2026; Claessens et al., 2026; Li et al., 2026), which removes acquisition-specific spacing from the model input and introduces interpolation before representation learning. We instead curate 472,084 quality-controlled CT volumes spanning diverse anatomical regions and contrast phases, canonicalizing anatomical axis order while retaining each volume's native voxel lattice and voxel-to-world geometry.

Retaining native geometry is useful only if the model can interpret it. Standard positional encodings operate on discrete patch indices, even though the same token offset can correspond to different physical displacements across scans. We therefore introduce physically anchored self-supervised pretraining with affine-aware 3D rotary positional encoding (RoPE) (Su et al., 2024). Rather than treating patch indices as spatial coordinates, TriPoD maps patch centers into view-centered physical space through the tracked voxel-to-world affine. Its complete 3×3 linear component preserves physical scale and orientation, allowing token relationships to reflect physical displacement while retaining fixed-size training views.

The next challenge is to introduce clinical semantics without compromising the transferability of the visual representation. Existing multimodal CT models align volumetric features with radiology reports to support language-guided recognition (Hamamci et al., 2026; Claessens et al., 2026; Wald et al., 2026). We adopt post-hoc alignment with a frozen visual encoder and finding-level supervision from atomic report statements. However, opposite-sentence learning (OSL) (Wald et al., 2026) can associate individual clinical concepts with a dominant textual polarity, allowing the training label to be inferred without visual evidence. A patient-disjoint text-only audit exposes this concept-conditioned shortcut. We consequently propose Polarity-Debiased Opposite Sentence Learning (PD-OSL), combining evidence-constrained atomic findings, explicit-negative dilution, and weak counterfactual supplementation to reduce phrase-conditioned polarity leakage during locked-image alignment.

We evaluate TriPoD across eight classification benchmarks, eleven anatomical and lesion segmentation benchmarks, and zero-shot recognition on internal and external thoracic CT cohorts. These

experiments test global, dense, and language-aligned transfer, while controlled analyses examine the effects of physical-coordinate encoding and polarity-debiased supervision. Our principal contributions are threefold:

(1) We introduce physically anchored volumetric pretraining at clinical scale, using $472{,}084$ quality-controlled CT volumes while preserving native acquisition geometry and encoding physical patch coordinates with affine-aware 3D RoPE.

(2) We identify concept-conditioned polarity leakage in opposite-sentence learning and propose PD-OSL, which combines evidence-constrained atomic supervision with polarity-debiased sampling to promote visually grounded finding-level alignment.

(3) We establish a triadic evaluation spanning global classification, dense segmentation, and language-aligned zero-shot recognition, demonstrating strong transfer from a single reusable 3D CT encoder across all three regimes.

2 Related Work

Self-supervised pretraining and foundation models for volumetric CT. Self-supervised learning enables volumetric models to exploit large unlabeled cohorts, progressing from handcrafted pretext tasks to masked image modeling and self-distillation. General-purpose objectives such as MAE (He et al., 2022), iBOT (Zhou et al., 2022), and the DINO family (Caron et al., 2021; Oquab et al., 2024; Siméoni et al., 2026), together with early volumetric methods (Zhou et al., 2019; Tang et al., 2022) and medical-specific designs (Xie et al., 2024; Zhou et al., 2023; Jiang et al., 2023), have established broad transfer across anatomies and tasks. Recent CT foundation models further scale pretraining: CT-FM and VoCo use large-scale contrastive or geometry-aware objectives (Pai et al., 2025; Wu et al., 2025b); FM-HCT specializes in head CT (Zhu et al., 2026); TAP-CT and CoralBay emphasize frozen task-agnostic evaluation (Veenboer et al., 2026; Gatopoulos et al., 2026); and supervised or segmentation-oriented models provide strong dense transfer (Li et al., 2024; He et al., 2025; Hu et al., 2026). Most existing systems, however, are developed primarily for global recognition, dense prediction, or language-aligned transfer, leaving the cross-regime generality of a single volumetric representation insufficiently characterized.

Spatial representation in 3D CT transformers. Voxel indices lack consistent physical meaning: variation in spacing, slice thickness, and orientation can make identical voxel displacements correspond to different anatomical distances. The common remedy is to canonicalize orientation and resample scans to a fixed resolution, such as 1-mm isotropic in FM-HCT (Zhu et al., 2026), $0.5 \times 0.5 \times 1.0$ mm in SPECTRE (Claessens et al., 2026), and 1.5-mm isotropic in FlexiCT (Li et al., 2026). Three-dimensional rotary embeddings are then applied on the resulting standardized grids (Claessens et al., 2026; Li et al., 2026), allowing physical geometry to enter only indirectly through preprocessing. TAP-CT retains native spacing but continues to index positions on the image grid (Veenboer et al., 2026). Consequently, how to encode token positions directly in physical space while preserving native acquisition geometry remains underexplored.

Vision-language alignment and task-agnostic transfer. Text supervision provides a complementary route to clinically meaningful representations by adapting objectives such as CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023) to volumetric CT. CT-CLIP, COLIPRI, and Merlin align CT volumes with reports or structured labels for zero-shot recognition, retrieval, generation, and segmentation (Hamamci et al., 2026; Wald et al., 2026; Blankemeier et al., 2026), while generalist radiology models extend report supervision across 2D and 3D modalities (Wu et al., 2025a). Finer-grained approaches additionally match local regions or multi-scale visual structures with report descriptions (Shui et al., 2025; Ni et al., 2026). Two limitations remain. First, systems that jointly optimize visual self-supervision and language alignment, including SPECTRE and FlexiCT (Claessens et al., 2026; Li et al., 2026), make it difficult to separate image-derived representation quality from the effects of report supervision. Locked-image alignment offers a cleaner separation between these stages (Jose et al., 2025). Second, opposite-sentence objectives such as that used by COLIPRI (Wald et al., 2026) can expose concept-conditioned polarity regularities, allowing supervision labels to be inferred from textual form rather than visual evidence. The polarity structure of finding-level alignment supervision has received comparatively little attention.

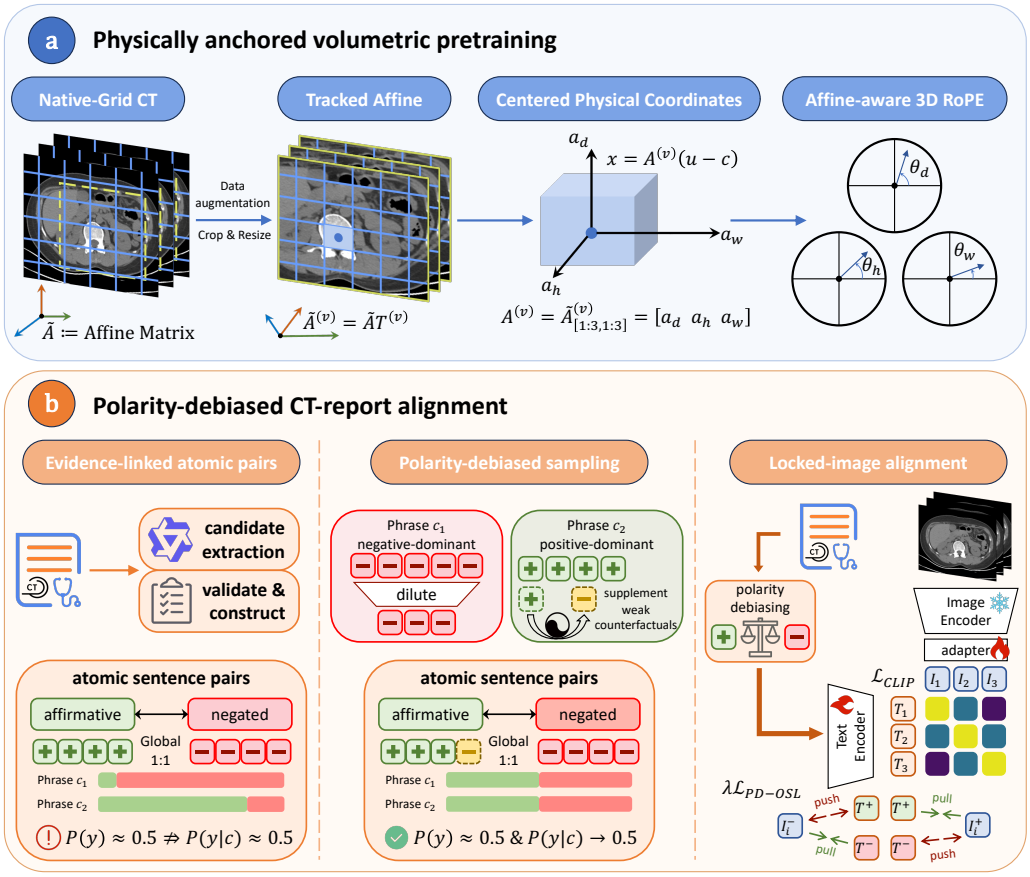

Figure 2: Overview of TriPoD. **(a)** Physically anchored volumetric pretraining preserves native acquisition geometry by propagating the voxel-to-world affine through view construction, mapping patch centers to view-centered physical coordinates, and using them in affine-aware 3D RoPE. **(b)** PD-OSL combines evidence-constrained atomic supervision with polarity-debiased sampling to reduce concept-conditioned polarity leakage during locked-image CT-report alignment.

3 Methodology

3.1 Overview

TriPoD consists of two stages: physically anchored volumetric pretraining and subsequent polarity-debiased CT–report alignment (Fig. 2). The first stage encodes scan-specific acquisition geometry with affine-aware 3D RoPE, while the second aligns the frozen visual representation with atomic clinical findings using PD-OSL.

3.2 Clinical CT Corpus Construction

We assemble routine clinical CT archives from nine medical centers spanning several years. A rule-based pipeline selects diagnostically valid volumetric series, assigns anatomical-region and acquisition-phase categories for corpus statistics, and removes localizers, derived images, incomplete reconstructions, duplicate series, and volumes with invalid spatial metadata. After quality control, 472,084 volumes remain, covering diverse anatomies, contrast phases, diseases, and acquisition protocols. Each series is decoded as a complete 3D volume and canonicalized to a common anatomical axis convention without spacing normalization; its native voxel lattice and voxel-to-world affine are preserved for physically anchored pretraining.

3.3 PHYSICALLY ANCHORED VOLUMETRIC PRETRAINING

Native-Grid View Construction. We pretrain a 3D ViT-B (Dosovitskiy et al., 2021) with a DINOv3-style teacher–student framework (Siméoni et al., 2026). Starting from each curated CT volume and its voxel-to-world affine, we sample stochastic 3D regions and regrid them to fixed tensor sizes for multi-crop self-supervision. Every crop, resize, and optional flip is accompanied by the corresponding affine update, which continues to map coordinates in the transformed view to physical space. Fixed-size inputs therefore share a tensor shape without being treated as if they shared a physical sampling grid. Here, "native-grid" describes the complete-volume representation before stochastic view construction; sampled views may still undergo interpolation. Intensity augmentation, CT windowing, and normalization leave the tracked spatial mapping unchanged.

Affine-Aware 3D Rotary Positional Encoding. Let v index a training view constructed from volume i, and let

$$A_i^{(v)} = \begin{bmatrix} \mathbf{a}_d^{(v)} & \mathbf{a}_h^{(v)} & \mathbf{a}_w^{(v)} \end{bmatrix} \in \mathbb{R}^{3\times 3} \tag{1}$$

denote the complete linear component of its tracked affine, where $d/h/w$ refers to physical axis $z/y/x$ respectively. Each column $\mathbf{a}_a^{(v)} \in \mathbb{R}^3$ gives the physical displacement associated with a unit step along one view axis. Unlike the diagonal spacing matrix $\mathrm{diag}(s_d, s_h, s_w)$, the complete linear component also retains the orientation of the view through its off-diagonal entries. For patch center $\mathbf{u}_p$ and geometric view center $\mathbf{c}_v$, we compute the view-centered physical displacement

$$\mathbf{x}_{ip}^{(v)} = A_i^{(v)} \left(\mathbf{u}_p - \mathbf{c}_v\right). \tag{2}$$

The translation component of the homogeneous affine is intentionally omitted. The resulting coordinate therefore represents relative physical displacement from the view center rather than absolute scanner-space location. For physical axis a and rotary frequency index f, the corresponding phase is

$$\theta_{ip,a,f}^{(v)} = 2\pi \frac{x_{ip,a}^{(v)}}{T_{a,f}}, \tag{3}$$

where $T_{a,f}$ is drawn from a fixed axis-specific period bank shared across studies. The resulting rotations are applied to patch-token queries and keys in every Transformer block. Study-specific spacing and orientation therefore enter through $\mathbf{x}_{ip}^{(v)}$, while the frequency basis remains shared across the pretraining corpus. This formulation injects scan-specific physical scale and orientation into every self-attention block without requiring whole-volume isotropic resampling.

Visual Pretraining Objectives. Visual pretraining proceeds in two stages. Stage 1 combines global and local DINO distillation, masked-patch iBOT (Zhou et al., 2022), and KoLeo regularization (Sablayrolles et al., 2019). Stage 2 initializes from the Stage 1 representation and adds high-resolution Gram-based anchoring from a frozen teacher to refine dense patch relations. Affine-aware 3D RoPE is retained throughout both stages, and the refined Stage 2 encoder is used for downstream transfer and locked-image text alignment. Architecture, objective weights, crop configurations, and optimization schedules are reported in Appendix A.2.

3.4 POLARITY-DEBIASED CT-REPORT ALIGNMENT

Locked-image alignment and phrase–polarity leakage. We adopt locked-image text tuning (Zhai et al., 2022), freezing the Stage-2 3D visual encoder while optimizing a lightweight visual adaptation head, CXR-BERT (Nicolson et al., 2024), modality-specific projections, and a shared logit scale. Report-level text is trained with the symmetric CLIP objective and cross-rank negatives (Radford et al., 2021) to provide global semantic alignment, whereas short sentence pairs provide polarity-sensitive supervision in the same embedding space.

Following OSL (Wald et al., 2026), each affirmative phrase t^+ is paired with its negated counterpart t^-, and the target indicates which member is supported. Global balance between positive and negative targets does not ensure balance within each clinical phrase. A phrase that predominantly occurs with one polarity can therefore reveal the target without visual evidence, a failure mode we call *phrase–polarity leakage*.

Quantifying supervision leakage. We audit this shortcut with a patient-disjoint text-only probe. A token-presence naive Bayes classifier receives only the normalized affirmative phrase and predicts its sampled polarity. Its AUROC, denoted *LeakageAUC*, measures phrase-conditioned label predictability: values near 0.5 indicate little recoverable polarity information, whereas high values reveal a lexical shortcut. This diagnostic evaluates the supervision distribution rather than disease classification and is not used as a training objective.

Evidence-constrained phrase construction. PD-OSL operates on atomic findings anchored to verbatim report spans. We retain only definite, current, and visually assessable statements, never interpret report omissions as explicit negatives, and generate anatomy-preserving negations with deterministic rules. Section parsing, exclusion and quarantine criteria, anatomical-modifier handling, and section-level status assignment are detailed in Appendix A.5.

Phrase polarity-debiased sampling. Given the evidence-constrained phrase inventory, PD-OSL preserves all report-derived positives and reshapes only the negative distribution. We first address negative-dominant phrases arising from routine reporting. Let $n_-(u)$ denote the number of patients for whom phrase u occurs as an explicit negative, and let $\widetilde{n}_+(u)$ denote its positive-patient frequency in the section-specific lookup. Each explicit-negative instance is retained with probability:

$$p_{\text{keep}}(u) = \min\left(1, \frac{\widetilde{n}_+(u)+1}{n_-(u)+1}\right). \tag{4}$$

This rule increasingly suppresses explicit negatives as their patient frequency exceeds the corresponding positive frequency, while leaving all positive findings unchanged.

Negative dilution alone, however, cannot provide negative supervision for positive-dominant phrases. We therefore construct section-specific candidate banks from definite positive findings. Each entry $a = (s, u)$ consists of a clinical section s and a normalized phrase u, and is assigned a weight $n_+(a)$ equal to its number of positive donor patients. For CT i, the admissible set $\mathcal{A}_i$ contains only entries for which the corresponding section is not `unknown`, the phrase is not positive in the current record, the donor set excludes the current patient, and the phrase has not already been selected. We then sample an admissible entry according to

$$q(a \mid i) = \frac{n_+(a)}{\sum_{a' \in \mathcal{A}_i} n_+(a')}, \qquad a \in \mathcal{A}_i. \tag{5}$$

Weighting candidates by their positive-patient support preferentially exposes positive-dominant phrases to the opposite polarity, thereby reducing their phrase-level predictability. Because coverage of a report section does not establish the absence of every unmentioned finding, these samples are treated as *weak counterfactual negatives*, rather than verified negatives. Phrase matching remains lexical, without merging medical synonyms.

Each CT contains at most eight phrase slots: up to four report-derived positives and four negatives. Retained explicit negatives are assigned first, weak counterfactual negatives fill the remaining negative slots, and unfilled slots are ignored.

Joint alignment objective. For slot k of CT i, let $\mathbf{v}_i$ denote the normalized CT embedding and let $\mathbf{t}_{ik}^+$ and $\mathbf{t}_{ik}^-$ denote the normalized embeddings of the affirmative and negated phrases, respectively. A positive slot should favor t_{ik}^+ over t_{ik}^-, whereas a negative slot should favor the reverse. Reusing the learned CLIP scale α, we represent this relative preference with the polarity logit:

$$z_{ik} = \alpha\left[\cos\big(\mathbf{v}_i, \mathbf{t}_{ik}^+\big) - \cos\big(\mathbf{v}_i, \mathbf{t}_{ik}^-\big)\right]. \tag{6}$$

Thus, a positive value of z_{ik} indicates greater compatibility with the affirmative phrase, while a negative value favors its negated counterpart. Let $\mathcal{V}$ denote the set of non-padding slots across all distributed ranks. The polarity-sensitive objective is then

$$\mathcal{L}_{\text{PD-OSL}} = \frac{1}{|\mathcal{V}|} \sum_{(i,k)\in\mathcal{V}} \text{BCEWithLogits}(z_{ik}, y_{ik}). \tag{7}$$

Finally, we combine this finding-level objective with report-level contrastive alignment:

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{CLIP}} + \lambda(r)\mathcal{L}_{\text{PD-OSL}}, \tag{8}$$

where r denotes the optimization step and $\lambda(r)$ is a scheduled loss weight. The visual backbone remains frozen throughout alignment. The detailed configurations are in Appendix A.4.

4 EXPERIMENTS

We evaluate TriPoD from three complementary perspectives: broad visual transfer across classification and segmentation, the controlled effect of physical-coordinate positional encoding, and the ability of polarity-debiased supervision to reduce text-derived shortcuts during zero-shot alignment.

4.1 EXPERIMENTAL SETUP

Main visual-transfer experiments use the refined Stage-2 3D ViT-B encoder from Sec. 3.3. We evaluate eight public classification benchmarks under both frozen linear probing and model-specific adaptation. On eleven segmentation benchmarks, every visual encoder is fully unfrozen and optimized end-to-end with its segmentation decoder. Decoder architecture, input preprocessing, and inference procedures follow the corresponding original papers and official implementations. The segmentation table therefore compares fully fine-tuned systems rather than frozen encoders. Full segmentation benchmark names are provided in Appendix F. CT–report alignment locks the Stage-2 visual encoder and trains only the alignment modules. Zero-shot recognition is evaluated on 1,564 CT-RATE scans from 1,304 patients with 18 labels and on 3,630 RAD-ChestCT volumes with 16 mapped labels, using macro AUPRC and AUROC.

Controlled positional-encoding analyses compare two ViT-S encoders trained for 30k steps on the same fixed 10% corpus subset, differing only in the coordinates supplied to 3D RoPE. Baselines cover representative supervised, self-supervised, and vision–language CT models. Dataset definitions, splits, label harmonization, evaluation protocols, and available result provenance are provided in Appendix B; classification adaptation details are reported in Appendix C.1, segmentation protocols in Appendix B.3, and complete training configurations in Appendix A.

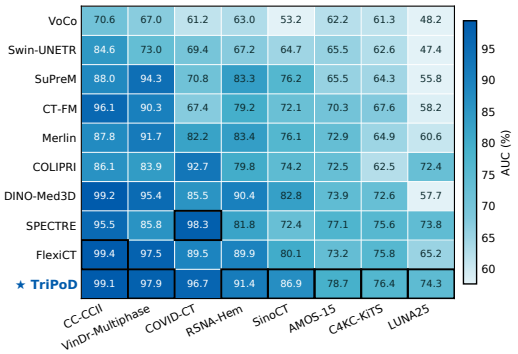

Figure 3: Frozen linear-probe AUROC (%) across 10 CT foundation models and 8 classification datasets. The per-dataset best is outlined, and ⋆ marks TriPoD with the highest mean AUROC.

4.2 GLOBAL SEMANTIC TRANSFER ACROSS CT CLASSIFICATION TASKS

Under the unified frozen linear-probing protocol, TriPoD achieves the highest mean AUROC of 87.68%, exceeding the strongest baseline, FlexiCT (83.83%), by 3.85 percentage points and ranking first on six of eight datasets (Fig. 3). This controlled comparison supports broad transfer across anatomical regions and clinical objectives without updating the backbone. Full AUC, accuracy, and F1 results under model-specific adaptation are provided in the Appendix C.2; they estimate attainable downstream performance but are not used for encoder-only attribution.

4.3 DENSE TRANSFER ACROSS ANATOMICAL AND LESION SEGMENTATION

Across eleven segmentation benchmarks, TriPoD obtains the highest overall mean Dice of 73.92, compared with 73.43 for VISTA3D and 73.14 for COLIPRI (Table 1). Its strongest results occur on broad anatomical tasks, including TotalSegmentator v2, AbdomenAtlas1.0Mini, and SegTHOR. Performance is more heterogeneous on the five smaller MSD lesion tasks, where TriPoD does not achieve the strongest aggregate result. We therefore interpret this table as evidence of broad anatomical transfer rather than universal superiority. Every backbone is fully fine-tuned; decoder architecture and baseline-specific inference procedures follow the corresponding original papers and official implementations. The table consequently compares complete fine-tuned systems, with protocol and result provenance summarized in Appendix B.3.

Table 1: Dense segmentation Dice (%) across eleven benchmarks. All encoders are fully fine-tuned; the first five columns are MSD tasks. Best and second-best results are shown in **bold** and underlined.

Method	Liver	Lung	Panc.	Vessel	Colon	AISD	WHS	SegThor	INST.	Atlas	TotalSeg	Avg
Swin-UNETR	64.72	71.64	48.31	60.72	26.19	56.48	89.06	87.23	77.62	90.85	86.47	69.03
SuPreM	68.20	72.55	50.02	64.71	38.78	57.54	**90.88**	89.70	81.00	89.16	86.95	71.77
VISTA3D	**71.90**	73.50	**63.13**	62.50	44.72	52.85	89.84	90.20	78.80	90.98	89.30	73.43
VoCo (nnUNet)	70.09	68.99	50.43	**70.61**	41.93	**66.36**	89.53	88.82	70.77	90.66	89.48	72.52
VoCo (Swin-B)	69.64	70.11	60.87	61.89	30.97	62.95	87.33	89.16	81.17	91.00	87.74	72.08
Merlin	67.14	69.27	62.55	66.66	29.32	64.69	89.19	90.65	**84.57**	89.03	76.66	71.77
CT-FM	64.48	72.52	54.22	61.99	41.21	56.68	86.55	88.12	81.14	89.42	89.81	71.47
DINO-Med3D	67.96	62.02	53.85	66.18	44.84	63.46	86.20	89.72	66.17	90.67	85.86	70.63
COLIPRI	67.34	72.97	62.57	62.99	**49.14**	54.78	88.17	89.23	81.50	90.45	85.37	73.14
SPECTRE	52.91	61.66	32.08	55.20	31.06	50.31	85.10	86.60	75.92	88.18	76.18	63.20
FlexiCT	61.35	**73.88**	56.50	61.64	40.80	55.06	88.98	87.88	79.15	90.51	82.58	70.76
Ours (TriPoD)	68.14	72.17	58.67	66.81	42.89	59.77	90.33	**90.76**	81.74	**91.35**	**90.50**	**73.92**

Table 2: Controlled ViT-S ablation on two large multi-organ segmentation cohorts. Dice and NSD are per-case macro percentages; intervals are case-level paired 95% confidence intervals.

Dataset	Metric	Normal PE	Physical PE	Δ	95% CI
TotalSegmentator v2 ($n = 248$)	Macro Dice ↑	84.67	86.57	$+1.90$	$[+1.16, +2.73]$
	NSD@1 mm ↑	69.58	71.63	$+2.04$	$[+1.45, +2.71]$
AbdomenAtlas1.0Mini ($n = 1{,}038$)	Macro Dice ↑	90.827	90.807	-0.020	$[-0.087, +0.047]$
	NSD@1 mm ↑	80.797	80.836	$+0.039$	$[-0.042, +0.121]$

4.4 Dataset-Dependent Effects of Physical Coordinates

The final ViT-B encoder combines native-grid inputs with physical-coordinate 3D RoPE. To isolate coordinate parameterization within the available compute budget, we compare two ViT-S encoders trained for 30k steps on the same fixed 10% corpus subset. Normal PE uses normalized voxel-grid coordinates, whereas Physical PE uses centered physical coordinates derived from the transformed affine. This proxy-scale experiment does not compare native-grid training with isotropic resampling and does not by itself attribute the final ViT-B performance to positional encoding.

Physical PE produces small, directionally mixed changes on global recognition. Under frozen linear probing, AUROC changes by -0.66 points on SinoCT and $+0.39$ on VinDr-Multiphase, while RSNA-Hem accuracy changes by -0.50. With lightweight adaptation, the corresponding changes are $+1.15$ AUROC, $+0.04$ AUROC, and $+0.71$ accuracy. These results indicate that physical coordinates preserve global semantic transfer but do not uniformly improve linear separability.

The dense-transfer effect is therefore dataset dependent. Physical PE improves both region overlap and millimeter-scale boundary accuracy on TotalSegmentator v2, whereas all observed changes on AbdomenAtlas1.0Mini are near zero. TotalSegmentator v2 is evaluated on volumes already standardized to 1.5-mm isotropic spacing, leaving no identifiable case-level spacing variation; we consequently make no claim about spacing robustness or an interaction between anisotropy and PE benefit.

4.5 Polarity Debiasing Mitigates Language Shortcuts

Text-supervision quality audit. We first test whether OSL polarity labels can be inferred from text alone. Although the evaluated sampling strategies have similar marginal positive fractions (0.414–0.456), their phrase-conditional label predictability differs substantially (Table 3). Explicit-negative and normal-section sampling remain almost perfectly predictable from normalized affirmative phrases, while paper-style bank sampling only partially reduces the shortcut. PD-OSL lowers LeakageAUC to 0.573 and balanced accuracy to 0.558, both close to chance. This reduction cannot be explained by global class rebalancing and instead reflects a more balanced phrase-conditional polarity distribution.

Table 3: Patient-disjoint text-only audit of OSL supervision. Lower LeakageAUC indicates less phrase-conditioned label predictability.

Sampling strategy	Pos. frac.	LeakageAUC	Bal. acc.
Explicit negatives	0.456	0.998	0.985
Normal-section bank	0.445	0.982	0.953
Paper-style bank	0.414	0.787	0.732
PD-OSL	0.424	**0.573**	**0.558**

Table 4: Zero-shot recognition on CT-RATE (18 labels) and RAD-ChestCT (16 mapped labels), reported as macro percentages. TriPoD variants share the checkpoint, dataset-specific prompt, and metrics; published baselines retain their original prompts.

Model	Loss	CT-RATE		RAD-ChestCT	
		AUPRC	AUROC	AUPRC	AUROC
CT-CLIP	$\mathcal{L}_{\text{CLIP}}$	36.24	73.10	35.81	62.90
COLIPRI-C	$\mathcal{L}_{\text{CLIP}} + \mathcal{L}_{\text{OSL}}$	42.04	75.02	41.05	68.71
COLIPRI-CM	$\mathcal{L}_{\text{CLIP}} + \mathcal{L}_{\text{OSL}} + \mathcal{L}_{\text{MAE}}$	48.62	79.78	44.45	72.20
COLIPRI-CRM	$\mathcal{L}_{\text{CLIP}} + \mathcal{L}_{\text{OSL}} + \mathcal{L}_{\text{RRG,MAE}}$	47.53	78.53	45.15	72.98
TriPoD	$\mathcal{L}_{\text{CLIP}}$	45.36	77.64	45.61	72.29
TriPoD	$\mathcal{L}_{\text{CLIP}} + \mathcal{L}_{\text{OSL}}$	50.25	80.92	46.28	73.21
TriPoD	$\mathcal{L}_{\text{CLIP}} + \mathcal{L}_{\text{PD-OSL}}$	**52.18**	**81.48**	**48.37**	**74.64**

Zero-shot recognition. We evaluate locked-image alignment on CT-RATE and RAD-ChestCT. Within each dataset, all TriPoD variants share the visual checkpoint, dataset-specific prompt, and metric implementation; published baselines retain their original prompts.

Under the matched prompt, PD-OSL improves over standard OSL on all four dataset–metric combinations: by 1.93 AUPRC and 0.56 AUROC points on CT-RATE, and by 2.09 AUPRC and 1.43 AUROC points on RAD-ChestCT. At the finding level, PD-OSL improves CT-RATE AUROC over CLIP-only alignment for 15 of 18 abnormalities and over paper-style OSL for 12 of 18; full category-wise results are reported in Table 12. Together with the text-only audit, the matched-prompt results support the narrower conclusion that polarity-debiased supervision reduces a measurable textual shortcut and is associated with stronger zero-shot recognition.

5 Conclusion

We introduced TriPoD, a unified framework combining physically anchored volumetric pretraining with polarity-debiased text alignment. Pretraining retains native acquisition geometry and encodes physical patch relationships through affine-aware 3D RoPE. The shared visual encoder transfers across eight classification and eleven segmentation benchmarks. Controlled ViT-S analyses show task-dependent global effects and dataset-dependent dense effects, with clear gains on TotalSegmentator but negligible changes on AbdomenAtlas. A text-only audit further exposes phrase-conditioned polarity leakage in conventional OSL. PD-OSL reduces label predictability from text alone and improves matched-checkpoint, matched-prompt zero-shot recognition on CT-RATE and the external RAD-ChestCT cohort. These results position acquisition geometry and supervision construction as complementary design principles for transferable 3D CT representations.

Limitations. The positional-encoding study uses ViT-S on 10% of the corpus and isolates coordinate parameterization rather than native-grid training versus isotropic resampling; larger factorial studies remain needed. Segmentation fully fine-tunes all backbones but retains architecture-specific decoders and inference pipelines. Alignment comparisons lack repeated runs and confidence intervals, while published baselines differ in prompts and label taxonomies. Language evaluation is limited to thoracic findings.

AI USE STATEMENT

Generative AI tools were used to assist with cleaning and extracting textual data, providing feedback on experimental design, and drafting and editing parts of the manuscript. All AI-assisted data-processing outputs, experimental decisions, analyses, and manuscript text were reviewed and verified by the authors. The authors take full responsibility for the final content of this work, including all claims, results, and artifacts produced with the assistance of generative AI.

ETHICS AND PRIVACY STATEMENT

The retrospective use of clinical CT volumes and radiology reports was approved by the institutional review boards of the participating centers, with informed consent waived where applicable. All data were de-identified before research use and processed within approved secure computing environments. Patient-level information was used only for constructing disjoint data partitions and was not provided to the evaluated models. The clinical pretraining corpus cannot be publicly released because of institutional data-governance and privacy restrictions.

REPRODUCIBILITY STATEMENT

Architectural, optimization, and data-construction details are provided in Appendix A. Dataset definitions and evaluation protocols are described in Appendix B. Complete classification results, controlled positional-encoding analyses, and category-level alignment results are reported in Appendices C–E. Code, configuration files, and evaluation prompts will be released upon publication; clinical data remain subject to institutional governance and cannot be redistributed.

REFERENCES

Louis Blankemeier, Ashwin Kumar, Joseph Paul Cohen, Jiaming Liu, Longchao Liu, Dave Van Veen, Syed Jamal Safdar Gardezi, Hongkun Yu, Magdalini Paschali, Zhihong Chen, et al. Merlin: a computed tomography vision–language foundation model and dataset. *Nature*, 652 (8112):1318–1328, 2026. 1, 2, 3

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pp. 9650–9660, 2021. 3

Cris Claessens, Christiaan Viviers, Giacomo D'Amicantonio, Egor Bondarev, and Fons van der Sommen. Scaling self-supervised and cross-modal pretraining for volumetric ct transformers. In *CVPR*, pp. 13636–13647, 2026. 2, 3

Binh T Dao, Thang V Nguyen, Hieu H Pham, and Ha Q Nguyen. Phase recognition in contrast-enhanced ct scans based on deep learning and random sampling. *Medical Physics*, 49(7):4518–4528, 2022. 16

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 5

Adam E Flanders et al. Construction of a machine learning dataset through collaboration: the rsna 2019 brain ct hemorrhage challenge. *Radiology: Artificial Intelligence*, 2(3):e190211, 2020. 16

Ioannis Gatopoulos, Nicolas Känzig, Sebastian Otálora, and Fei Tang. Coralbay: A self-supervised ct foundation model. *arXiv preprint arXiv:2606.03888*, 2026. 3

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, et al. Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature Biomedical Engineering*, pp. 1–19, 2026. 1, 2, 3, 17

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pp. 16000–16009, 2022. 3

Yufan He, Pengfei Guo, Yucheng Tang, Andriy Myronenko, Vishwesh Nath, Ziyue Xu, Dong Yang, Can Zhao, Benjamin Simon, Mason Belue, et al. Vista3d: A unified segmentation foundation model for 3d medical imaging. In *CVPR*, pp. 20863–20873, 2025. 3

Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical image analysis*, 67:101821, 2021. 17

Sarah M. Hooper, Jared A. Dunnmon, Matthew P. Lungren, Domenico Mastrodicasa, Daniel L. Rubin, Christopher Ré, Adam Wang, and Bhavik N. Patel. Impact of upstream medical image processing on downstream performance of a head ct triage neural network. *Radiology: Artificial Intelligence*, 3(4):e200229, 2021. doi: 10.1148/ryai.2021200229. URL https://doi.org/10.1148/ryai.2021200229. PMID: 34350412. 16

Haoyu Hu, Xiyao Ma, Shiqi Liu, Linsen Zhang, Xiaoliang Xie, Xiaohu Zhou, and Zeng-Guang Hou. Dino-med3d: Bridging dimension and domain gaps in volumetric segmentation via progressive adaptation. In *MICCAI*, 2026. 3

Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022. 17

Yankai Jiang, Mingze Sun, Heng Guo, Xiaoyu Bai, Ke Yan, Le Lu, and Minfeng Xu. Anatomical invariance modeling and semantic alignment for self-supervised learning in 3d medical image analysis. In *ICCV*, pp. 15859–15869, 2023. 3

Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, et al. Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment. In *CVPR*, pp. 24905–24916, 2025. 3

Wenxuan Li, Alan Yuille, and Zongwei Zhou. How well do supervised 3d models transfer to medical imaging tasks? In *ICLR*, volume 2024, pp. 16389–16404, 2024. 3

Yuheng Li, Yuan Gao, Haoyu Dong, Yuxiang Lai, Shansong Wang, Mojtaba Safari, James E Baciak, and Xiaofeng Yang. Universal ct representations from anatomy to disease phenotype through agglomerative pretraining. *arXiv preprint arXiv:2605.21906*, 2026. 2, 3

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023. 1

Xuefeng Ni, Linshan Wu, Jiaxin Zhuang, Qiong Wang, Mingxiang Wu, Varut Vardhanabhuti, Lihai Zhang, Hanyu Gao, and Hao Chen. Mg-3d: Multi-grained knowledge-enhanced vision-language pre-training for 3d medical image analysis. *Medical Image Analysis*, 111:104027, 2026. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2026.104027. 3

Aaron Nicolson, Jason Dowling, Douglas Anderson, and Bevan Koopman. Longitudinal data and a semantic similarity reward for chest x-ray report generation. *Informatics in Medicine Unlocked*, 50:101585, 2024. 5

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. 3

Suraj Pai, Ibrahim Hadzic, Dennis Bontempi, Keno Bressem, Benjamin H Kann, Andriy Fedorov, Raymond H Mak, and Hugo JWL Aerts. Vision foundation models for computed tomography. *arXiv preprint arXiv:2501.09001*, 2025. 1, 2, 3, 16

D Peeters, B Obreja, N Antonissen, and C Jacobs. Benchmarking of artificial intelligence and radiologists for lung cancer screening in ct: The luna25 challenge. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2025. 17

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021. 3, 5

Mohammad Rahimzadeh, Abolfazl Attar, and Seyed Mohammad Sakhaei. A fully automated deep learning-based network for detecting covid-19 from a new and large lung ct scan dataset. *Biomedical Signal Processing and Control*, 68:102588, 2021. 17

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Spreading vectors for similarity search. In *ICLR*, 2019. 5

Zhongyi Shui, Jianpeng Zhang, Weiwei Cao, Sinuo Wang, Ruizhe Guo, Le Lu, Lin Yang, Xianghua Ye, Tingbo Liang, Qi Zhang, and Ling Zhang. Large-scale and fine-grained vision-language pre-training for enhanced CT image understanding. In *ICLR*, 2025. 3

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel HAZIZA, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Herve Jegou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. 3, 5

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568(C), 2024. 2

Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *CVPR*, pp. 20698–20708, 2022. 3

Tim Veenboer, George Yiasemis, Eric Marcus, Vivien van Veldhuizen, Cees G. M. Snoek, Jonas Teuwen, and Kevin B. W. Groot Lipman. TAP-CT: 3d task-agnostic pretraining of computed tomography foundation models. In *Medical Imaging with Deep Learning*, 2026. 1, 3

Tassilo Wald, Ibrahim Ethem Hamamci, Yuan Gao, Sam Bond-Taylor, Harshita Sharma, Maximilian Ilse, Cynthia Lo, Olesya Melnichenko, Anton Schwaighofer, Noel CF Codella, et al. Comprehensive language–image pre-training for 3d medical image understanding. In *ECCV*, pp. 388–406, 2026. 1, 2, 3, 5

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):62385, 2025a. doi: 10.1038/s41467-025-62385-7. 3

Linshan Wu, Jiaxin Zhuang, and Hao Chen. Large-scale 3d medical image pre-training with geometric context priors. *IEEE TPAMI*, 2025b. 3, 16

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, pp. 432–448, 2018. 18

Yutong Xie, Jianpeng Zhang, Yong Xia, and Qi Wu. Unimiss+: Universal medical self-supervised learning from cross-dimensional unpaired data. *IEEE TPAMI*, 46(12):10021–10035, 2024. 3

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, pp. 18102–18112, 2022. 5

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pp. 11975–11986, 2023. 3

Kang Zhang 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. 16

Hong-Yu Zhou, Chixiang Lu, Chaoqi Chen, Sibei Yang, and Yizhou Yu. A unified visual information preservation framework for self-supervised pre-training in medical image analysis. *IEEE TPAMI*, 45(7):8020–8035, 2023. 3

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image BERT pre-training with online tokenizer. In *ICLR*, 2022. 3, 5

Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B Gotway, and Jianming Liang. Models genesis: Generic autodidactic models for 3d medical image analysis. In *MICCAI*, pp. 384–393, 2019. 3

Weicheng Zhu, Haoxu Huang, Huanze Tang, Rushabh Musthyala, Boyang Yu, Long Chen, Emilio Vega, Thomas O'Donnell, Reya Hayek, Lindsey Kuohn, et al. 3d foundation model for generalizable disease detection in head computed tomography. *Nature Biomedical Engineering*, pp. 1–12, 2026. 1, 2, 3

A Training and Data-Construction Details

A.1 Corpus accounting

The corpus pipeline starts from 533,205 candidate NIfTI volumes collected across nine centers over multiple years. Rule-based cleaning, series categorization, and quality control retain 472,084 eligible CT volumes for pretraining (Section 3.2). Here, a "volume" denotes an eligible reconstructed CT series rather than a unique patient or an independent clinical study. Partitioning volumes with variable through-plane extents produces 956,876 derived z-axis chunks used by the training sampler; these chunks are not counted as independent CT volumes. Spatial dimensions below are reported in $D \times H \times W$ order.

A.2 Visual pretraining configuration

Both visual stages use a 3D ViT-B with 12 transformer blocks, 768-dimensional embeddings, 12 attention heads, a SwiGLU feed-forward network with ratio 3, four storage tokens, and an $8 \times 16 \times 16$ patch embedding. The input consists of three CT window channels with $(c, w) = (40, 400)$, $(40, 80)$, and $(200, 2400)$ HU. Stage 1 is trained from scratch for 300k steps using two $64 \times 256 \times 256$ global crops and eight $32 \times 128 \times 128$ local crops. Stage 2 preserves the student crop configuration and introduces an additional undistorted $128 \times 512 \times 512$ view for the frozen Gram teacher. The Gram weight increases from 0 to 2.0 during the first refinement epoch, while the local DINO weight decreases from 1.0 to 0.5.

A.3 Controlled positional-encoding ablation

The positional-encoding ablation uses 3D ViT-S models with 12 transformer blocks, 384-dimensional embeddings, six attention heads, and 23.7M deployable parameters. Both variants are trained from scratch for 30k steps on the same fixed 10% subset of the pretraining corpus and share identical input channels, crops, masking strategy, objectives, optimizer, learning-rate schedule, and distributed-training configuration. The only experimental difference is the coordinate system supplied to 3D RoPE: the physical-coordinate variant derives patch locations from the transformed affine, whereas the normalized-grid control replaces these coordinates with normalized patch-grid indices. The control therefore removes study-specific physical scale and orientation while preserving the architecture and frequency bank; it is not a comparison between native-grid input and isotropic resampling.

A.4 CT-report alignment configuration

CT-report alignment uses the refined Stage-2 ViT-B checkpoint as a locked visual backbone. Whole-volume CT inputs have a spatial size of $240 \times 480 \times 480$ and use the same three HU window channels as visual pretraining. The trainable visual head contains two transformer blocks and combines the CLS token with mean-pooled patch tokens before projection to a 512-dimensional shared embedding space. The text branch uses the fully trainable `microsoft/BiomedVLP-CXR-BERT-specialized` encoder with a maximum report length of 512 tokens. Optimization uses AdamW with a visual-head learning rate of 5×10^{-5}, a text-encoder learning rate of 5×10^{-6}, and a 300-step warm-up. PD-OSL uses at most eight sentence pairs per CT, a maximum sentence length of 96 tokens, and polarity-debiased balanced negative sampling.

A.5 Evidence-constrained phrase construction

Reports are first divided into eight clinical sections. Candidate atomic findings must be anchored to verbatim evidence spans and describe definite, current, and visually assessable observations. Uncertain, historical, resolved, recommendation-only, acquisition-related, purely comparative, or otherwise nonvisual statements are quarantined. An omitted finding is never treated as negative evidence. For a retained concept c, the affirmative phrase is the evidence-supported statement and its opposite is produced by a deterministic, anatomy-preserving negation operator. Cases with unsafe negation, unresolved scope, or contradictory polarity are excluded. Anatomical elaboration is removed only

Table 5: Complete configurations of the two visual pretraining stages. Stage 2 preserves the Stage-1 student crop resolution and introduces an additional high-resolution view observed by the frozen Gram teacher.

Configuration	ViT-B Stage 1	ViT-B Stage 2
Training role	Volumetric self-supervised pretraining from scratch	High-resolution Gram-anchored visual refinement
Initialization	Random initialization	Student initialized from the Stage-1 EMA teacher at step 300k; frozen Stage-1 teacher at step 200k used as the Gram anchor
Pretraining data	472,084 quality-controlled CT volumes, yielding 956,876 derived z-axis chunks	Same as Stage 1
Backbone	3D ViT-B with 12 blocks, 768-dimensional embeddings, 12 attention heads, SwiGLU FFN with ratio 3, and four storage tokens	Same as Stage 1
Backbone parameters	89,784,576 (89.8M)	89,784,576 (89.8M)
Trainable parameters	127,542,016 (127.5M), including the student backbone and DINO/iBOT heads; EMA teacher excluded	127,542,016 (127.5M), including the student backbone and DINO/iBOT heads; EMA and Gram teachers excluded
CT intensity channels	Three window channels with $(c, w) = (40, 400)$, $(40, 80)$, and $(200, 2400)$ HU	Same as Stage 1
Patch size	$8 \times 16 \times 16$	$8 \times 16 \times 16$
Positional encoding	Physical-coordinate 3D RoPE with axis-wise coordinate normalization, coordinate jitter 1.1, and maximum periods 800/500/500 for $D/H/W$	Same as Stage 1
Global crops	Two $64 \times 256 \times 256$ crops; in-plane scale 0.32–1.0; depth scale 0.8–1.0	Same as Stage 1
Local crops	Eight $32 \times 128 \times 128$ crops; in-plane scale 0.10–0.32; depth scale 0.3–0.6	Same as Stage 1
Gram-teacher view	Not used	One $128 \times 512 \times 512$ undistorted view; teacher features are trilinearly resized for feature matching
Self-supervised objective	DINO with weight 1.0, iBOT with weight 1.0, and KoLeo with weight 0.1	DINO with weight 1.0, iBOT with weight 1.0, KoLeo with weight 0.1, and Gram anchoring increasing from 0 to 2.0 during the first refinement epoch
Local DINO weight	1.0	Decayed from 1.0 to 0.5 during the first refinement epoch
DINO/iBOT prototypes	65,536/32,768	65,536/32,768
iBOT masking	Applied with probability 0.5; masking ratio sampled from 0.1–0.5	Same as Stage 1
Training length	300 epochs $\times$ 1,000 iterations = 300,000 steps	60 epochs $\times$ 1,000 iterations = 60,000 steps
Hardware allocation	48 GPUs with batch size 24 per GPU	88 GPUs with batch size 13 per GPU
Global batch size	1,152	1,144
Optimizer	AdamW, $\beta = (0.9, 0.99)$, weight decay 0.04, gradient clipping 30, and layer-wise LR decay 0.98	Same as Stage 1
Learning rate	Base peak LR 1.5×10^{-4}; effective scaled LR 6×10^{-4}; 30-epoch warm-up	Base peak LR 1×10^{-4}; effective scaled LR 4×10^{-4}; one-epoch warm-up
EMA teacher momentum	0.994	0.999
Numerical precision	BF16 parameters, FP32 gradient reduction, and FP8 transformer blocks	Same as Stage 1
Distributed training	FSDP with `SHARD_GRAD_OP`	Same as Stage 1

when the span contains one safely identifiable finding head; arbitrary multi-fact conjunctions are not split.

Before reconstruction, the extraction yields 172,876 report-level phrase pairs. Mapping these reports to CT reconstructions produces 339,984 pair instances across 47,144 CT rows. Removing only safe anatomical modifiers recovers another 89,224 instances (26.2%), yielding 429,208 instances: 223,885 affirmative and 205,323 explicitly negative. Each CT admits at most eight PD-OSL slots, comprising up to four affirmative and four negative slots. Explicit negatives are selected first, safe weak counterfactuals fill the remaining negative slots, and excess candidates are ignored. Overall, 315,373 of 377,152 possible slots are valid (83.6%); 134,013 are affirmative and 181,360 are negative.

The section-level status (abnormal, normal, or unknown) constrains weak-counterfactual construction but is not interpreted as an exhaustive disease label. During polarity-debiased sampling, the negative branch receives only the normalized affirmative phrase. It does not receive the CT, the negated sentence, a section identifier, or patient metadata. Sampling and evaluation remain patient-

Table 6: Configurations of the controlled ViT-S positional-encoding ablation. All settings are held fixed except the coordinate system supplied to 3D RoPE.

Configuration	ViT-S with Physical-Coordinate RoPE	ViT-S with Normalized-Grid RoPE
Experimental role	Controlled proxy-scale positional-encoding ablation	Controlled proxy-scale positional-encoding ablation
Pretraining data	Fixed 10% subset of the visual pretraining corpus (819 shards)	Identical subset
Backbone	3D ViT-S with 12 blocks, 384-dimensional embeddings, six attention heads, SwiGLU FFN with ratio 4, and four storage tokens	Same
Backbone parameters	23,658,624 (23.7M)	23,658,624 (23.7M)
Trainable parameters	39,654,016 (39.7M)	39,654,016 (39.7M)
Patch size	$8 \times 16 \times 16$	$8 \times 16 \times 16$
Positional encoding	Physical coordinates derived from the transformed affine	Normalized voxel-grid coordinates
Coordinate settings	Coordinate jitter 1.1; maximum periods 800/500/500 for $D/H/W$	Identical settings
Global crops	Two $64 \times 256 \times 256$ crops; in-plane scale 0.32–1.0; depth scale 0.8–1.0	Same
Local crops	Eight $32 \times 128 \times 128$ crops constrained to lie inside global crops; in-plane scale 0.10–0.32; depth scale 0.3–0.6	Same
Objective	DINO (1.0) + iBOT (1.0) + KoLeo (0.1)	Same
DINO/iBOT prototypes	32,768/16,384	Same
iBOT masking	Probability 0.5; masking ratio 0.1–0.5	Same
Training length	30,000 steps	30,000 steps
Global batch size	1,024	1,024
Optimizer	AdamW, $\beta = (0.9, 0.99)$, weight decay 0.04, gradient clipping 30, and layer-wise LR decay 0.98	Same
Learning rate	Base peak LR 1.5×10^{-4}; effective scaled LR 6×10^{-4}; three-epoch warm-up	Same
EMA teacher momentum	0.994	0.994
Numerical precision	BF16 parameters and RoPE, FP32 gradient reduction, and FP8 transformer blocks	Same
Distributed training	32 GPUs, batch size 32 per GPU, FSDP with `SHARD_GRAD_OP`	Same

disjoint. This restriction prevents the balancing model from exploiting the wording of the generated negative itself and makes its selection decision depend only on textual propensity estimated from the affirmative concept.

B Datasets and Evaluation Protocols

B.1 Classification benchmarks

CC-CCII (Zhang et al., 2020) is a three-class chest CT dataset for COVID-19 detection, comprising novel coronavirus pneumonia (NCP), common pneumonia (CP), and normal controls (Normal). Following VoCo (Wu et al., 2025b), we adopt the same patient-level split, with 2,514 training scans (930 NCP, 938 CP, and 646 Normal) and 1,664 validation scans (614 NCP, 618 CP, and 432 Normal).

RSNA-Hem (Flanders et al., 2020) is a large-scale, multi-label brain CT dataset from the RSNA 2019 Brain CT Hemorrhage Challenge, providing annotations for acute intracranial hemorrhage and its five subtypes (subarachnoid, intraventricular, subdural, epidural, and intraparenchymal hemorrhage) typically encountered at brain CT. Following VoCo (Wu et al., 2025b), we use 8,464 training, 1,059 validation, and 1,059 test scans.

SinoCT (Hooper et al., 2021) contains 9,779 head CT scans with a balanced distribution of normal (45%) and abnormal (55%) cases. We adopt the 8:1:1 split used in CT-FM (Pai et al., 2025), resulting in 7,823 training, 978 validation, and 978 test scans.

VinDr-Multiphase (Dao et al., 2022) comprises 1,188 phase-annotated CT scans for four-class classification. The official split includes 830 training scans (138 non-contrast, 279 venous, 340 arterial, and 73 other) and 358 test scans (45 non-contrast, 133 venous, 151 arterial, and 29 other).

Table 7: Complete configuration of CT-report alignment. The Stage-2 visual backbone is frozen throughout alignment.

Configuration	CT-report Alignment
Training role	Locked-image CT-report contrastive alignment
Initialization	Visual encoder initialized from the Stage-2 checkpoint at step 60k; text encoder initialized from `microsoft/BiomedVLP-CXR-BERT-specialized`
Training data	47,144 paired CT volumes and radiology reports after removing two CTs associated with one mismatched report
Frozen components	Complete 3D ViT-B visual backbone, including its final normalization layer
Trainable components	Two-block visual adaptation head, complete text encoder, visual and textual projection layers, and logit-scale parameter
Visual backbone	3D ViT-B with 12 blocks, 768-dimensional embeddings, 12 attention heads, SwiGLU FFN, four storage tokens, and 89.8M frozen parameters
Visual head	Two transformer blocks with drop-path 0.3; combines the CLS token with mean-pooled patch tokens; output dimension 512
Text encoder	CXR-BERT-specialized with maximum sequence length 512; fully fine-tuned with gradient checkpointing
Shared embedding dimension	512
CT intensity channels	Three window channels with $(c, w) = (40, 400)$, $(40, 80)$, and $(200, 2400)$ HU
Input volume and patch size	Whole-volume input of $240 \times 480 \times 480$; patch size $8 \times 16 \times 16$
Positional encoding	Physical-coordinate 3D RoPE inherited from the frozen visual encoder; maximum periods 800/500/500 for $D/H/W$
CT normalization	Each window is clipped to $[0, 1]$; channel means $(0.2199, 0.2037, 0.2769)$ and standard deviations $(0.2998, 0.3637, 0.2487)$
Contrastive objective	Symmetric CT-to-text and text-to-CT CLIP InfoNCE with cross-rank negatives
PD-OSL objective	Up to eight opposite-sentence pairs per CT; loss weight 0.5 with a 300-step linear warm-up
Negative sampling	Polarity-debiased balanced sampling from the reconstructed phrase pairs
Text-pair maximum length	96 tokens
Text microbatch	Four sentence pairs
Temperature	Initialized to 0.07; logit scale constrained to the configured numerical range
Training length	6,000 optimizer steps
Checkpoint interval	Every 200 steps
Hardware allocation	Eight GPUs across two nodes; batch size 16 per GPU
Global batch size	128 contrastive candidates
Optimizer	AdamW, $\beta = (0.9, 0.98)$, $\epsilon = 10^{-8}$, weight decay 0.05, gradient clipping 1.0, and no gradient accumulation
Learning rate	5×10^{-5} for the visual head and 5×10^{-6} for the text encoder; 300-step warm-up
Numerical precision	BF16 parameters with FP32 gradient reduction
Data augmentation	Disabled during whole-volume CT-report alignment

C4KC-KiTS (Heller et al., 2021) comprises 143 renal CT cases of clear cell renal cell carcinoma with ISUP grade annotations. We formulate a binary classification task comparing 97 low-grade cases with 46 high-grade cases using stratified five-fold cross-validation.

We use **COVID-CT**, **LUNA25**, and **AMOS-15** from the CVPR 2026 3D CT Foundation Model Challenge release[1]. COVID-CT (Rahimzadeh et al., 2021) contains 372 chest CT scans, including 308 training scans (248 negative and 60 positive) and 64 validation scans (32 negative and 32 positive). LUNA25 (Peeters et al., 2025) contains 1,200 CT scans for the classification of lung nodule malignancy, including 1,000 training scans (900 benign and 100 malignant) and 200 validation scans (100 benign and 100 malignant). AMOS-15 is an abdominal CT benchmark derived from 1,463 volumes in AMOS (Ji et al., 2022), covering 15 binary disease classification tasks. We follow the official challenge training and validation splits for all three datasets.

B.2 Zero-shot CT–text benchmarks

CT-RATE (Hamamci et al., 2026) contains 1,564 evaluation scans from 1,304 patients and provides 18 thoracic finding labels. **RAD-ChestCT** (Hamamci et al., 2026) contains 3,630 CT volumes without paired reports; evaluation uses 16 mapped labels. Mosaic attenuation is unavailable in RAD-ChestCT, while arterial-wall and coronary-artery calcification are merged to match its taxonomy. We report macro AUPRC and AUROC over the labels available in each benchmark.

For the aggregate results in Table 4, TriPoD-CLIP, OSL, and PD-OSL use the same visual checkpoint, the prompt native to each evaluation dataset, and the same metric implementation. Table 12

[1] https://huggingface.co/datasets/kmin06/CVPR26-3DCTFMCompetition

instead reports two matched prompt regimes: a CLIP-prompt block and a COLIPRI-prompt block. MI2, Curia, and the reproduced COLIPRI variants use a common short-prompt protocol within their comparison block. CT-CLIP AUPRC is reproduced locally, whereas its AUROC is taken from the original report because the reproduced value was comparable. Missing entries remain blank when evaluation is unavailable or not comparable: BUID is not publicly released, Merlin was not trained on CT-RATE, and fVLM uses a different 16-category formulation.

B.3 Downstream evaluation protocols

For classification, all frozen linear-probe evaluations use identical data splits and a shared hyperparameter search space. We additionally evaluate each method under its original adaptation protocol, keeping the encoder frozen or fine-tuning it as prescribed. Results from the two settings are reported separately. AUROC serves as the primary metric, with ACC and F1 reported as secondary metrics. For segmentation, every visual encoder is fully unfrozen and optimized end-to-end with its decoder. Decoder architecture, input preprocessing, test-time augmentation, and post-processing follow the corresponding original paper or official implementation unless stated otherwise. Input spacing and other task-specific settings are selected using training or validation data only. Dice is the primary metric.

Our segmentation decoder uses a 3D adaptation of UPerNet (Xiao et al., 2018). UPerNet builds on a feature pyramid: a top-down pathway and lateral connections combine features at multiple semantic levels, while a pyramid pooling module applied to the highest-level backbone feature supplies global context before pyramid fusion. This design is well suited to ViT encoders because intermediate transformer blocks can be exposed as a multi-level feature hierarchy for dense prediction. DINO-Med3D also uses a UPerNet decoder, making it the most directly aligned decoder-level comparison among the DINO-based baselines.

We do not force every baseline to use the same decoder. The evaluated encoders include convolutional networks, hierarchical Swin Transformers, and plain ViTs, which differ in the number, resolution, and semantics of their feature stages. A single shared decoder would require architecture-specific feature adapters and could advantage the backbone for which it was originally designed. We therefore pair each encoder with its standard or officially recommended decoder while fully fine-tuning the complete encoder–decoder system. Accordingly, the broad segmentation comparison should be interpreted as a comparison of complete fine-tuned segmentation systems; the comparison with DINO-Med3D provides the closest assessment under a matched UPerNet decoder family.

The reported VoCo Swin-B segmentation results are obtained by rerunning its public self-supervised checkpoint, rather than copied from the paper. The VoCo nnU-Net entry uses the VoComni weights. Following CT-FM, TotalSegmentator performance is macro-averaged over organs for each CT; the remaining reported segmentation scores follow the aggregation prescribed by the corresponding evaluation pipeline. Test-time augmentation and post-processing are used only when prescribed by the baseline implementation.

C Additional Visual-Transfer Results

C.1 Experimental setup

Frozen linear probing. All pretrained encoders are frozen in evaluation mode, and a single linear classifier is trained on the pooled features without updating any backbone parameters or normalization statistics. Feature extraction follows each model's native architecture: VoCo, SuPreM, and Swin-UNETR use the deepest SwinViT features with global average pooling; CT-FM, Merlin, and DINO-Med3D use their respective pretrained encoder pooling; FlexiCT, COLIPRI, and SPECTRE retain their shipped pooling or projection heads, with only a fresh linear classifier appended. For our model, the entire ViT-3D backbone is frozen and each input volume is resized to $64 \times 256 \times 256$. All linear probes are trained with AdamW ($\text{lr} = 1 \times 10^{-4}$) and cosine annealing for up to 100 epochs.

Model-specific classification adaptation. For Tables 8 and 9, each method follows the adaptation strategy, input preprocessing, and pooling or classification interface specified by its original paper

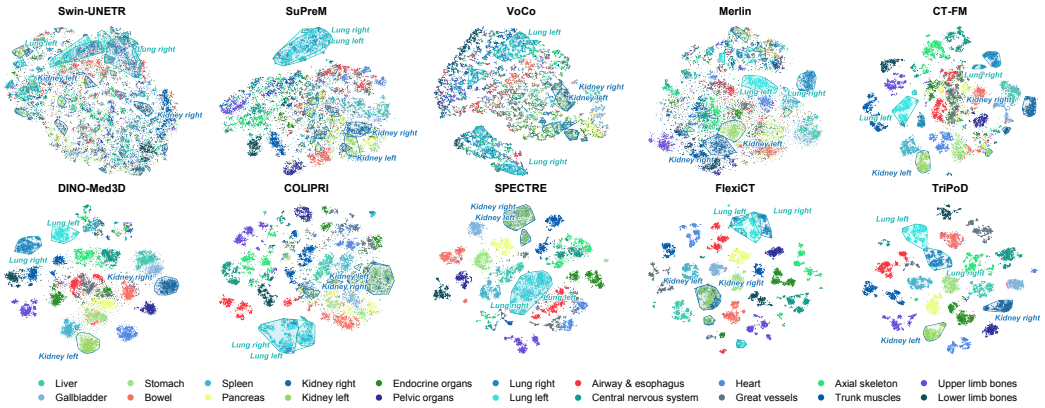

Figure 4: Frozen anatomical feature organization across pretrained CT encoders. Each panel shows an independently fitted two-dimensional t-SNE projection of approximately 18,600 feature vectors extracted from 1,228 TotalSegmentator v2 scans and grouped into 20 semantic organ categories. Colors denote anatomical groups; left and right kidneys and lungs are retained separately to expose laterality structure. The visualization is qualitative because each encoder is projected independently.

Table 8: Model-specific adaptation results on CC-CCII, VinDr-Multiphase, RSNA-Hem, and COVID-CT.

	CC-CCII			VinDr			RSNA-Hem			COVID-CT		
Model	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
Swin-UNETR	98.32	91.71	91.63	97.49	93.01	89.50	94.22	92.01	57.84	88.87	75.00	66.67
SuPreM	99.25	94.47	94.40	97.69	90.50	86.63	94.12	93.01	61.11	94.43	87.50	85.71
VoCo	98.45	94.29	94.18	95.59	91.90	88.12	93.73	93.53	63.50	97.07	93.75	93.33
Merlin	99.03	93.69	93.41	98.05	92.46	88.22	95.33	94.75	71.08	97.66	92.19	91.53
CT-FM	99.11	94.41	94.24	97.77	91.34	87.42	93.22	93.16	66.94	92.09	81.25	77.78
DINO-Med3D	99.48	95.97	95.88	97.15	93.29	89.87	95.34	94.63	68.78	97.66	93.75	93.33
COLIPRI	86.06	68.21	67.99	83.87	65.92	46.81	79.81	87.38	32.99	92.68	50.00	0.00
SPECTRE	95.50	85.28	85.20	85.80	73.18	66.71	81.83	88.41	35.98	98.34	90.62	89.66
FlexiCT	99.42	95.31	95.18	97.53	92.46	88.95	89.94	91.19	52.23	89.45	70.31	59.57
TriPoD	99.41	96.63	96.52	98.10	93.85	89.97	96.20	94.60	71.95	99.80	95.31	95.08

and official implementation. These results estimate attainable downstream performance under each method's intended adaptation setting and are reported separately from the controlled frozen probes.

Frozen-feature visualization. We extracted approximately 18,600 anatomical feature vectors from 1,228 TotalSegmentator v2 CT scans across 20 semantic organ groups. Features were obtained from each model's frozen pretrained encoder and pooled using the fractional occupancy of the corresponding anatomical masks. Each model was independently projected via two-dimensional t-SNE with cosine distance, perplexity of 30, early exaggeration of 4.0, 1,000 iterations, and a fixed random seed of 42. Left and right kidneys and lungs were retained as separate groups to evaluate anatomical laterality preservation. Figure 4 visualizes the resulting frozen-feature organization. Because t-SNE was fitted independently for each encoder, the plots support only within-model qualitative assessment of anatomical grouping and should not be compared by absolute coordinates across panels.

C.2 Complete model-specific adaptation results

Tables 8 and 9 report ACC, AUROC, and F1 for the model-specific adaptation setting summarized in the main paper. Metrics are percentages.

Table 9: Model-specific adaptation results on SinoCT, LUNA25, C4KC-KiTS, and AMOS-15.

	SinoCT			LUNA25			C4KC-KiTS			AMOS-15		
Model	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
Swin-UNETR	83.33	75.15	77.31	50.24	50.00	0.00	67.90	67.68	60.16	73.07	63.84	53.14
SuPreM	85.24	77.20	77.81	56.72	52.00	20.00	72.82	73.33	66.73	75.92	67.05	58.59
VoCo	85.86	77.91	78.74	54.74	51.00	15.52	74.31	75.49	65.66	74.99	65.34	56.70
Merlin	85.84	78.02	79.31	64.69	56.50	43.87	59.35	69.83	52.00	79.91	68.50	58.87
CT-FM	84.20	75.15	78.01	53.76	53.50	37.58	67.93	73.45	62.62	76.08	64.33	55.76
DINO-Med3D	86.53	77.40	77.74	53.88	51.50	28.14	71.08	69.88	64.18	74.71	61.96	42.12
COLIPRI	74.16	69.53	72.66	72.44	65.50	56.05	62.45	69.26	65.30	72.53	71.79	70.91
SPECTRE	72.41	67.28	73.38	73.83	63.50	50.34	75.60	74.83	73.56	77.08	75.18	75.01
FlexiCT	80.12	72.60	73.98	65.20	56.00	30.16	75.75	81.13	78.95	73.24	71.96	71.66
TriPoD	87.79	79.75	81.59	76.91	68.00	65.59	78.12	76.51	67.88	81.31	73.20	70.98

Table 10: Controlled ViT-S positional-encoding ablation on classification. Values are percentages.

Protocol	PE	SinoCT		VinDr		RSNA
		ACC	AUC	ACC	AUC	ACC
Frozen	Normal	73.51	81.36	88.26	97.35	89.22
Frozen	Physical	72.18	80.70	89.38	97.74	88.72
Adapter	Normal	75.25	84.34	91.89	98.27	92.12
Adapter	Physical	76.89	85.49	91.34	98.31	92.83

D Additional Positional-Encoding Analyses

D.1 Classification transfer

Table 10 gives the complete controlled ViT-S classification ablation. The physical-coordinate variant is not uniformly better under frozen probing, while adapter tuning yields gains on SinoCT and RSNA-Hem and nearly unchanged VinDr AUROC. We therefore interpret physical coordinates as task-dependent spatial information rather than a universal improvement.

D.2 Boundary-sensitive segmentation metrics

Table 11 expands the main-text Dice/NSD@1 mm result on the two large cohorts. Physical coordinates improve every reported metric on TotalSegmentator v2. On AbdomenAtlas1.0Mini, all paired intervals include zero, indicating practically indistinguishable performance under this protocol. Intervals are case-level paired 95% confidence intervals.

D.3 Fixed-spacing audit

The evaluated TotalSegmentator v2 pipeline standardizes all volumes to $1.5 \times 1.5 \times 1.5\,\text{mm}$ before inference. Residual variation in the recorded evaluation-grid z spacing (1.76×10^{-7}), voxel anisotropy (1.20×10^{-7}), and token anisotropy (3.17×10^{-7}) is numerical noise. Consequently, these outputs cannot support a meaningful spacing-stratified or spacing-correlation analysis; we report no such claim.

E Additional Language-Alignment Analysis

E.1 Category-level effect of polarity-debiased sampling

Table 12 reports two separate matched-prompt comparisons. The left block evaluates CLIP and PD-OSL with the CLIP prompt; the right block evaluates paper-style OSL and PD-OSL with the COLIPRI prompt. PD-OSL improves AUROC for 15 of 18 categories relative to CLIP and for 12 of 18 categories relative to OSL. Each delta is computed within its block; the two PD-OSL columns use different prompts and should not be compared across blocks. This analysis is distinct from Table 4, where all TriPoD variants use the same dataset-specific prompt.

Table 11: Complete per-case macro metrics for the controlled ViT-S segmentation ablation. Higher is better for Dice and NSD; lower is better for HD95.

Dataset	Metric	Normal PE	Physical PE	Δ	95% CI
TotalSegmentator v2 ($n = 248$)	Macro Dice ↑	84.67	86.57	+1.90	[+1.16,+2.73]
	NSD@1 mm ↑	69.58	71.63	+2.04	[+1.45,+2.71]
	NSD@2 mm ↑	85.91	87.94	+2.03	[+1.30,+2.86]
	NSD@3 mm ↑	88.61	90.54	+1.93	[+1.20,+2.76]
	HD95 ↓	14.01	6.82	-7.19	[-9.13,-5.46]
AbdomenAtlas1.0Mini ($n = 1{,}038$)	Macro Dice ↑	90.827	90.807	-0.020	[-0.087,+0.047]
	NSD@1 mm ↑	80.797	80.836	+0.039	[-0.042,+0.121]
	NSD@2 mm ↑	89.688	89.718	+0.030	[-0.053,+0.113]
	NSD@3 mm ↑	92.841	92.862	+0.021	[-0.065,+0.106]
	HD95 ↓	10.383	10.423	+0.040	[-0.354,+0.434]

Table 12: CT-RATE category-level AUROC under two matched prompt regimes. The left block uses the CLIP prompt for both methods; the right block uses the COLIPRI prompt for both methods. N_+ is the number of positive examples, and Δ is the absolute percentage-point change from the corresponding baseline within the same block.

		CLIP prompt			**COLIPRI prompt**		
Finding	N_+	CLIP	PD-OSL	Δ	OSL	PD-OSL	Δ
Medical material	158	83.5	91.7	+8.19	87.0	91.3	+4.30
Arterial-wall calcification	435	87.9	92.7	+4.74	93.1	92.7	-0.36
Cardiomegaly	164	88.8	92.7	+3.91	92.3	92.9	+0.58
Pericardial effusion	110	85.3	86.5	+1.24	85.0	86.2	+1.27
Coronary-artery calcification	388	88.5	91.9	+3.38	92.4	90.3	-2.16
Hiatal hernia	218	67.6	73.1	+5.52	71.2	73.4	+2.24
Lymphadenopathy	398	72.9	72.5	-0.44	72.7	72.6	-0.18
Emphysema	305	73.7	78.8	+5.15	78.2	79.1	+0.98
Atelectasis	364	67.4	72.3	+4.93	73.1	73.6	+0.50
Lung nodule	704	57.7	60.7	+3.01	65.1	61.3	-3.78
Lung opacity	606	81.4	80.2	-1.15	77.0	79.6	+2.64
Pulmonary fibrotic sequela	427	60.0	68.4	+8.45	68.5	68.8	+0.28
Pleural effusion	189	95.3	97.3	+1.98	96.9	97.2	+0.32
Mosaic attenuation	126	71.0	85.0	+14.02	79.3	85.7	+6.38
Peribronchial thickening	175	73.5	79.0	+5.47	78.3	78.8	+0.43
Consolidation	298	83.2	86.5	+3.35	88.1	87.2	-0.86
Bronchiectasis	169	70.4	72.5	+2.06	76.3	73.3	-2.95
Interlobular septal thickening	124	84.8	83.7	-1.04	82.2	82.4	+0.20

F SEGMENTATION BENCHMARK REFERENCE

Table 13 expands the abbreviated names used in the main segmentation comparison.

Table 13: Public segmentation benchmarks used in Table 1. The middle column gives the complete dataset or challenge name; the final column summarizes the evaluated anatomy or lesion target.

Benchmark	Complete public benchmark name	Segmentation target
MSD Liver Tumor	Medical Segmentation Decathlon Task03: Liver	Liver and hepatic tumors
MSD Lung Tumor	Medical Segmentation Decathlon Task06: Lung	Lung tumors
MSD Pancreas Tumor	Medical Segmentation Decathlon Task07: Pancreas	Pancreas and pancreatic tumors
MSD Hepatic Vessel	Medical Segmentation Decathlon Task08: Hepatic Vessel	Hepatic vessels and liver tumors
MSD Colon Tumor	Medical Segmentation Decathlon Task10: Colon	Colon tumors
AISD	Acute Ischemic Stroke Dataset	Acute ischemic infarct lesions on non-contrast head CT
MM-WHS	Multi-Modality Whole Heart Segmentation Challenge	Whole-heart substructures in the CT subset
SegTHOR	Segmentation of Thoracic Organs at Risk in CT Images	Heart, aorta, trachea, and esophagus
INSTANCE 2022	Intracranial Hemorrhage Segmentation Challenge (INSTANCE) 2022	Intracranial hemorrhage on non-contrast head CT
AbdomenAtlas1.0Mini	AbdomenAtlas 1.0 Mini	Abdominal organ structures
TotalSegmentator v2	TotalSegmentator version 2	Whole-body anatomical structures