mDINO: A SEQUENTIAL 2D–3D LATENT-SPACE SELF-DISTILLATION PRE-TRAINING FRAMEWORK FOR MEDICAL FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical foundation models based on self-supervised learning (SSL) follow two paradigms: reconstruction-based methods capture the local texture at the expense of global semantics, and contrastive methods yield discriminative representations at the cost of fine-grained local details. Latent-space self-distillation offers a unified alternative to both paradigms, yet has not been adopted in medical foundation models, especially 3D domain, because of two main challenges: 1) Latent-space self-distillation methods demands large-scale pre-training corpora that far exceed the scale of available medical datasets. 2) Existing latent-space self-distillation models, such as DINOv3, are pre-trained for 2D natural images, leaving the domain gap of 3D medical foundation models unaddressed. In this paper, we propose mDINO, a sequential 2D-to-3D latent-space self-distillation pre-training framework for medical foundation models. To address data scarcity, we construct a large-scale CT corpus of 113M axial slices extracted from 1.13M volumes spanning diverse anatomical regions. mDINO first conducts 2D pre-training on 2D corpus, then leverages the spatial correspondence between 2D corpus and their source 3D volumes to transfer the learned representations as a structurally consistent initialization for 3D pre-training. To bridge the domain gap and enable latent-space self-distillation on 3D volumes, we further introduce 3D Hilbert flatten for locality-preserving token serialization and isotropy 3D-medical-RoPE for volumetric positional encoding. mDINO achieves superior performance over all reconstruction-based and contrastive-based medical foundation models on 40+ widely used 2D and 3D medical benchmarks. Code and pre-trained weights will be released.

1 INTRODUCTION

Medical foundation models aim to learn generalizable representations that transfer across diverse clinical tasks, including organ segmentation, lesion detection and radiology report generation. Self-supervised learning (SSL) is the dominant pre-training strategy for such models. During SSL, reconstruction-based methods, such as masked autoencoders (2022; 2025b; 2022), recover masked image regions and yield strong local texture representations, yet their pixel-level objectives provide limited supervision for global semantic understanding. Contrastive methods (2024; 2025) learn instance-discriminative features by pulling together augmented views of the same image, yet their view-level objectives discard fine-grained spatial details that are critical for dense prediction tasks such as segmentation. As a result, neither paradigm alone produces representations that generalize across the different kinds of medical imaging tasks.

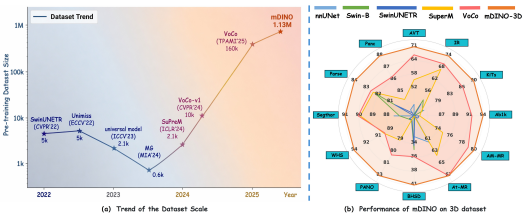


Figure 1: (a) 3D Pre-training dataset scales across recent medical SSL works. mDINO is trained on a large-scale medical CT corpus (1.13M volumes). (b) Performance comparison on 2D/3D segmentation across multiple modalities.

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.

Latent-space self-distillation(2025) offers a unified alternative by jointly optimizing image-level alignment and patch-level masked prediction in the feature space, yet this class of methods has not been adopted in medical imaging especially 3D domain. The first obstacle is the data scale: latent-space self-distillation methods require large-scale pre-training corpora to stabilize the self-distillation training, whereas existing medical imaging datasets, shown in Fig. 1, are substantially smaller in scale. The second obstacle is the domain gap: 2D models can only process volumetric data as independent slices, discarding the inter-slice spatial context that encodes critical anatomical structure. Volumetric pre-training is therefore necessary, yet extending latent-space self-distillation to 3D introduces substantially training complexity, especially in locality-preserving token serialization and anisotropy-aware positional encoding.

To address both obstacles, we propose mDINO, a sequential 2D-to-3D latent-space self-distillation pre-training framework consisting of mDINO-2D and mDINO-3D. We construct a large-scale CT corpus of 113M axial slices extracted from 1.13M volumes, where the spatial correspondence between 2D slices and their source 3D volumes permits direct weight transfer from mDINO-2D to initialize mDINO-3D, substantially reducing volumetric optimization difficulty. To enable latent-space self-distillation on 3D volumes, we also introduce 3D Hilbert flatten for locality-preserving token serialization and propose isotropy 3D-medical-RoPE for volumetric positional encoding. The contributions of this work are as follows:

- **Large-scale medical CT corpus.** A corpus of 1.13M 3D volumes and 113M 2D axial slices is constructed, with 2D slices extracted from their corresponding 3D volumes to preserve spatial correspondence for sequential pre-training.
- **3D architectural design and sequential pre-training.** 3D Hilbert flatten and isotropy 3D-medical-RoPE are utilized to extend latent-space self-distillation to 3D medical volumes. A sequential strategy transfers mDINO-2D representations to initialize mDINO-3D, exploiting the structural consistency between 2D slices and 3D volumes.
- **State-of-the-art performance.** mDINO achieves superior performance over reconstruction-based and contrastive-based medical foundation models on the 40+ widely used 2D and 3D segmentation benchmarks, with transferability further validated on the radiology report generation task.

2 Related Work

Medical Foundation Models. Medical foundation models based on reconstruction-based methods and contrastive methods aim to learn generalizable representations transferable across diverse clinical tasks. Reconstruction-based methods (2022; 2022; 2025b) recover masked inputs at the pixel level, yielding representations rich in local structures but limited in semantic abstraction. Contrastive methods (2020; 2020; 2024; 2025) enforce instance-level discrimination, producing globally separable embeddings at the expense of fine-grained spatial details required for dense prediction. Latent-space self-distillation, as instantiated by the DINO family (2021; 2023; 2025), bridges this divide by jointly optimizing image-level semantic alignment and patch-level masked prediction in the feature space. Recent work has demonstrated that transferring DINOv2/v3 pre-trained on natural images to 2D medical imaging yields effective representations (2025; 2025b). Training dedicated medical latent-space self-distillation foundation models from scratch is therefore important. Empirically, mDINO trained from scratch yields superior performance over DINOv3 models transferred from natural 2D images. However, this remains unattained: available medical datasets fall short of the scale required to stabilize self-distillation objectives, and no existing implementation supports volumetric self-distillation pre-training for 3D medical data.

3D Medical Self-supervised Learning. Among native 3D approaches that train directly on volumetric data, VoCo (2024; 2025) achieves state-of-the-art performance in 3D medical contrastive pre-training with comprehensive benchmark coverage, and SuPreM (2024) pursues supervised pre-training at scale as an alternative. Another line of work investigates 2D-to-3D transfer (2023) through convolution inflation or temporal attention replication during SSL, yet spatial information is inevitably lost when volumetric data is reduced to independent 2D slices. On the technical side, space-filling curves have been adopted in SSM-based architectures (2024; 2024) to preserve spatial locality during token serialization, and 3D extensions of RoPE have been explored in video Transformers (2024; 2023) for volumetric positional encoding. However, latent-space self-distillation with medical-special flatten and RoPE remains largely unexplored.

3 METHOD

Our mDINO comprises two sequentially trained models, mDINO-2D and mDINO-3D, as shown in Fig. 2. Stage 1 pre-trains mDINO-2D on CT-2D (113M axial slices) under the DINOv3 objective. Stage 2 transfers all ViT block weights from mDINO-2D to initialize mDINO-3D, then continues pre-training on CT-3D (1.13M volumes). To extend latent-space self-distillation to 3D volumes, Stage 2 introduces two contributions: 3D Hilbert flatten replaces the default Raster-order flatten to produce locality-preserving token sequences for volumetric masked patch prediction, and isotropy 3D-medical-RoPE replaces 2D positional encoding with a three-axis isotropic formulation suited to the near-isotropic voxel geometry of medical CT. These two modules operate jointly within the ViT of mDINO-3D: 3D Hilbert flatten governs the token order passed to the iBOT masking module(2025), while isotropy 3D-medical-RoPE injects spatial position into every self-attention layer.

3.1 DATASET CONSTRUCTION

The 2D-3D unsupervised high-quality dataset is visualized in Fig. 3 and constructed as follows:

- CT-3D: Dataset aggregates 2.93M CT scans (including limited CTA) with scan ranges of 150–2000 mm, covering the thorax, abdomen–pelvis, and cranium. The dataset contains 1.13M clinical cases after rigorous quality control. Inclusion criteria are: 1) patients $\geq$12 years old to exclude developmental skeletal variations; 2) axial depth of 32–128 slices ensuring volumetric completeness; 3) slice thickness $\leq$2 mm for near-isotropic voxel resolution; 4) matrix size of 512$\times$512 pixels; 5) exclusion of scans with $>$30% invalid voxels (HU$\leq$-1024) to remove artifacts, motion blur, and out-of-field contamination. The dataset is collected from multi centers to support cross-anatomical deep learning pre-training.
- CT-2D: The 2D high-quality dataset is sampled from CT-3D and contains 113M CT slices. It comprises three clinical CT modalities with strict quality criteria: 1) chest CT—non-contrast volumetric chest scans covering pulmonary parenchyma, mediastinal structures, and thoracic wall soft tissues; 2) lumbar spine CT—imaging sequences encompassing L1–L5 vertebrae, intervertebral discs, paravertebral musculature, and adjacent soft tissues; 3) brain CT—cranial volumes including the cerebral cortex and white matter, ventricular system, basal ganglia, and pathological lesions. All sampled cases satisfy the following inclusion criteria: 1) slice thickness $\leq$10 mm, 2) matrix size of 512$\times$512 pixels, and 3) axial slice count $\geq$16, ensuring adequate spatial resolution and volumetric continuity for clinical interpretation.

The dataset contains 1.13M 3D volumes and 113M 2D slices, largely surpassing the scale of existing medical imaging datasets, as shown in Fig. 1(a). This scale attains the data requirement for latent-space self-distillation pre-training, a dataset seldom accessible to prior medical imaging corpora. The multi-center provenance, rigorous quality control, and cross-anatomical scope further promote representation diversity and clinical applicability.

3.2 STAGE 1: PRE-TRAINING MDINO-2D ON CT SLICES

mDINO-2D is initialized from scratch and pre-trained on CT-2D under the DINOv3 pipeline including the loss function and optimizer. The mDINO-2D processes every sample of 113M axial slices as 16×16 patches. Pre-training on CT-2D shapes patch-level representations for medical CT intensity distributions rather than natural-image statistics, establishing a domain-aligned initialization for Stage 2. All ViT block weights of mDINO-2D are subsequently transferred to mDINO-3D after pre-training in Stage 1.

3.3 STAGE 2: MDINO-3D PRE-TRAINING ON VOLUMETRIC DATA

The ViT of mDINO-3D retains all block-level weights from mDINO-2D, including self-attention projection matrices, feed-forward network weights, and LayerNorm parameters. Three dimension-dependent components are replaced: the 2D patch embed is upgraded to a 3D patch embed to accept volumetric inputs; the raster-order flatten is replaced with a 3D Hilbert flatten; and RoPE-2D is replaced with an isotropic 3D medical RoPE. Since the CT-2D slices used in Stage 1 are extracted from the same CT-3D volumes that constitute the Stage 2 pre-training corpus, the two stages share both

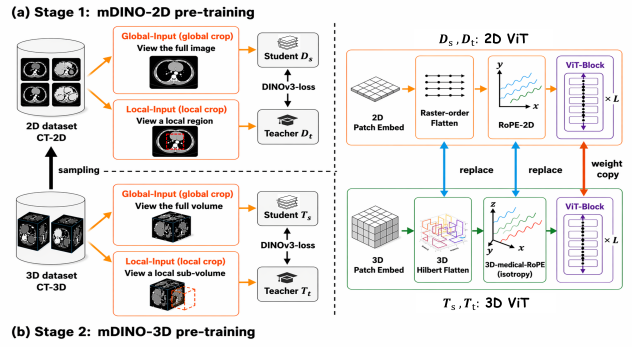


Figure 2: Overview of the mDINO two-stage pre-training pipeline. **(a) Stage 1: mDINO-2D pre-training.** A ViT student network (D_s) learns from an EMA-updated teacher (D_t) through the DINOv3 pipeline. Training is conducted on 113M axial slices extracted from 3D CT volumes. **(b) Stage 2: mDINO-3D pre-training.** All ViT block weights are transferred from mDINO-2D; dimension-dependent components are re-initialized, including 3D patch embedding, 3D Hilbert flatten for locality-preserving serialization, and isotropy 3D-medical-RoPE for volumetric positional encoding.

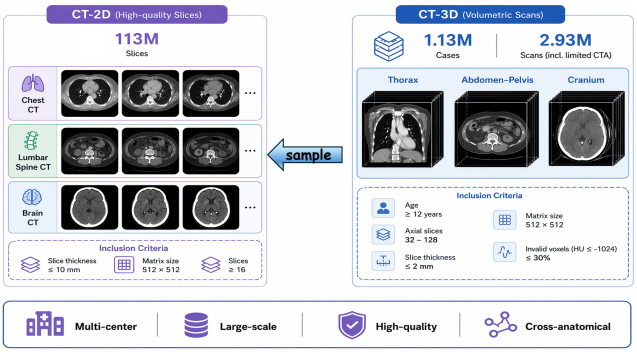


Figure 3: The construction pipeline of CT-2D and CT-3D. CT-3D aggregates 1.13M clinical CT volumes after quality control, covering thorax, abdomen–pelvis, and cranium. CT-2D contains 113M axial slices sampled from CT-3D across three modalities: chest CT, lumbar spine CT, and brain CT.

the imaging modality and the underlying data distribution. The patch-level representations learned in Stage 1 are thus already adapted to this distribution, and inheriting them yields a considerably more informative initialization than random initialization, easing the subsequent 3D pre-training of mDINO-3D. mDINO-3D is subsequently pre-trained on 1.13M 3D CT volumes under the same pipeline, including the loss function, as Stage 1 (see Supplementary A for full training details).

3D Hilbert Flatten for Locality-preserving Serialization. A 3D patch embedding partitions the input volume into non-overlapping patches and projects each to a token vector, yielding a set of tokens indexed by their 3D voxel coordinates. mDINO-3D serializes these tokens by traversing a 3D Hilbert space-filling curve $\mathcal{H} : \{0, \ldots, N-1\}^3 \to \{0, \ldots, N^3-1\}$, satisfying unit-step adjacency:

$$\|\mathcal{H}(i+1) - \mathcal{H}(i)\|_1 = 1, \quad \forall i. \tag{1}$$

Every consecutive token pair in the 1D sequence thus corresponds to face-sharing 3D neighbors. In the DINOv3 framework, iBOT(2025) applies block masking over contiguous 1D spans; under Hilbert flatten, such spans map to compact 3D neighborhoods, producing spatially coherent masked regions for volumetric patch prediction.

Isotropy 3D-medical-RoPE for Volumetric Positional Encoding. Standard 2D RoPE encodes only (x, y) positions, leaving tokens at different depth levels positionally indistinguishable. mDINO-3D extends RoPE to three spatial dimensions by assigning independent rotation generators to each axis:

$$R_{x,y,z} = \mathrm{diag}(R(x),\, R(y),\, R(z)), \tag{2}$$

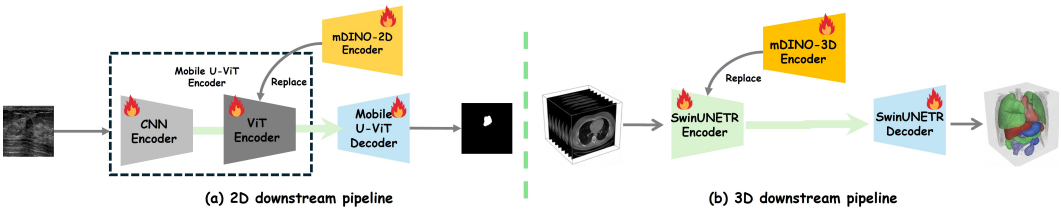


Figure 4: Downstream adaptation protocol. For 2D segmentation, the ViT encoder of Mobile U-ViT is replaced with the mDINO-2D pre-trained ViT encoder. For 3D segmentation and radiology report generation, the encoder of SwinUNETR is replaced with the mDINO-3D pre-trained ViT encoder. All full networks are then fine-tuned in an end-to-end manner.

where each rotation block is defined as:

$$R(t) = \begin{pmatrix} \cos t\theta & -\sin t\theta \\ \sin t\theta & \cos t\theta \end{pmatrix}. \tag{3}$$

The isotropic design assigns a single shared frequency parameter θ to all three axes, reflecting the near-isotropic voxel resolution of medical CT volumes where each spatial axis carries equivalent geometric meaning. This shared frequency constitutes a principled domain-specific inductive bias for volumetric medical data. The relativity property $R_{x_1}^T R_{x_2} = R_{x_2 - x_1}$ is preserved, enabling translation-equivariant attention. Derivation is provided in Supplementary A.6.

4 Experiments

mDINO is evaluated on 40+ benchmarks spanning 2D and 3D medical segmentation, with additional evaluation on radiology report for transferability validation. Benchmark descriptions are provided in Supplementary A.3.

4.1 Implementation Details

Pre-training Configuration. Pre-training is conducted in two sequential stages. In Stage 1, mDINO-2D is trained on CT-2D (113M slices) for 5 epochs under the DINOv3 pipeline. In Stage 2, all ViT block weights are transferred from mDINO-2D to initialize mDINO-3D, which is then trained on CT-3D (1.13M volumes) for an additional 5 epochs under the same objective with identical loss weights. We adopt ViT-S with patch size 16 as a representative configuration. Optimization follows AdamW with learning rate 2×10^{-3} and cosine decay. Total pre-training requires approximately 15,000+ H200 GPU hours.

Downstream evaluation protocol. All downstream evaluations are inherited from established public pipelines(2025a; 2025) rather than custom settings, so that every result is obtained under the same held-out protocol as the corresponding prior work. For mdino in 2D segmentation, shown in Fig. 4(a), Mobile U-ViT (2025a) is adopted as the downstream framework, following its published data splits, preprocessing, training configurations, and baseline settings; only its ViT encoder is replaced by the mDINO-2D pre-trained ViT-S backbone, with the decoder architecture, loss function, and all training hyperparameters unchanged. For mdino in 3D segmentation and radiology report generation, shown in Fig. 4(b), SwinUNETR (2021) is adopted as the downstream framework with its encoder replaced by the mDINO-3D pre-trained ViT-S backbone, following the splits, preprocessing, and training settings of VoCo (2025). Both pipelines are publicly released with complete training code and configurations. Full fine-tuning is performed for all experiments, and all baselines are integrated and fine-tuned under this same encoder-replacement protocol to ensure fair comparison.

4.2 Representation Analysis

Fig. 5 visualizes the semantic features from the final ViT layer via PCA projections and cosine similarity maps. The PCA projections show that mDINO can assign consistent color codes to regions of the same anatomical structure, while color transitions are abrupt at organ boundaries.

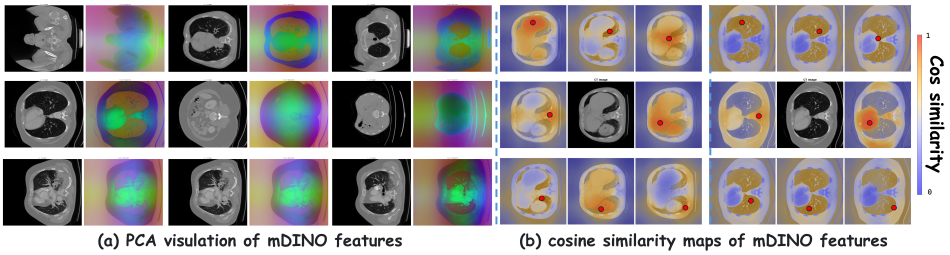


Figure 5: (a) PCA visualization of semantic features from the mDINO encoder. The first three principal components are mapped to RGB values. Semantically consistent regions exhibit coherent color patterns, while organ boundaries correspond to clear feature transitions. (b) Cosine similarity maps of mDINO feature. We visualize the cosine similarity maps between the patches marked with a red dot and all other patches. High similarity responses are concentrated within the same anatomical structure.

And the cosine similarity maps of mDINO further show that high-response regions concentrate within individual anatomical structures, indicating that the learned representations are organized by anatomical structure rather than local appearance.

4.3 2D Segmentation

Table 1: Results on 2D medical segmentation tasks (mean IoU). † denotes deep supervision.

Network	Params↓	GFLOPs↓	IoU (%)				
			Ultrasound			Dermoscopy	Avg
			BUS	BUSI	TNSCUI	ISIC	
heavy-weight medical network							
U-Net (2015)	34.52 M	65.52	86.73	68.61	75.88	82.18	78.35
CMU-Net (2023)	49.93 M	91.25	87.18	71.42	77.12	82.16	79.97
nnUNet (2021)	26.10 M	12.67	87.51	72.11	78.99	83.31	80.48
Swin-Unet (2022)	27.14 M	5.91	85.27	63.59	75.77	82.15	76.70
TransUnet (2021)	105.32 M	38.52	87.35	71.39	77.63	83.17	79.89
UCTransNet (2022)	66.24 M	32.98	85.98	68.47	73.87	82.67	77.75
MissFormer (2021)	35.45 M	7.25	81.92	63.29	68.26	80.99	73.62
EMCAD (2024)	26.76 M	5.60	87.81	71.09	77.48	82.91	79.82
light-weight natural network							
MobileViT-s (2021)	10.66 M	2.13	82.57	64.28	71.64	80.12	74.65
EdgeViT-s (2022)	16.49 M	2.71	81.32	61.12	68.74	79.06	72.56
EMO-6m	9.14 M	1.85	84.89	67.50	74.02	81.31	76.93
EfficientViT-b1 (2023)	4.79 M	9.18	84.81	67.21	71.18	82.22	76.36
RepViT-m3 (2024)	14.37 M	2.79	76.51	56.07	66.21	78.34	69.78
UniRepLKNet (2024)	5.83 M	9.39	84.13	65.26	67.73	81.64	74.69
light-weight medical network							
MedT (2021)	1.37 M	2.40	80.81	63.36	71.00	81.79	74.24
UNeXt (2022)	1.47 M	0.58	84.73	65.04	71.04	82.10	75.73
EGE-Unet (2023)	0.072 M	0.045	84.72	58.90	74.47	82.19	75.07
ERDUnet (2023)	10.21 M	10.29	85.92	68.67	76.02	82.03	78.16
CMUNeXt (2024a)	3.14 M	7.41	87.27	71.20	76.96	82.47	79.48
TinyU-Net (2024)	0.48 M	1.66	85.67	66.21	74.03	81.95	76.97
Universal Model							
MedSAM (2024)	91M	-	88.83	72.91	80.27	-	-
I-JEPA (2023)	22M	-	88.59	68.34	74.70	-	-
MM-DINOv2 (2025)	86M	-	88.16	73.41	77.58	-	-
MedDINOv3 (2025b)	86M	-	88.05	73.92	79.01	-	-
DINO U-Net (2025a)	27M	-	86.34	72.06	77.70	-	-
SegDino (2025)	24M	-	87.78	72.12	78.36	-	-
Mobile U-ViT-L† (2025a)	7.89 M	3.71	87.87	73.80	78.62	83.53	80.96
+DINOv3-LVD-1689M (ViT-S)(2025)	22.08 M	-	86.68	72.31	74.56	83.1	79.16
+DINOv3-LVD-1689M (ViT-B)(2025)	86.00 M	-	87.77	69.00	77.16	-	-
+DINOv3-LVD-1689M (ViT-L)(2025)	300.00 M	-	89.01	69.76	77.96	-	-
+mDINO-2D (ViT-S)	22.08 M	3.71	**89.43**	**74.56**	**80.47**	**84.35**	**82.20**
+mDINO-2D (ViT-B)	86.00 M	-	**90.02**	**75.34**	**81.07**	**85.51**	**82.99**
△ *vs. Mobile U-ViT-L*	-	-	**+2.15**	**+1.54**	**+2.45**	**+1.98**	**+2.03**

Tab. 1 evaluates mDINO-2D on four out-of-domain benchmarks, covering three ultrasound datasets and one dermoscopy dataset, to show representation transferability beyond the CT pre-training domain. mDINO-2D achieves 82.99% on average IoU, surpassing DINOv3, indicating that CT-domain pre-training yields representations more transferable to medical segmentation than large-

scale natural-image pre-training. Furthermore, mDINO-2D outperforms task-specific architectures including CMU-Net (2023), UCTransNet (2022), and TransUnet (2021) with 2–5× fewer parameters, demonstrating that a general-purpose pre-trained backbone can match or exceed architectures designed and optimized for medical image segmentation.

4.4 3D SEGMENTATION

mDINO-3D is evaluated across 39 segmentation tasks covering abdominal, thoracic, cardiac, and neurological structures across CT and MRI modalities, including organ delineation, tumor detection, and lesion segmentation.

Table 2: Results of 3D CT multi-organ segmentation (mean DSC).

Method	Model (Params)	Data Scale	BTCV(2015)	AMOS(2022)	FLA22(2022)	Total(2023)	Atlas(2023)	△%(AVG)
From Scratch	3D-UNet(2016) (19M)		80.98	84.02	87.58	-	-	-
	UNETR (2022) (115M)		79.82	82.52	89.02	-	-	-
	nnUNet (2021) (31M)		79.29	88.79	91.38	82.26	88.77	-
	MedNext(2023)(63M)		82.36	88.21	90.13	-	-	-
	V-JEPA(2025)(–)		76.95	83.75	87.30	-	-	-
	Swin-B (2021) (72M)		82.79	87.19	89.18	80.97	88.12	85.65
	Swin-L (2021) (290M)		83.52	87.53	89.49	82.04	88.56	86.23
	Swin-H (2021) (1.2B)		79.36	86.14	87.96	82.78	88.97	85.04
General Pre-training								
MAE3D (2023)	UNETR	10k	82.48	82.71	89.31	-	-	-
MoCo v3 (2023)	Swin-B	1.6k	79.54	80.95	83.22	-	-	-
Medical Pre-training								
MG (2021)	3D-UNet	0.6k	81.45	81.27	85.02	-	-	-
DoDNet (2021)	3D-UNet	1k	81.10	79.63	86.19	-	-	-
Unimiss (2024)	MiT(48M)	5k	82.05	86.26	89.17	-	-	-
SwinUNETR (2022)	Swin-B	5k	82.58	85.68	89.31	-	-	-
Universal Model (2023)	Swin-B	2.1k	83.74	88.01	91.11	-	-	-
SuPreM (2024)	Swin-B	2.1k	85.32	88.14	91.37	82.96	89.16	87.39
VoCo-v1 (2024)	Swin-B	10k	84.51	88.06	91.29	80.46	89.13	86.69
VoCo-SSL(2025)	Swin-B	160k	85.34	89.45	92.14	86.32	89.57	88.56
mDINO-3D	ViT-S	1.13M	86.40	89.80	92.80	87.90	90.42	89.43
△ *vs . VoCo-SSL*	-	-	**↑1.06**	**↑0.35**	**↑0.66**	**↑1.38**	**↑0.85**	**↑0.87**

Large-Scale Multi-Organ Segmentation. Tab. 2 evaluates five large-scale benchmarks covering 13–104 anatomical structures per dataset. mDINO-3D achieves 89.43% average DSC, surpassing VoCo-SSL by +0.87% with a smaller backbone. Tab. 3 further evaluates 24 tasks spanning CT, MRI, and diverse anatomical regions. mDINO-3D outperforms VoCo-SSL on 21 of 24 tasks, with consistent margins across abdominal organs (+1.27% on Ab1k, +2.18% on WHS), vascular structures (+2.85% on AVT, +2.92% on IR), and tumor segmentation (+3.01% on BHSD). Cross-modality generalization is further confirmed on MRI tasks: +0.63% on AMOS-MR, +0.89% on WHS-MR, and +0.61% on BraTS.

Table 3: Results of 3D CT/MRI/PET multi-organ segmentation (mean DSC).

Method	Ab1k(2021)	WHS(2018)	AVT(2022)	CHAOS(2021)	Sliver.(2019)	IR.(2010)	KiTs(2023)	Kipa.(2021)
nnUNet (2021)	85.74	88.72	50.19	94.53	94.87	51.26	78.92	88.99
Swin-B (2021)	85.76	89.11	46.76	94.10	94.96	57.19	78.61	85.18
SwinUNETR (2022)	86.32	89.06	46.18	94.98	94.67	55.69	76.82	85.14
SuPrem (2024)	86.40	90.88	58.85	96.42	96.72	68.48	78.38	85.76
VoCo-SSL(2025) (Swin-B)	92.38	90.83	67.57	96.46	97.90	70.52	87.32	87.46
mDINO-3D (ViT-S)	93.65	93.01	70.42	96.84	97.74	73.44	89.64	88.18
△ *vs. nnUNet*	**↑7.91**	**↑4.29**	**↑20.23**	**↑2.31**	**↑2.87**	**↑22.18**	**↑10.72**	**↓0.81**
△ *vs . VoCo-SSL (Swin-B)*	**↑1.27**	**↑2.18**	**↑2.85**	**↑0.38**	**↓0.16**	**↑2.92**	**↑2.32**	**↑0.72**
Method	**Panc.(2016)**	**PANO.(2024)**	**Segthor(2020)**	**BHSD(2023)**	**Struct.(2023)**	**Verse.(2021)**	**Covid.(2022)**	**FUMPE (2018)**
nnUNet (2021)	84.68	78.06	88.15	35.02	70.60	65.13	62.42	48.62
Swin-B (2021)	84.38	78.40	87.90	36.40	**76.42**	62.01	63.91	50.31
SwinUNETR (2022)	84.53	78.34	87.23	35.97	53.36	87.33	65.90	51.72
SuPrem (2024)	85.19	79.92	89.70	32.82	59.85	89.54	63.29	51.98
VoCo-SSL (Swin-B)(2025)	86.29	81.05	89.80	37.15	53.97	-	68.38	49.30
mDINO-3D (ViT-S)	87.33	82.64	90.40	40.16	56.25	-	67.77	51.23
△ *vs. nnUNet*	**↑2.65**	**↑4.58**	**↑2.25**	**↑5.14**	**↓14.35**	↑–	**↑5.35**	**↑2.61**
△ *vs . VoCo-SSL (Swin-B)*	**↑1.04**	**↑1.59**	**↑0.6**	**↑3.01**	**↑2.28**	↑-	**↓0.61**	**↑1.93**
Method	**Parse.(2023)**	**AIIB.(2023)**	**Auto.(2022)**	**AM-MR.(2022)**	**WHS-MR(2018)**	**ACDC(2018)**	**At-MR(2023)**	**BraTs.(2014)**
nnUNet (2021)	80.55	88.72	35.84	72.56	85.36	92.12	63.23	**91.02**
Swin-B (2021)	82.78	89.09	25.25	72.46	86.13	87.22	60.40	89.05
SwinUNETR (2022)	81.66	89.05	22.09	72.89	86.29	89.47	60.51	87.33
SuPrem (2024)	82.88	89.96	24.68	75.69	85.79	89.10	64.64	89.54
VoCo-SSL (Swin-B)(2025)	83.02	89.49	–	78.56	87.16	–	66.80	89.94
mDINO-3D (ViT-S)	83.96	89.83	–	79.19	88.05	–	66.82	90.55
△ *vs. nnUNet*	**↑3.41**	**↑1.11**	↑–	**↑6.63**	**↑2.69**	↑–	**↑3.59**	**↓0.47**
△ *vs . VoCo-SSL (Swin-B)*	**↑0.94**	**↑0.34**	↑–	**↑0.63**	**↑0.89**	↑-	**↓0.02**	**↑0.61**

Medical Segmentation Decathlon. Tab. 4 reports results on the standardized MSD 10-task benchmark using identical data folds as nnUNet (2021). mDINO-3D achieves particularly strong results on challenging tumor segmentation tasks: **78.45%** on Task 03 (liver tumor, +5.45% over VoCo-SSL (2025)) and **72.43%** on Task 07 (pancreas tumor, +4.23%). These results indicate that mDINO

Table 4: The DSC (%) of MSD 10-Task Challenge (2022). Results are reported on identical folds as nnUNet for fair comparison. Tumor DSC is reported for Task03 and Task07.

Method	Task01	Task02	Task03	Task04	Task05	Task06	Task07	Task08	Task09	Task10
nnUNet (2021)	71.25	91.84	67.43	87.08	79.16	64.73	46.08	68.92	92.57	41.69
Swin-B (2021)	71.74	92.28	67.85	88.66	72.64	70.28	47.88	64.80	94.90	35.13
SwinUNETR (2022)	72.79	92.06	64.72	87.01	73.76	71.64	48.31	60.72	95.02	26.19
SuPrem (2024)	70.07	92.55	68.20	87.40	72.92	72.55	50.02	64.71	96.01	38.78
VoCo-SSL (Swin-B) (2025)	73.72	93.41	73.00	88.03	72.83	75.76	68.20	66.38	96.01	38.20
mDINO-3D (ViT-S)	**74.39**	**93.53**	**78.45**	**88.46**	**75.66**	72.78	**72.43**	66.57	**96.38**	**45.35**
△ *vs. VoCo-SSL*	**+0.67**	**+0.12**	**+5.45**	**+0.43**	**+2.83**	-2.98	**+4.23**	**+0.19**	**+0.37**	**+7.15**

representations encode both global anatomical context necessary for organ localization and fine-grained structural detail required for tumor boundary delineation, a capability that reconstruction-based or contrastive-based methods address only partially. mDINO-3D further achieves the highest score on Task 10 (colon tumor, 45.35%), one of the most challenging tasks due to small tumor sizes and high inter-patient variability.

Table 5: Report generation, classification results on CTRG, CC-CCII and CT-RATE benchmark (2024b). BLEU, accuracy, and AUC (%) are reported.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Report Generation				
From Scratch				
Mesh-Memory (2020)	–	–	–	21.0
GSKET (2022)	–	–	–	23.5
Swin-B (2021)	58.90	47.84	40.71	35.63
Swin-H (2021)	59.98	49.00	41.28	36.95
With Pre-training				
PCRLv2 (2023)	58.11	46.96	40.08	35.26
SwinUNETR (2022)	57.15	46.73	40.72	34.28
SuPreM (2024)	58.70	48.13	39.96	35.23
VoCo (2025)	60.45	49.35	42.30	37.42
mDINO-3D (ViT-S)	**62.16**	**49.79**	**43.01**	**37.50**
△ *vs. VoCo*	**+1.71**	**+0.44**	**+0.71**	**+0.08**
Classification				
CC-CCII (Acc)	VoCo 93.80	→	**mDINO 94.73**	△ **+0.93**
CT-RATE (AUC)	VoCo 71.28	→	**mDINO 72.34**	△ **+1.06**

4.5 Vision-Language Task

The experimental protocol of VoCo (2025) is adopted for report generation, including the feature aggregation strategy, language model architecture, training data, and loss function; the only modification is replacing the SwinUNETR encoder with the mDINO-3D pre-trained ViT-S backbone. As shown in Tab. 5, mDINO-3D achieves **62.16%** BLEU-1 (**+1.71%** over VoCo) and **49.79%** BLEU-2, outperforming both Swin-H (59.98%, 49.00%) and VoCo (60.45%, 49.35%).To further probe this transfer, we evaluate mDINO on classification and vocabulary classification under the same encoder-replacement protocol: on CC-CCII, mDINO attains **94.73%** accuracy (**+0.93%** over VoCo), and on CT-RATE vocabulary classification it attains **72.34%** AUC (**+1.06%**).

4.6 Ablation Study

Each architectural component is validated through controlled ablations, as shown in Tab. 6. Pre-training loss serves as the primary metric, with downstream DSC on BTCV(2015) as the evaluation criterion. Due to the huge computational cost, all ablation experiments are conducted with 1 epoch which contains 113M slices and 1.13M volumes. One epoch here means a full pass over 1.13M CT volumes—already exceeding the total pre-training scale of most prior 3D medical SSL works—so this controlled one-epoch experiment is a meaningful fairness check. ; the final mDINO-3D model is trained for 5 epochs (15,216 H200 GPU hours), achieving 86.40% DSC.

Sequential 2D-3D Transfer. Both configurations are trained for 1 epoch under identical settings. Initializing from mDINO-2D reduces pre-training DINOv3 loss(2025) by 34%, from 14.68 to 9.62, and improves DSC by +3.01%, from 82.31% to 85.32%, confirming that 2D-to-3D weight transfer provides a substantially better optimization starting point than random initialization.

Table 6: Ablation study on mDINO-3D components. Pre-training loss and downstream DSC (%) on BTCV validate the contribution of each design choice. All ablation experiments use 1 epoch which contains 113M slices and 1.13M volumes; the final model (last row) uses 5 epochs.

Component	Configuration	Epoch	GPU-h	Loss↓	DSC↑
(a) Sequential 2D→3D Transfer					
Initialization	From scratch	1	3,968	14.68	82.31
	With mDINO-2D	1	3,968	**9.62**	**85.32**
	Improvement	–	–	**-34%**	**+3.01**
(b) Serialization Strategy					
Flatten	Raster-order	1	3,072	11.33	83.44
	Morton/Z-order	1	3,072	10.21	83.96
	Hilbert (ours)	1	3,072	**9.62**	**85.32**
	vs. Raster / Morton	–	–	**-15% / -5.8%**	**+1.88 / +1.36**
(c) Positional Encoding					
Position	2D-RoPE	1	3,072	10.40	84.48
	3D-RoPE (shared θ)	1	3,072	**9.62**	**85.32**
	3D-RoPE (axis-spec. θ)	1	3,072	9.85	84.97
	shared vs. axis-spec.	–	–	–	**+0.35**
(d) Backbone Architecture					
Backbone	Swin-S (49M)	1	3,456	10.58	84.01
	ViT-S (22M)	1	3,072	**9.62**	**85.32**
	Improvement	–	–	**-9%**	**+1.31**
(e) Scale-Controlled Dataset					
SSL Method	VoCo	1	3,072	10.85	83.88
	mDINO	1	3,072	**9.62**	**85.32**
	Equal data gain	–	–	**-11%**	**+1.44**
Baseline (full training)					
	mDINO (full)	**5**	**15,216**	**8.60**	**86.40**

3D Hilbert Flatten vs. Alternative Serialization Strategies. Raster-order, Morton/Z-order, and Hilbert curve serialization are compared under identical configurations. Hilbert achieves the lowest pre-training loss of 9.62 and the highest DSC of 85.32%, outperforming Morton by +1.36% DSC and raster by +1.88% DSC, validating that locality-preserving serialization is the appropriate design choice for iBOT block masking in volumetric data.

Isotropy 3D-Medical-RoPE. Replacing 2D-RoPE with 3D-RoPE under shared θ reduces pre-training loss by 7.5% and improves DSC by 0.84%. A further comparison between shared and axis-specific θ shows that shared parameterization achieves 85.32% DSC, outperforming axis-specific θ by +0.35%, consistent with the near-isotropic voxel spacing of the pre-training corpus.

Backbone Architecture. Swin-S and ViT-S are trained under the identical mDINO objective for 1 epoch, which contains 113M slices and 1.13M volumes. ViT-S achieves 85.32% DSC, outperforming Swin-S by +1.31%, indicating that the ViT architecture is better suited to the mDINO self-distillation objective.

Scale-Controlled Comparison with VoCo. To show the priority of the proposed mDINO, VoCo(2025) with ViT-S is selected as comparison and pre-trained on identical data with ViT-S for 1 epoch. mDINO achieves 85.32% DSC, outperforming VoCo by +1.44%, confirming that the performance gain is attributable to framework design rather than data scale.

5 Conclusion

This paper presents mDINO, a latent-space self-distillation framework adapted for medical CT imaging through a sequential 2D-to-3D pre-training pipeline. A large-scale CT corpus (1.13M volumes, 113M slices) is constructed to support domain-specific pre-training. Two architectural components are introduced for volumetric data: 3D Hilbert flatten for locality-preserving token serialization and isotropy 3D-medical-RoPE for volumetric positional encoding. Compared with random initialization, initializing mDINO-3D from mDINO-2D weights achieves better performance. Experiments across 40+ benchmarks demonstrate consistent improvements in 2D cross-modality segmentation, 3D multi-organ segmentation, radiology report generation and ocabulary classification.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing of the manuscript, including improving the clarity and professionalism of the language. We have not used generative AI tools for any scientific component of this work: all research ideas, method design, experimental implementation, data analysis, and result interpretation were conducted solely by the authors, and no generative AI was involved in writing code, producing figures or tables, or formulating scientific claims or conclusions. Additionally, we used generative AI tools to assist in searching for and retrieving information. We have reviewed all AI-assisted work: every piece of information retrieved with AI assistance was manually verified against the original sources by the authors, and all polished text was carefully checked to ensure that it accurately reflects our intended meaning. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All CT volumes used in this work, together with the 2D slices derived from them, were provided by our collaborating medical centers. The use of these data for research has been approved by the corresponding ethics review boards, and the required data-usage permissions were obtained prior to this study. All data are used exclusively for the research purposes described in this paper and are processed in accordance with the terms of the approvals.

To protect patient privacy, all data were de-identified by the source institutions before being accessed by the authors, and no patient-identifying information is disclosed in this paper or in the released artifacts. Model weights and code released with this paper do not contain any identifiable patient information. Any public release of data will be limited to a de-identified subset of the pre-training corpus and will be carried out in accordance with the relevant IRB protocols and data-sharing agreements with the collaborating centers.

REFERENCES

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.

Nathalie Alves, Merlijn Schuurmans, Dominik Rutkowski, Derya Yakar, Ingfrid Haldorsen, Marjolein Liedenbaum, Anders Molven, Phillip Vendittelli, Geert Litjens, Jurgen Hermans, et al. The panorama study protocol: Pancreatic cancer diagnosis-radiologists meet ai. *Zenodo*, 2024.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

Han Cai, Junyan Li, Muyan Hu, Chuang Gan, and Song Han. Efficientvit: Lightweight multi-scale attention for high-resolution dense prediction. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 17256–17267. IEEE, 2023.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pp. 205–218. Springer, 2022.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

Junren Chen, Rui Chen, Wei Wang, Junlong Cheng, Lei Zhang, and Liangyin Chen. Tinyu-net: Lighter yet better u-net with cascaded multi-receptive fields. In *International conference on medical image computing and computer-assisted intervention*, pp. 626–635. Springer, 2024.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Zekai Chen, Devansh Agarwal, Kshitij Aggarwal, Wiem Safta, Mariann Micsinai Balan, and Kevin Brown. Masked image modeling advances 3d medical image analysis. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1969–1979. IEEE, 2023.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pp. 424–432. Springer, 2016.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10575–10584. IEEE, 2020.

Xiaohan Ding, Yiyuan Zhang, Yixiao Ge, Sijie Zhao, Lin Song, Xiangyu Yue, and Ying Shan. Unireplknet: A universal perception large-kernel convnet for audio, video, point cloud, time-series and image recognition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5513–5524. IEEE, 2024.

Quentin Garrido, Nicolas Ballas, Mahmoud Assran, Adrien Bardes, Laurent Najman, Michael Rabbat, Emmanuel Dupoux, and Yann LeCun. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *arXiv preprint arXiv:2502.11831*, 2025.

Sergios Gatidis, Tobias Hepp, Marcel Früh, Christian La Fougère, Konstantin Nikolaou, Christina Pfannenberg, Bernhard Schölkopf, Thomas Küstner, Clemens Cyran, and Daniel Rubin. A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. *Scientific Data*, 9(1):601, 2022.

Xu Han, Yuan Tang, Zhaoxuan Wang, and Xianzhi Li. Mamba3d: Enhancing local features for 3d point cloud analysis via state space model. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 4995–5004, 2024.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pp. 272–284. Springer, 2021.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *2022 IEEE/CVF winter conference on applications of computer vision (WACV)*, pp. 1748–1758. Ieee, 2022.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9726–9735. IEEE, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Yuting He, Guanyu Yang, Jian Yang, Rongjun Ge, Youyong Kong, Xiaomei Zhu, Shaobo Zhang, Pengfei Shao, Huazhong Shu, Jean-Louis Dillenseger, et al. Meta grayscale adaptive network for 3d integrated renal structures segmentation. *Medical image analysis*, 71:102055, 2021.

Yuting He, Guanyu Yang, Rongjun Ge, Yang Chen, Jean-Louis Coatrieux, Boyu Wang, and Shuo Li. Geometric visual similarity learning in 3d medical image self-supervised pre-training. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9538–9547. IEEE, 2023.

Nicholas Heller, Fabian Isensee, Dasha Trofimova, Resha Tejpaul, Zhongchen Zhao, Huai Chen, Lisheng Wang, Alex Golts, Daniel Khapun, Daniel Shats, et al. The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct. *arXiv preprint arXiv:2307.01984*, 2023.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pp. 289–305. Springer, 2024.

Xiaohong Huang, Zhifang Deng, Dandan Li, and Xueguang Yuan. Missformer: An effective medical image segmentation transformer. *arXiv preprint arXiv:2109.07162*, 2021.

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

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.

A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical image analysis*, 69:101950, 2021.

Zoé Lambert, Caroline Petitjean, Bernard Dubray, and Su Kuan. Segthor: Segmentation of thoracic organs at risk in ct images. In *2020 Tenth international conference on image processing theory, tools and applications (IPTA)*, pp. 1–6. Ieee, 2020.

Bennett Landman, Zhoubing Xu, Juan Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge*, volume 5, pp. 12. Munich, Germany, 2015.

Hao Li, Di-Hua Zhai, and Yuanqing Xia. Erdunet: An efficient residual double-coding unet for medical image segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4):2083–2096, 2023.

Haoyue Li, Yifan Gao, Feng Yuan, Xiaosong Wang, and Xin Gao. Dino u-net: Exploiting high-fidelity dense features from foundation models for medical image segmentation. *arXiv preprint arXiv:2508.20909*, 2025a.

Wenxuan Li, Alan Yuille, and Zongwei Zhou. How well do supervised models transfer to 3d image segmentation. In *The Twelfth International Conference on Learning Representations*, volume 1, pp. 1, 2024.

Yuheng Li, Yizhou Wu, Yuxiang Lai, Mingzhe Hu, and Xiaofeng Yang. Meddinov3: How to adapt vision foundation models for medical image segmentation? *arXiv preprint arXiv:2509.02379*, 2025b.

Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 21152–21164, 2023.

Gongning Luo, Kuanquan Wang, Jun Liu, Shuo Li, Xinjie Liang, Xiangyu Li, Shaowei Gan, Wei Wang, Suyu Dong, Wenyi Wang, et al. Efficient automatic segmentation for multi-level pulmonary arteries: The parse challenge. *arXiv preprint arXiv:2304.03708*, 2023.

Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6695–6714, 2021.

Jun Ma, Yao Zhang, Song Gu, Xingle An, Zhihe Wang, Cheng Ge, Congcong Wang, Fan Zhang, Yu Wang, Yinan Xu, et al. Fast and low-gpu-memory abdomen ct organ segmentation: the flare challenge. *Medical Image Analysis*, 82:102616, 2022.

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

Mojtaba Masoudi, Hamid-Reza Pourreza, Mahdi Saadatmand-Tarzjan, Noushin Eftekhari, Fateme Shafiee Zargar, and Masoud Pezeshki Rad. A new dataset of computed-tomography angiography images for computer-aided detection of pulmonary embolism. *Scientific data*, 5(1): 180180, 2018.

Sachin Mehta and Mohammad Rastegari. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10): 1993–2024, 2014.

Yang Nan, Javier Del Ser, Zeyu Tang, Peng Tang, Xiaodan Xing, Yingying Fang, Francisco Herrera, Witold Pedrycz, Simon Walsh, and Guang Yang. Fuzzy attention neural network to tackle discontinuity in airway segmentation. *IEEE transactions on neural networks and learning systems*, 35(6):7391–7404, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Junting Pan, Adrian Bulat, Fuwen Tan, Xiatian Zhu, Lukasz Dudziak, Hongsheng Li, Georgios Tzimiropoulos, and Brais Martinez. Edgevits: Competing light-weight cnns on mobile devices with vision transformers. In *European conference on computer vision*, pp. 294–311. Springer, 2022.

Chongyu Qu, Tiezheng Zhang, Hualin Qiao, Yucheng Tang, Alan L Yuille, Zongwei Zhou, et al. Abdomenatlas-8k: Annotating 8,000 ct volumes for multi-organ segmentation in three weeks. *Advances in Neural Information Processing Systems*, 36:36620–36636, 2023.

Félix Quinton, Romain Popoff, Benoît Presles, Sarah Leclerc, Fabrice Meriaudeau, Guillaume Nodari, Olivier Lopez, Julie Pellegrinelli, Olivier Chevallier, Dominique Ginhac, et al. A tumour and liver automatic segmentation (atlas) dataset on contrast-enhanced magnetic resonance imaging for hepatocellular carcinoma. *Data*, 8(5):79, 2023.

Lukas Radl, Yuan Jin, Antonio Pepe, Jianning Li, Christina Gsaxner, Fen-hua Zhao, and Jan Egger. Avt: Multicenter aortic vessel tree cta dataset collection with ground truth segmentation masks. *Data in brief*, 40:107801, 2022.

Md Mostafijur Rahman, Mustafa Munir, and Radu Marculescu. Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11769–11779. IEEE, 2024.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Holger Roth, Amal Farag, Evrim B Turkbey, Le Lu, Jiamin Liu, and Ronald M Summers. Data from pancreas-ct. *(No Title)*, 2016.

Holger R Roth, Ziyue Xu, Carlos Tor-Díez, Ramon Sanchez Jacob, Jonathan Zember, Jose Molto, Wenqi Li, Sheng Xu, Baris Turkbey, Evrim Turkbey, et al. Rapid artificial intelligence solutions in a pandemic—the covid-19-20 lung ct lesion segmentation challenge. *Medical image analysis*, 82: 102605, 2022.

Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F Jaeger, and Klaus H Maier-Hein. Mednext: transformer-driven scaling of convnets for medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 405–415. Springer, 2023.

Jiacheng Ruan, Mingye Xie, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. Ege-unet: an efficient group enhanced unet for skin lesion segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 481–490. Springer, 2023.

Daniel Scholz, Ayhan Can Erdur, Viktoria Ehm, Anke Meyer-Baese, Jan C Peeken, Daniel Rueckert, and Benedikt Wiestler. Mm-dinov2: adapting foundation models for multi-modal medical image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 320–330. Springer, 2025.

Anjany Sekuboyina, Malek E Husseini, Amirhossein Bayat, Maximilian Löffler, Hans Liebl, Hongwei Li, Giles Tetteh, Jan Kukačka, Christian Payer, Darko Štern, et al. Verse: a vertebrae labelling and segmentation benchmark for multi-detector ct images. *Medical image analysis*, 73:102166, 2021.

Jun Shi. Structseg2019 gtv segmentation, 2023. URL https://dx.doi.org/10.21227/h75x-gt46.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Luc Soler, Alexandre Hostettler, Vincent Agnus, Arnaud Charnoz, Jean-Baptiste Fasquel, Johan Moreau, Anne-Blandine Osswald, Mourad Bouhadjar, and Jacques Marescaux. 3d image reconstruction for comparison of algorithm database. *URL: https://www. ircad. fr/research/data-sets/liver-segmentation-3d-ircadb-01*, 13:4, 2010.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Fenghe Tang, Lingtao Wang, Chunping Ning, Min Xian, and Jianrui Ding. Cmu-net: a strong convmixer-based medical ultrasound image segmentation network. In *2023 IEEE 20th international symposium on biomedical imaging (ISBI)*, pp. 1–5. IEEE, 2023.

Fenghe Tang, Jianrui Ding, Quan Quan, Lingtao Wang, Chunping Ning, and S Kevin Zhou. Cmunext: An efficient medical image segmentation network based on large kernel and skip fusion. In *2024 IEEE international symposium on biomedical imaging (ISBI)*, pp. 1–5. IEEE, 2024a.

Fenghe Tang, Bingkun Nian, Jianrui Ding, Wenxin Ma, Quan Quan, Chengqi Dong, Jie Yang, Wei Liu, and S Kevin Zhou. Mobile u-vit: revisiting large kernel and u-shaped vit for efficient medical image segmentation. In *Proceedings of the 33rd ACM international conference on multimedia*, pp. 3408–3417, 2025a.

Fenghe Tang, Bingkun Nian, Yingtai Li, Zihang Jiang, Jie Yang, Wei Liu, and S Kevin Zhou. Mambamim: Pre-training mamba with state space token interpolation and its application to medical image segmentation. *Medical Image Analysis*, 103:103606, 2025b.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20698–20708. IEEE, 2022.

Yuhao Tang, Haichen Yang, Liyan Zhang, and Ye Yuan. Work like a doctor: Unifying scan localizer and dynamic generator for automated computed tomography report generation. *Expert Systems with Applications*, 237:121442, 2024b.

Jeya Maria Jose Valanarasu and Vishal M Patel. Unext: Mlp-based rapid medical image segmentation network. In *International conference on medical image computing and computer-assisted intervention*, pp. 23–33. Springer, 2022.

Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 36–46. Springer, 2021.

Bram van Ginneken. Sliver07, March 2019. URL https://doi.org/10.5281/zenodo.2597908.

Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Rep vit: revisiting mobile cnn from vit perspective. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15909–15920. IEEE, 2024.

Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R Zaiane. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 2441–2449, 2022.

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

Biao Wu, Yutong Xie, Zeyu Zhang, Jinchao Ge, Kaspar Yaxley, Suzan Bahadir, Qi Wu, Yifan Liu, and Minh-Son To. Bhsd: A 3d multi-class brain hemorrhage segmentation dataset. In *International workshop on machine learning in medical imaging*, pp. 147–156. Springer, 2023.

Linshan Wu, Jiaxin Zhuang, and Hao Chen. Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22873–22882. IEEE, 2024.

Linshan Wu, Jiaxin Zhuang, and Hao Chen. Large-scale 3d medical image pre-training with geometric context priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Yutong Xie, Jianpeng Zhang, Yong Xia, and Qi 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.

Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical image analysis*, 80:102510, 2022.

Sicheng Yang, Hongqiu Wang, Zhaohu Xing, Sixiang Chen, and Lei Zhu. Segdino: An efficient design for medical and natural image segmentation with dino-v3. *arXiv preprint arXiv:2509.00833*, 2025.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pp. 543–553, 2023.

Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1195–1204. Ieee, 2021.

Kang Zhang, Xiaohong Liu, Jun Shen, Zhihuan Li, Ye Sang, Xingwang Wu, Yunfei Zha, Wenhua Liang, Chengdi Wang, Ke 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.

Yingtao Zhang, Min Xian, Heng-Da Cheng, Bryar Shareef, Jianrui Ding, Fei Xu, Kuan Huang, Boyu Zhang, Chunping Ning, and Ying Wang. Busis: a benchmark for breast ultrasound image segmentation. In *Healthcare*, volume 10, pp. 729. Mdpi, 2022.

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 Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8020–8035, 2023.

Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical image analysis*, 67:101840, 2021.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

Xiahai Zhuang. Multivariate mixture model for myocardial segmentation combining multi-source images. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):2933–2946, 2018.

A Supplementary Material

A.1 Open-Source Release Plan

This section details the artifacts that will be released alongside mDINO, giving the community three capabilities: verifying every reported result, reusing the encoders at zero pre-training cost, and extending the recipe to new corpora and tasks. Table 7 summarizes the release scope.

Table 7: Summary of artifacts to be released.

Artifact	Scope
Pre-training code and configurations	Full sequential 2D→3D pipeline
Pre-trained checkpoints	mDINO-2D and mDINO-3D, all trained variants
Fine-tuning and evaluation code	All 2D and 3D benchmarks reported
Per-benchmark configuration files	2D (Mobile U-ViT) and 3D (VoCo/MONAI) protocols
Fine-tuned weights	Every result reported in the paper

Released artifacts. Upon publication, we will release (i) the complete pre-training code and configurations covering the full sequential 2D→3D pipeline of mDINO-2D and mDINO-3D; (ii) all pre-trained mDINO-2D and mDINO-3D checkpoints; (iii) the downstream fine-tuning and evaluation code for every benchmark reported in the paper, together with the complete per-benchmark configuration files; and (iv) the fine-tuned weights used to produce every reported result. The downstream code will inherit the two public pipelines adopted in our experiments (Mobile U-ViT for 2D and VoCo/MONAI for 3D), so each reported entry can be re-evaluated under the exact held-out protocol of the corresponding benchmark.

What the release enables. The release is designed to cover both pre-training stages and every downstream result, and will support three capabilities. **(a) Verification:** every reported number can be independently re-evaluated with the released fine-tuned weights, evaluation scripts, and public benchmarks; no access to the pre-training corpus is required anywhere on the evaluation path. **(b) Reuse:** the encoders are standard ViTs that plug directly into existing task networks, so downstream users will be able to adopt mDINO with zero pre-training cost. **(c) Extension:** the full training recipe, provided as code and configurations, is corpus-agnostic: any group with access to institutional CT data will be able to apply it to train on their own corpora.

Data governance. The 1.13M-volume pre-training corpus was acquired under IRB approval and data-use agreements with 97 collaborating institutions; in line with clinical data-governance standards, the raw scans are not redistributable. The release is designed so that this constraint does not touch the verification path: everything needed to inspect, reuse, re-evaluate, and extend mDINO will be public.

Toward partial data release. Beyond artifacts, we are working with our collaborating centers to prepare a de-identified subset of the pre-training corpus for public release under IRB-compliant protocols, with patient identifiers removed and de-identification procedures thoroughly verified. This release is intended to make our pre-training results fully reproducible and to facilitate large-scale medical pre-training for the broader community.

A.2 Pre-training Data Scale and Performance Overview

The pre-training data scale comparison and performance overview are presented in Fig. 6. The 2D-3D unsupervised high-quality dataset is constructed as follows:

- CT-3D: Dataset aggregates 2.93M CT scans (including limited CTA acquisitions) acquired at 97 collaborating medical institutions across diverse regions in Asia between 2016 and 2024, covering heterogeneous scanner vendors, slice thicknesses, and reconstruction kernels, with scan ranges of 150–2000 mm across the thorax, abdomen–pelvis, and cranium. The dataset contains 1.13M clinical cases after rigorous quality control. Inclusion criteria are: (1) patients $\geq$12 years old to exclude developmental skeletal variations; (2) axial depth of 32–128 slices ensuring volumetric

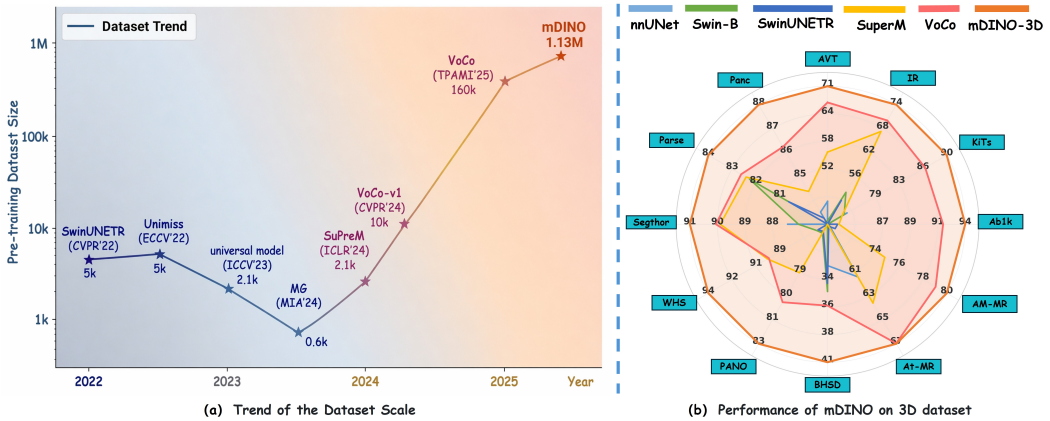


Figure 6: (a) Pre-training dataset scales across recent medical SSL works. mDINO is trained on a large-scale medical CT corpus (1.13M volumes). (b) Performance comparison on 2D/3D benchmarks across multiple modalities. Note: radar chart compresses differences for visualization; quantitative values are provided in Tables 1–4.

completeness; (3) slice thickness $\leq$2 mm for near-isotropic voxel resolution; (4) matrix size of 512$\times$512 pixels; (5) exclusion of scans with $>$30% invalid voxels (HU$\leq$-1024) to remove artifacts, motion blur, and out-of-field contamination. This multi-center dataset supports cross-anatomical deep learning pre-training.

- CT-2D: The 2D high-quality dataset is sampled from CT-3D and contains 113M CT slices. It comprises three clinical CT modalities with strict quality criteria: (1) Chest CT—non-contrast volumetric chest scans covering pulmonary parenchyma, mediastinal structures, and thoracic wall soft tissues; (2) Lumbar Spine CT—imaging sequences encompassing L1–L5 vertebrae, intervertebral discs, paravertebral musculature, and adjacent soft tissues; (3) Brain CT—cranial volumes including the cerebral cortex and white matter, ventricular system, basal ganglia, and pathological lesions. All sampled cases satisfy the following inclusion criteria: (i) slice thickness $\leq$10 mm, (ii) matrix size of 512$\times$512 pixels, and (iii) axial slice count $\geq$16, ensuring adequate spatial resolution and volumetric continuity for clinical interpretation.

Benefiting from large-scale cross-dimensional pre-training, mDINO achieves state-of-the-art performance across more than 40 downstream benchmarks spanning both 2D and 3D datasets. These benchmarks cover multiple imaging modalities, including ultrasound, MRI, CT, and dermoscopy, as well as diverse tasks such as segmentation and radiology report generation. For clarity of presentation, representative results are reported in the main text.

Leakage and De-duplication Analysis. The pre-training corpus is separated from all evaluation benchmarks at three levels. **(1) Provenance.** All 1.13M volumes were collected directly from the 97 collaborating institutions and were never sourced from any public benchmark repository; the corpus and the evaluation sets are therefore independent collections by construction. **(2) Objective.** mDINO is trained fully self-supervised: the corpus carries no segmentation masks, diagnostic labels, or any downstream annotation whatsoever, so label-level leakage is structurally absent. **(3) Data.** Hash-based de-duplication was performed against every public benchmark used for evaluation—BTCV, AMOS, TotalSegmentator, KiTS, and all MSD tasks—at the image, study, and patient level, removing exact duplicates between the corpus and the evaluation sets. Beyond these measures, the evaluation design itself argues against contamination-driven inflation: the downstream suite spans ultrasound, MRI, and dermoscopy benchmarks that share no modality with the CT-only corpus and hence cannot overlap with it, and improvements are consistent across all 40+ benchmarks—the pattern expected of a general-purpose representation, whereas contamination typically inflates a few specific leaked datasets. We note that, as with any independently collected clinical corpus, exact-hash matching cannot provably exclude every near-duplicate across institutions; the measures above follow standard practice for large-scale clinical data.

Table 8: Downstream datasets. All datasets are unseen during pre-training.

Dataset	Modality	Task
2D Segmentation		
BUS(2022)	US	Breast Seg.
BUSI(2020)	US	Breast Seg.
TNSCUI[1]	US	Thyroid Seg.
ISIC2018(2019)	Derm	Skin Lesion Seg.
3D Segmentation		
BTCV(2015)	CT	Abd. Seg.
AMOS22(2022)	CT	Abd. Seg.
FLARE22(2022)	CT	Abd. Seg.
Abdomen1k(2021)	CT	Abd. Seg.
AbdomenAtlas(2023)	CT	Abd. Seg.
TotalSegmentator(2023)	CT	Multi-org. Seg.
MM-WHS(2018)	CT	Heart Seg.
AVT(2022)	CT	Aorta Seg.
CHAOS(2021)	CT	Liver Seg.
Sliver07(2019)	CT	Liver Seg.
IRCADb(2010)	CT	Liver Tumor Seg.
KiTS(2023)	CT	Kidney Tumor Seg.
KiPA22(2021)	CT	Kidney Tumor Seg.
TCIA-Panc.(2016)	CT	Panc. Seg.
PANORAMA(2024)	CT	Panc. Tumor Seg.
Segthor(2020)	CT	Thoracic Risk Seg.
BHSD(2023)	CT	Brain Hem. Seg.
StructSeg19(2023)	CT	NPC Seg.
Covid-19-20(2022)	CT	COVID Seg.
FUMPE(2018)	CT	PE Seg.
Parse22(2023)	CT	PA Seg.
AIIB23(2023)	CT	Fibrotic Lung Seg.
CC-CCII(2020)	CT	COVID Cls.
AMOS-MRI(2022)	MRI	Abd. Seg.
MM-WHS-MRI(2018)	MRI	Heart Seg.
ATLAS-MRI(2023)	MRI	Liver Tumor Seg.
BraTs21(2014)	MRI	Brain Tumor Seg.
MSD(2022)		
Task01 Heart	MRI	Heart Seg.
Task02 Heart	MRI	Heart Seg.
Task03 Liver	CT	Liver Tumor Seg.
Task04 Hip	MRI	Hip Seg.
Task05 Pros.	MRI	Pros. Seg.
Task06 Lung	CT	Lung Cancer Seg.
Task07 Panc.	CT	Panc. Tumor Seg.
Task08 Vessel	CT	Vessel Tumor Seg.
Task09 Spleen	CT	Spleen Seg.
Task10 Colon	CT	Colon Cancer Seg.
Report Generation		
CTRG-Chest	VLP	Report Gen.

Ethics Statement. The pre-training corpus was aggregated from multiple collaborating medical institutions under institutional review board (IRB) approval, with all data de-identified according to HIPAA guidelines prior to aggregation. Patient consent was obtained following institutional protocols, and no personally identifiable information was retained in the final dataset. The complete open-source release plan (code, configurations, checkpoints, and fine-tuned weights) is detailed in Appendix A.1.

A.3 Downstream Datasets

Downstream evaluations are conducted on four medical tasks, including 2D segmentation, 3D segmentation, and 3D report generation. The evaluation benchmarks are summarized in Tab. 8. Cross-modality generalization is assessed on ultrasound and MRI modalities. All downstream datasets remain unseen during pre-training.

A.4 Data-Driven Priors of CT Volumes

The design of mDINO is purpose-built for the statistical properties of CT volumes rather than a direct application of generic SSL techniques: each component of the training pipeline and architecture is motivated by a quantitatively measured CT prior and verified by a dedicated ablation. To ground this rationale rather than leave it as a post-hoc claim, this section directly quantifies the CT statistics motivating each design, and the strong effectiveness of the resulting foundation model across more than 40 downstream benchmarks (Tables 1–4 of the main text) provides direct empirical evidence for the value of this framework.

On a mixture of volumes sampled from BTCV and AMOS-CT, we construct the centered patch-feature matrix $X = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n \times d}$ with $X = U\Sigma V^\top$, where x_i is the intensity–gradient descriptor of the i-th 3D patch, and measure the effective rank $r_{95} = \min\{r : \sum_{i=1}^{r} \sigma_i^2 / \sum_i \sigma_i^2 \geq 0.95\}$, finding only $12.4\%/13.7\%$ of dimensions explain 95% of the energy. We further measure the energy sparsity $k_{90} = \min_S\{|S|/n : \sum_{i \in S} \|x_i\|_2^2 / \sum_i \|x_i\|_2^2 \geq 0.90\}$, finding only $18.6\%/21.3\%$ of patches (ranked by energy) account for 90% of it, while six-neighbor feature similarity is high ($0.82/0.80$) versus random-neighbor similarity ($0.19/0.17$)— informative CT features are sparse but locally clustered.

These priors motivate mDINO directly: sparse, locally clustered features call for masked self-distillation over compact 3D neighborhoods rather than fragmented tokens; the low-rank spectrum, consistent across slices and volumes, supports sequential 2D→3D pre-training to estimate the dominant anatomical subspace before full volumetric optimization; and near-identical correlation decay across the axial, coronal, and sagittal axes supports a shared-frequency 3D positional encoding over axis-specific video-style encoding. Ablations verify these choices (Tab. 6 of the main text): 2D→3D initialization reduces the pre-training loss from 14.68 to 9.62 and lifts BTCV DSC from 82.31% to 85.32%; Hilbert flatten outperforms raster/Morton serialization; and shared-frequency 3D RoPE outperforms both 2D RoPE and axis-specific 3D RoPE.

A.5 Training Details

All experiments use mixed precision training with bfloat16 for model parameters and fp32 for gradient reduction. Fully Sharded Data Parallel (FSDP) with SHARD_GRAD_OP is adopted for distributed training. Model compilation is enabled to improve runtime efficiency, while CUDA graphs are disabled. Sinkhorn-Knopp centering is applied to stabilize teacher outputs. All ablation experiments use a fixed random seed to ensure reproducibility.

Pre-training config

Pre-training

Both stages follow the DINOv3 self-distillation setup: a ViT student learns from an EMA-updated teacher under combined DINO and iBOT losses with KoLeo regularization. Stage 2 differs from Stage 1 in the positional encoding (isotropy 3D-medical-RoPE), the student initialization (from the mDINO-2D checkpoint), and the volumetric crops. All configurations are listed in Tab. 9.

Downstream Fine-tuning config

All downstream settings strictly inherit the two established public pipelines adopted in our experiments: 2D benchmarks use Mobile U-ViT (2025a) and 3D benchmarks use VoCo/MONAI (2025). The exact values are restated in Tab. 10; per-benchmark differences (e.g., number of classes, in-plane spacing, foreground sampling ratio) are inherited unchanged from the two repositories, and the complete per-benchmark configuration files are included in the release plan (Appendix A.1).

A.6 Detailed Derivation of 3D RoPE

For completeness, we provide the derivation of extending RoPE (2024) to three spatial dimensions. The extension follows the standard framework; the domain-specific design choice is the use of shared frequency parameters across all three axes, motivated by the near-isotropic resolution of resampled medical CT volumes, in contrast to video Transformers that typically require distinct temporal and spatial frequencies.

Table 9: Pre-training hyperparameters for the two stages.

Setting	**Stage 1 (2D)**	**Stage 2 (3D)**
Backbone	ViT-S, patch 16, FFN 4.0	ViT-S, patch 16, FFN 4.0
Stoch. depth / LayerScale	0.3 / 1×10^{-5}	0.3 / 1×10^{-5}
Positional encoding	RoPE (base 100)	isotropy 3D-medical-RoPE
Student initialization	from scratch	mDINO-2D checkpoint
EMA momentum	$0.992 \rightarrow 1.0$ (cosine)	$0.992 \rightarrow 1.0$ (cosine)
Teacher temperature	$0.04 \rightarrow 0.07$ (10 ep.)	$0.04 \rightarrow 0.07$ (10 ep.)
Objective	DINO + iBOT + KoLeo (0.1)	DINO + iBOT + KoLeo (0.1)
Projection heads	65,536 proto., bn 256, MLP 2048	65,536 proto., bn 256, MLP 2048
iBOT masking	prob. 0.5, ratio $[0.1, 0.5]$	prob. 0.5, ratio $[0.1, 0.5]$
Optimizer	AdamW (β_1=0.9, β_2=0.999)	AdamW (β_1=0.9, β_2=0.999)
Learning rate	2×10^{-3}, cosine $\rightarrow 1 \times 10^{-6}$	2×10^{-3}, cosine $\rightarrow 1 \times 10^{-6}$
Warm-up epochs	10	10
Weight decay	$0.05 \rightarrow 0.2$	$0.05 \rightarrow 0.2$
Grad. clip / layer-wise decay	3.0 / 0.9	3.0 / 0.9
Patch-embed lr multiplier	0.2	0.2
Global crops	scale $[0.7, 1.0]$	scale $[0.7, 1.0]$
Local crops	scale $[0.3, 0.7]$	scale $[0.3, 0.7]$
Other augmentation	h-flip, color jitter	pairing ratio 1.0
Batch size per GPU	144	64

Table 10: Downstream fine-tuning hyperparameters, inherited from Mobile U-ViT (2D) and VoCo/MONAI (3D).

Setting	**2D (Mobile U-ViT)**	**3D (VoCo/MONAI)**
Optimizer	SGD (momentum 0.9)	AdamW
Learning rate	1×10^{-2}	3×10^{-4}
LR schedule	Poly $(1 - t/T)^{0.9}$	Linear-warmup cosine (20 ep.)
Weight decay	1×10^{-4}	1×10^{-5}
Loss	0.5 BCE + Dice	DiceCE
Input / ROI	256×256	$96 \times 96 \times 96$
Preprocessing	Flip/Rot90, Normalize	Spacing 1.5^3, clip $[-175, 250] \rightarrow [0, 1]$
Inference	Full-image	Sliding window, overlap 0.75
Seed	Fixed	Fixed

Both 1D RoPE and 2D RoPE are special cases of the following general formulation (2024):

$$R = \exp\Big(\sum_{i=1}^{N} x_i B_i\Big), \tag{4}$$

where x_i denotes the spatial coordinate along the i-th axis, and B_i is the generator matrix associated with that axis. And a valid set of matrices $\{B_i\}$ must make sure that R satisfies the following relativity and reversibility properties:

$$R_{x_1}^T R_{x_2} = R_{x_2 - x_1}, \tag{5}$$

$$R_{x_1} = R_{x_2} \Rightarrow x_1 = x_2, \tag{6}$$

where x_1 and x_2 are two absolute positions in N-dimensional Euclidean space representing the positions of a query and a key, respectively.

For 1D position, a valid choice of B is:

$$B = \theta \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \tag{7}$$

and its corresponding matrix exponential gives the standard 1D RoPE formulation:

$$R_m = \exp(mB) = \begin{pmatrix} \cos m\theta & -\sin m\theta \\ \sin m\theta & \cos m\theta \end{pmatrix}, \tag{8}$$

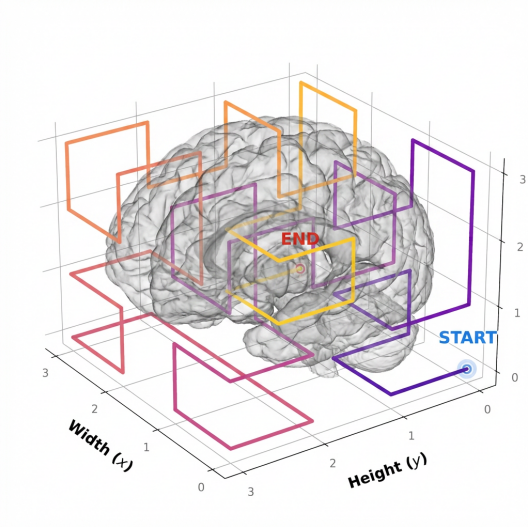


Figure 7: Visualization of the 3D Hilbert traversal on a $4 \times 4 \times 4$ voxel grid. The color gradient (blue to red) represents the temporal order of traversal. The path moves strictly between neighboring voxels, preserving geometric continuity and preventing fragmentation of local features during serialization.

where m is the absolute position of each token.

To extend this formulation to 3D medical imaging, we solve:

$$R_{x,y,z} = \exp(xB_1 + yB_2 + zB_3). \tag{9}$$

We construct block-diagonal generator matrices: $B_1 = \begin{pmatrix} J & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, $B_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & J & 0 \\ 0 & 0 & 0 \end{pmatrix}$, $B_3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & J \end{pmatrix}$, where $J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$.

The resulting 3D RoPE formulation is:

$$R_{x,y,z} = \exp(xB_1 + yB_2 + zB_3) = \mathrm{diag}(R(x),\, R(y),\, R(z)), \tag{10}$$

where $R(t) = \begin{pmatrix} \cos t\theta & -\sin t\theta \\ \sin t\theta & \cos t\theta \end{pmatrix}$.

Proof of properties. Since B_1, B_2, B_3 are block-diagonal with non-overlapping support, they commute pairwise ($B_iB_j = B_jB_i$). Therefore $\exp(xB_1+yB_2+zB_3) = \exp(xB_1)\exp(yB_2)\exp(zB_3)$. Each factor is an independent rotation matrix satisfying both relativity and reversibility, and the Kronecker product of valid RoPE matrices preserves both properties. This completes the proof.

A.7 Visualization of Hilbert Flatten

As shown in Fig. 7, the 3D Hilbert traversal is illustrated on a $4 \times 4 \times 4$ voxel grid. Each voxel is visited exactly once, and the color gradient (blue to red) encodes the temporal order of traversal. Unlike raster scanning, the Hilbert curve moves strictly between neighboring voxels, ensuring that spatially adjacent elements in the 3D volume remain proximate in the serialized token sequence. This property preserves geometric continuity and prevents fragmentation of local anatomical structures during flattening, supporting more effective spatial reasoning in volumetric representation learning.

A.8 PCA Visualization of mDINO

PCA Feature Visualization. Patch tokens from the final ViT encoder layer are projected onto the first three principal components computed across all tokens within each image, and the resulting

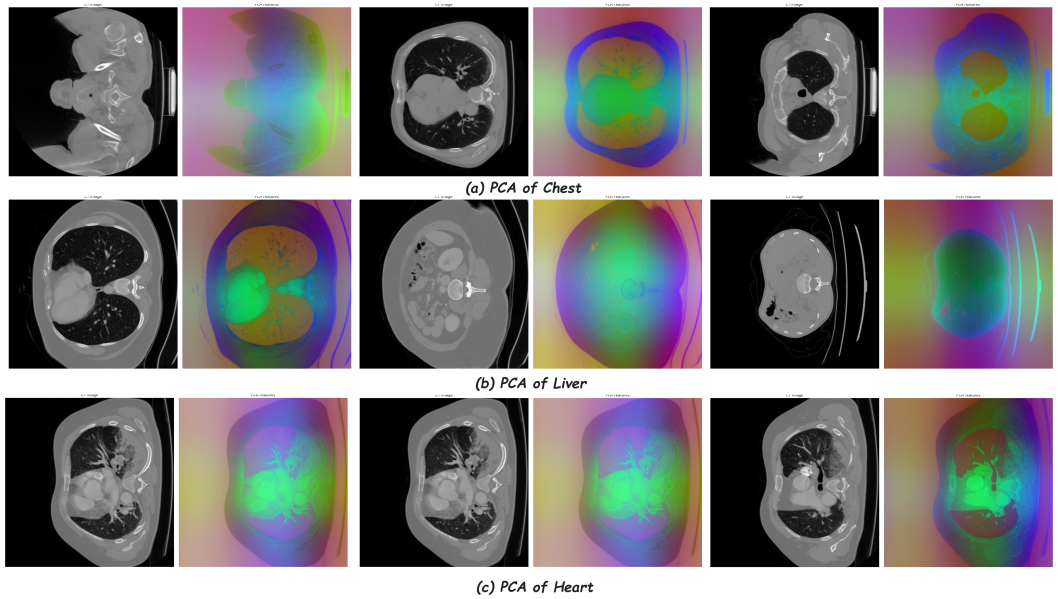
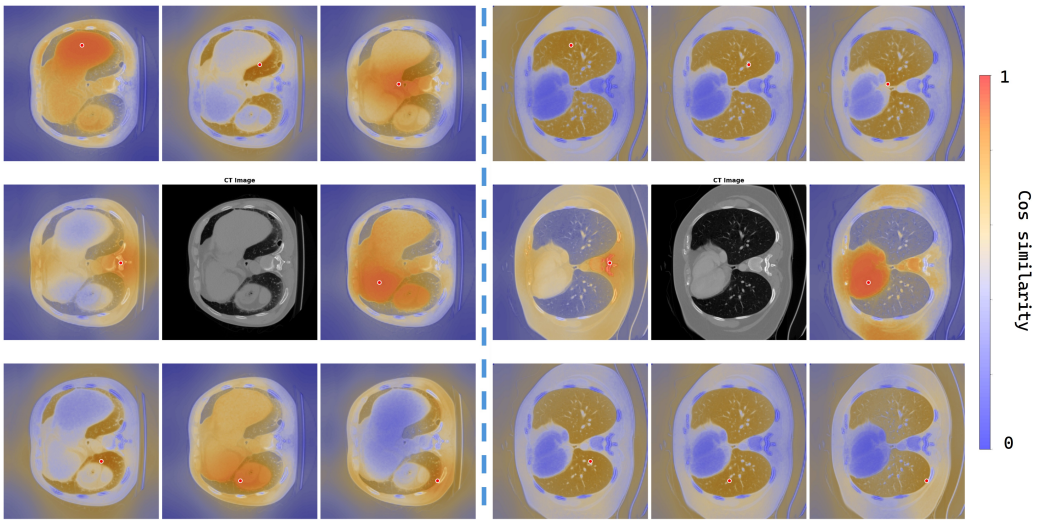


Figure 8: PCA visualization of patch-level features from the mDINO encoder. The first three principal components are mapped to RGB values. Semantically consistent regions exhibit coherent color patterns, while organ boundaries correspond to clear feature transitions.



(a) High-resolution dense features of mdino. We visualize the cosine similarity maps between the patches marked with a red dot and all other patches. Input image at 2048 × 2048.

Figure 9: Dense cosine similarity maps of mDINO on chest CT. Red dots indicate query patch locations. High similarity responses are concentrated within the same anatomical structure.

vectors are mapped to RGB values, which is shown in Fig. 8. Regions belonging to the same anatomical structure exhibit consistent color patterns, while organ boundaries correspond to clear feature transitions. Within homogeneous regions, the projected features change smoothly, indicating that the representation space respects the spatial continuity of anatomical structures. Background regions exhibit relatively uniform responses, whereas clinically relevant structures display richer feature variation, suggesting that the model implicitly emphasizes anatomically informative content.

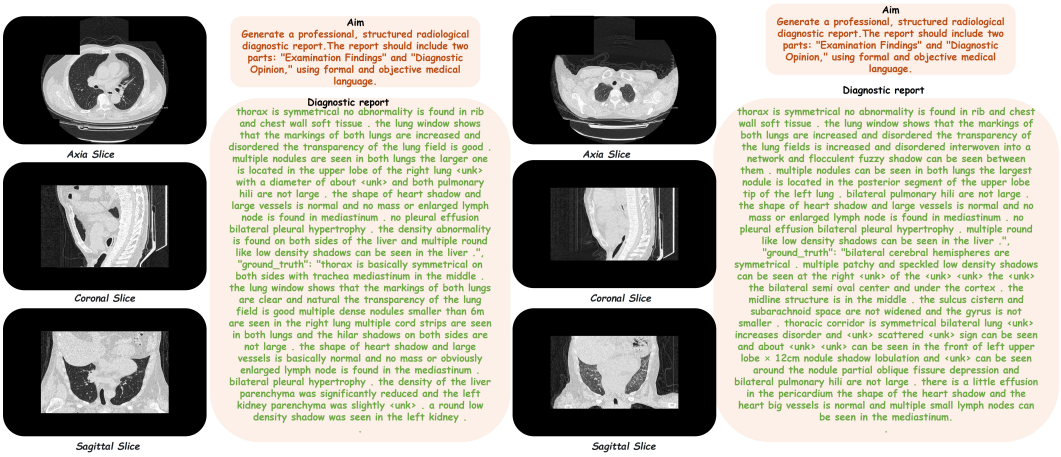


Figure 10: Qualitative examples of radiology report generation on chest CT. Multi-view inputs (axial, coronal, sagittal) are used to generate structured diagnostic reports.

A.9 Cosine Similarity Visualization of mDINO

Dense cosine similarity maps are computed between a selected query patch token and all other patch tokens in the image. Given patch features $\mathbf{f}_i \in \mathbb{R}^d$ extracted from the final encoder layer and ℓ_2-normalized, cosine similarity is:

$$s_{ij} = \mathbf{f}_i^\top \mathbf{f}_j. \tag{11}$$

Similarity scores are reshaped to the spatial patch grid, upsampled to the original resolution, and normalized via sigmoid centering for visualization.

As shown in Fig. 9, patches within the same anatomical structure exhibit consistently high similarity, while clear transitions occur at organ boundaries. Internal structures such as vessels and bronchi remain distinguishable, indicating that the representations capture both global organ-level semantics and fine structural details. These patterns confirm that mDINO organizes medical images according to anatomically meaningful spatial structure without explicit supervision.

A.10 Qualitative Analysis of Report Generation

Qualitative Analysis on Radiology Report Generation. Given multi-view CT slices (axial, coronal, and sagittal), the model generates structured diagnostic reports comprising Examination Findings and Diagnostic Opinion. As illustrated in Fig. 10, the generated reports correctly identify key anatomical structures and clinically relevant findings including pulmonary nodules, increased lung markings, and low-density shadows. The descriptions remain anatomically coherent across views, indicating that the mDINO representations preserve cross-view structural information and encode implicit volumetric context. These results demonstrate that the learned representations contain sufficient semantic information to support downstream language generation and high-level medical reasoning.