ANYCT-Ω: SCALING CROSS-VIEW REPRESENTATIONS FOR HETEROGENEOUS SPARSE-VIEW CT RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse-view computed tomography (CT) reconstruction aims to recover a volumetric attenuation field from a small number of X-ray projections. Existing learning-based methods are often optimized for a single dataset or anatomical domain, which limits their robustness under changes in anatomy, preprocessing, and acquisition statistics. We present AnyCT-Ω, a unified framework for heterogeneous sparse-view CT reconstruction that adapts transferable cross-view representations from a pretrained VGGT-Ω backbone to projection-based medical imaging. The model uses known CT geometry to align multi-view features with arbitrary three-dimensional query points and employs a content-routed residual mixture-of-experts head to adapt dense features across heterogeneous inputs. We further construct AnyCT-70K, a standardized corpus containing 73,610 processed samples from seven public CT sources covering thoracic, abdominal, pelvic, and musculoskeletal anatomy. Under a common joint-training protocol, AnyCT-Ω outperforms the strongest baseline on the mixed AnyCT-70K benchmark by 1.50, 1.72, and 1.64 dB PSNR at 6, 8, and 10 views, respectively. The corresponding SSIM improvements are 3.76, 3.63, and 3.06 points. Additional experiments show positive transfer when increasing the number of training sources and demonstrate that the proposed model reaches competitive reconstruction quality with 30,000 optimization steps. These results suggest that geometry-guided adaptation of transferable cross-view representations is a promising direction for heterogeneous sparse-view CT reconstruction.

1 INTRODUCTION

Computed tomography (CT) is an important three-dimensional imaging modality for clinical diagnosis, intervention planning, and disease screening. Conventional CT reconstruction typically relies on densely sampled X-ray projections. Sparse-view CT and cone-beam CT (CBCT) aim to reduce the number of acquired projections, thereby lowering radiation exposure and acquisition time. However, severe angular undersampling makes the reconstruction problem highly ill-posed: analytic methods such as filtered back-projection and FDK reconstruction can produce strong streak artifacts, while learning-based methods must infer missing anatomical structures from limited projection evidence(Feldkamp et al., 1984; Andersen & Kak, 1984).

Deep learning has substantially advanced sparse-view CT reconstruction. Early approaches either learn to post-process analytical reconstructions or directly learn inverse mappings from measurements to images(Jin et al., 2017). More recent methods represent the target volume as a continuous intensity or attenuation field and predict arbitrary three-dimensional points from sparse projection images(Lin et al., 2023; Zha et al., 2022). Cross-view and cross-regional architectures further improve CBCT reconstruction by explicitly aggregating projection features across views and anatomical regions(Lin et al., 2024b;a). Nevertheless, most existing methods are trained and evaluated on a single dataset or a limited number of anatomical domains. Their performance can therefore be sensitive to changes in anatomy, intensity normalization, preprocessing, and source-specific acquisition statistics. DeepSparse demonstrates the potential of foundation-style pretraining for sparse-view CBCT reconstruction(Lin et al., 2026), but large-scale joint training across heterogeneous CT sources remains insufficiently studied.

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.

In parallel, foundation models for multi-view three-dimensional vision have learned transferable geometric representations from large-scale natural-image data. DUSt3R, VGGT, and AnySplat demonstrate that multi-view pretraining can produce useful cross-view correspondences and geometry-aware features(Wang et al., 2024; 2025; Jiang et al., 2025). Dense prediction transformers further provide a general interface for converting intermediate transformer tokens into spatially aligned feature maps(Ranftl et al., 2021). These developments raise an important question: can transferable cross-view representations learned from natural images benefit projection-based medical reconstruction?

The transfer is non-trivial because natural-image reconstruction and CT reconstruction involve fundamentally different image-formation processes. Natural-image models are primarily developed for perspective projection and visible-surface geometry, whereas CT measurements are X-ray line integrals through a transparent and overlapping volume. The target of CT reconstruction is therefore a dense internal attenuation field rather than an explicitly visible surface. Directly interpreting natural-image depth or point predictions as CT attenuation is inappropriate. Instead, the pretrained representation must be aligned with the known CT acquisition geometry and adapted to point-wise attenuation prediction.

We address this problem with AnyCT-Ω, a unified framework for heterogeneous sparse-view CT reconstruction. AnyCT-Ω transfers cross-view representations from a pretrained VGGT-Ω backbone(Wang et al., 2025; 2026) and uses known CT geometry to project each three-dimensional query point onto the detector planes of all input views. Dense features and intermediate tokens are sampled at the projected locations, aggregated across views, and decoded into the normalized attenuation value of the query point. This geometry-guided formulation preserves the transferable cross-view representation while replacing the original natural-image geometric interpretation with a CT-specific reconstruction mechanism.

To handle distributional heterogeneity, we introduce a content-routed residual mixture-of-experts head(Shazeer et al., 2017). A shared predictor captures source-invariant reconstruction structure, while residual experts provide complementary corrections for pointwise, local, dilated, and high-frequency features. The routing weights are predicted from the input content rather than the dataset identity, allowing a single model to adapt to different anatomical regions and source statistics without using dataset-specific prediction heads.

We further construct AnyCT-70K, a standardized heterogeneous benchmark containing 73,610 processed samples from seven public CT sources. The benchmark covers thoracic, abdominal, pelvic, and musculoskeletal anatomy under a common sparse-view reconstruction protocol. On the mixed-data evaluation, AnyCT-Ω outperforms the strongest completed baseline under 6, 8, and 10 view settings. Data-scaling experiments further show that increasing the number of training sources can provide positive transfer rather than necessarily causing negative interference.

Our contributions are threefold:

- We introduce AnyCT-Ω and construct AnyCT-70K, a unified framework and standardized seven-source benchmark for heterogeneous sparse-view CT reconstruction. AnyCT-Ω adapts pretrained cross-view representations through known-geometry feature sampling and point-wise attenuation prediction, while AnyCT-70K contains 73,610 processed reconstruction samples and 630 held-out test samples across multiple anatomical regions.
- Building on the pretrained VGGT-Ω encoder and its inherited DPT dense transformation, we develop a CT-specific adaptation that combines known-geometry feature sampling, multi-level raw-map fusion, a PointDecayDecoder, and a content-routed residual MoE for heterogeneous inputs.
- Extensive experiments demonstrate that AnyCT-Ω performs strongly under heterogeneous joint training and consistently benefits from increasing the diversity of training sources.

2 Related Work

2.1 Sparse-View CT Reconstruction

Classical CT reconstruction inverts the Radon transform from line-integral measurements. Filtered back-projection provides an efficient analytical solution under sufficiently sampled projections,

whereas algebraic reconstruction methods such as SART iteratively enforce data consistency(Slaney & Kak, 1988; Andersen & Kak, 1984). For cone-beam CT, the Feldkamp-Davis-Kress (FDK) algorithm provides an efficient approximate reconstruction method, but its quality deteriorates substantially under severe angular undersampling(Feldkamp et al., 1984).

Recent learning-based approaches compensate for missing measurements using data-driven priors. Representative methods either post-process analytical reconstructions(Jin et al., 2017), unroll iterative optimization into trainable networks(Adler & Oktem, 2018), or represent the target volume as a continuous intensity or attenuation field(Zha et al., 2022; Lin et al., 2023). More recent feed-forward methods, including DIF-Net, C^2RV, and DIF-Gaussian, aggregate projection-conditioned features across views to enable point-wise volumetric prediction(Lin et al., 2023; 2024b;a). However, these methods are generally trained for a single source distribution or a small number of anatomical datasets. The effect of jointly training one reconstruction model across heterogeneous CT sources remains insufficiently explored.

2.2 Foundation Models for Cross-View Reconstruction

Large-scale pretraining has produced transferable representations for medical image analysis and multi-view 3D vision. Medical foundation models such as Models Genesis, Med3D, and Swin UNETR demonstrate that large-scale self-supervised or supervised learning can improve volumetric feature representations(Zhou et al., 2019; Chen et al., 2019; Tang et al., 2022). These models, however, are primarily designed for segmentation, classification, or general volumetric understanding rather than projection-based inverse problems.

In natural-image 3D vision, DPT converts intermediate transformer tokens into dense spatial predictions(Ranftl et al., 2021), while DUSt3R, MASt3R, and VGGT demonstrate that multi-view training can learn transferable geometric representations(Wang et al., 2024; Leroy et al., 2024; Wang et al., 2025). VGGT-Ω further scales this paradigm through stronger geometry-aware representations(Wang et al., 2026), and AnySplat applies multi-view features to feed-forward three-dimensional scene representation(Jiang et al., 2025). Unlike natural-image reconstruction, CT image formation is governed by X-ray line integrals through an overlapping volume rather than perspective projection of visible surfaces(Slaney & Kak, 1988; Feldkamp et al., 1984). Therefore, our method retains pretrained cross-view token representations while using known CT geometry to align features for attenuation-field prediction.

2.3 Mixture-of-Experts for Heterogeneous Sources

Mixture-of-experts (MoE) models increase capacity by combining a shared model with specialized experts selected by an input-dependent router(Shazeer et al., 2017). Sparse routing has enabled efficient scaling in large language and vision models, including Switch Transformers and V-MoE(Fedus et al., 2022; Riquelme et al., 2021). Soft routing further allows multiple experts to contribute to each prediction, avoiding hard expert assignment(Puigcerver et al., 2024).

Our MoE differs from conventional sparse conditional computation. It is applied only to the output of the dense feature decoder, and all four experts are evaluated for every input. A shared predictor models source-invariant structure, while pointwise, local, dilated, and high-frequency residual experts provide complementary corrections. The router combines local dense-map evidence with a global descriptor from cross-view tokens, enabling content-dependent specialization without dataset-specific heads or source identifiers.

Taken together, prior work provides strong foundations for analytical reconstruction, sparse-view neural fields, transferable geometric representations, and conditional specialization(Kudo et al., 2013; Lin et al., 2023; Wang et al., 2025; Shazeer et al., 2017), but their combination for large-scale heterogeneous sparse-view CT reconstruction remains largely unexplored.

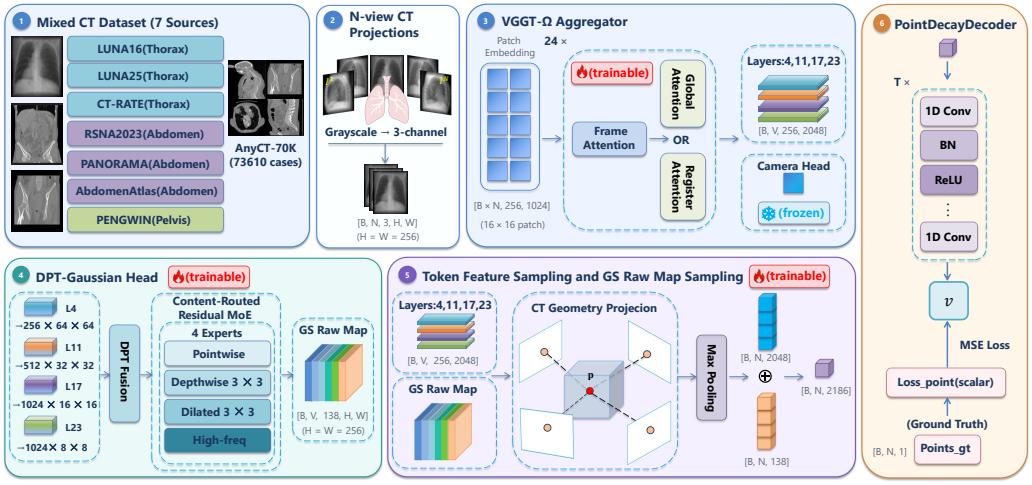


Figure 1: Overview of the AnyCT-Ω pipeline for geometry-guided sparse-view CT reconstruction. The pretrained VGGT-Ω encoder extracts multi-level cross view tokens, while the inherited DPT-Gaussian head produces dense projection-aligned features. A content-routed residual MoE adapts the dense features to heterogeneous CT inputs. Known CT geometry projects each 3D query point onto all detector views; the sampled token and dense features are max-pooled, fused, and decoded by the PointDecayDecoder into a normalized attenuation value.

3 METHODOLOGY

3.1 PROBLEM FORMULATION AND FRAMEWORK OVERVIEW

We formulate sparse-view cone-beam CT reconstruction as the recovery of a normalized volumetric attenuation field from a set of projection images. Let $\Omega \subset \mathbb{R}^3$ denote the normalized volume domain and $\mu : \Omega \to [0, 1]$ denote the target attenuation field. Given V projection images $\mathcal{P} = \{P_v\}_{v=1}^{V}$, acquisition angles $\{\theta_v\}_{v=1}^{V}$, and known CT geometry $\mathcal{G} = \{G_v\}_{v=1}^{V}$, our goal is to predict the attenuation at an arbitrary query point $\mathbf{x} \in \Omega$:

$$\hat{\mu}(\mathbf{x}) = f_\Theta \left(\mathbf{x}, \{P_v, \theta_v, G_v\}_{v=1}^{V}\right). \tag{1}$$

AnyCT-Ω combines a pretrained VGGT-Ω cross-view encoder(Wang et al., 2025; 2026), an inherited DPT-based dense feature transformation(Ranftl et al., 2021), a content-routed residual mixture-of-experts (MoE) module, and a point-wise attenuation decoder. As illustrated in Fig.1, VGGT-Ω extracts multi-level cross-view tokens from sparse CT projections. The inherited DPT-Gaussian head produces dense projection-aligned features, which are adapted by the proposed MoE. Known CT geometry maps each 3D query point to all detector views; sampled token and dense features are then pooled, fused, and decoded by the PointDecayDecoder into a normalized attenuation value.

3.2 VGGT-Ω CROSS-VIEW ENCODER

Each single-channel CT projection is replicated to three channels to match the pretrained patch embedding:$\tilde{P}_v = \mathrm{Replicate}(P_v, 3)$. The multi-view projections are jointly processed by the VGGT-Ω aggregator:

$$\left\{Z_v^\ell\right\}_{v=1}^{V} = E_\Omega^\ell \left(\tilde{P}_1, \ldots, \tilde{P}_V\right), \qquad \ell \in \mathcal{L}, \tag{2}$$

where $\mathcal{L} = \{4, 11, 17, 23\}$ denotes the selected intermediate layers and $Z_v^\ell \in \mathbb{R}^{N_p \times C_t}$ is the patch-token representation at layer ℓ for view v. The aggregator alternates within-view and inter-view attention, allowing each token to encode local projection evidence and cross-view context. The original VGGT-Ω camera head is retained only for architectural compatibility; all CT point projections use the known geometry $\{G_v\}$.

3.3 INHERITED DENSE FEATURE TRANSFORMATION

The selected token layers are reshaped into spatial maps, normalized, projected, resized, and fused by the inherited DPT transformation:

$$X_v^{\mathrm{DPT}} = U^{\mathrm{DPT}}\left(R_4(Z_v^4), R_{11}(Z_v^{11}), R_{17}(Z_v^{17}), R_{23}(Z_v^{23})\right), \tag{3}$$

where R_ℓ denotes layer-specific reshaping and projection operations. The fused feature is upsampled to detector resolution and combined with local appearance cues from the input projection:

$$X_v = \mathrm{Upsample}(X_v^{\mathrm{DPT}}) + \phi_{\mathrm{img}}(\tilde{P}_v), \tag{4}$$

where ϕ_{img} is a shallow 7×7 convolution followed by ReLU. This DPT transformation serves as an inherited representation interface, rather than a new contribution of AnyCT-Ω.

The inherited DPT-GS head produces a 138-channel dense map for each projection view. This output contains 137 Gaussian-compatible feature channels and one confidence channel. In the reported CT reconstruction pipeline, these channels are used as dense feature representations rather than directly rendered as Gaussian primitives. The CT objective is optimized through point-wise attenuation regression, and the Gaussian splatting decoder is not used to compute the reconstruction loss.

3.4 CONTENT-ROUTED RESIDUAL MIXTURE-OF-EXPERTS

To adapt dense features to heterogeneous CT inputs, we replace the final dense output projection with a content-routed residual MoE module(Shazeer et al., 2017). It contains a shared predictor and four residual experts. The global routing descriptor is derived from the normalized deepest-layer tokens $\bar{Z}^{23}$:

$$c = \left[\mathrm{Mean}_{v,p}(\bar{Z}_{v,p}^{23}) \parallel \mathrm{Mean}_{v,p}\left|\bar{Z}_{v,p}^{23} - \mathrm{AvgPool}(\bar{Z}_v^{23})_p\right|\right]. \tag{5}$$

For a detector location $\mathbf{u}$, the local router uses $X_v(\mathbf{u})$, while the global router uses c. The resulting soft expert weight is

$$g_{v,\mathbf{u},e} = \mathrm{softmax}_e\left(\frac{[R_{\mathrm{loc}}(X_v(\mathbf{u})) + \lambda_g \tanh(R_{\mathrm{glob}}(c))]_e}{\tau}\right), \tag{6}$$

where $\tau = 1$ and $\lambda_g = 0.1$.

The four residual experts model complementary pointwise, local 3×3, dilated 3×3, and high-frequency corrections, respectively. Let $S(\cdot)$ denote the shared predictor and $E_e(\cdot)$ denote expert e. The routed dense feature map is

$$F_v(\mathbf{u}) = S(X_v(\mathbf{u})) + \lambda_r \sum_{e=1}^{4} g_{v,\mathbf{u},e} E_e(X_v(\mathbf{u})), \tag{7}$$

where $\lambda_r = 0.1$. The routing and residual output layers are zero-initialized, such that training starts close to the shared predictor. All experts are evaluated for every input, yielding soft content-dependent specialization rather than sparse conditional computation.

3.5 GEOMETRY-GUIDED POINT QUERY AND ATTENUATION PREDICTION

For each query point $\mathbf{x}_i \in \Omega$, known CT geometry projects it to the normalized detector coordinate of view v. We bilinearly sample the routed dense feature map and aggregate evidence by view-wise max pooling:

$$\mathbf{u}_{v,i} = \pi_{G_v,\theta_v}(\mathbf{x}_i) \in [-1,1]^2, \qquad d_i = \max_{v=1,\ldots,V} B(F_v, \mathbf{u}_{v,i}), \tag{8}$$

where π_{G_v,θ_v} is determined by the known CT acquisition geometry(Feldkamp et al., 1984), and $B(\cdot)$ denotes bilinear sampling.

In parallel, patch tokens from every selected layer are reshaped into spatial maps T_v^ℓ and sampled at the same projected coordinates:

$$t_i^\ell = \max_{v=1,\dots,V} B(T_v^\ell, \mathbf{u}_{v,i}), \qquad \ell \in \{4, 11, 17, 23\}. \tag{9}$$

The sampled multi-level tokens are fused with the pooled dense feature, and the PointDecayDecoder predicts attenuation:

$$t_i = \phi_{\text{tok}}\left(t_i^4 \| t_i^{11} \| t_i^{17} \| t_i^{23}\right), \qquad \hat{\mu}_i = D_\psi\left([t_i \| d_i]\right). \tag{10}$$

Following the point-wise decoder design in sparse-view Gaussian CT reconstruction(Lin et al., 2024a), the PointDecayDecoder uses 1×1 convolutions, BatchNorm, LeakyReLU, and residual feature concatenation, with channel configuration $[C_h, C_h, 544, 136, 1]$. The coordinate $\mathbf{x}_i$ is not concatenated to the decoder input; it affects prediction through geometry-guided feature sampling.

3.6 Training and Inference

For each training sample, we randomly sample $N_q = 10{,}000$ query points and use $V \in \{6, 8, 10\}$ input views. With $\mu_i = \mu(\mathbf{x}_i)$ as the ground-truth attenuation, the reconstruction objective is

$$\mathcal{L}_{\text{rec}} = \frac{1}{N_q} \sum_{i=1}^{N_q} \left(\hat{\mu}_i - \mu_i\right)^2. \tag{11}$$

Additional details about volume normalization, block construction, projection generation, and view sampling are provided in AppendixA. The model is initialized from pretrained VGGT-Ω weights and jointly fine-tuned with the CT-specific modules. During inference, projections are encoded once; dense maps and token maps are reused while voxel centers are queried in chunks to form the final volume. No test-time optimization or ground-truth information is used at inference.

4 Experiments

In this section, we first conduct quantitative experiments on seven public CT datasets. We then provide qualitative comparisons, together with HU error maps to analyze the spatial distribution of reconstruction errors around tissue boundaries and high-frequency structures. Furthermore, we conduct comprehensive ablation studies to validate the effectiveness of each proposed module and the robustness of the pretrained model in non-ideal scenarios.

4.1 Experimental Setting

Datasets. AnyCT-70K combines CT-RATE(Hamamci et al., 2026), LUNA16(Setio et al., 2017), PANORAMA(Alves et al., 2024), PENGWIN(Liu et al., 2025), LUNA25(Peeters et al., 2026), RSNA2023(Hermans et al., 2024), and AbdomenAtlas(Qu et al., 2023). The processed corpus contains 72,670 training, 310 validation, and 630 test samples, 73,610 in total. Per-source train/validation/test splits are 53,077/50/100 for CT-RATE, 738/50/100 for LUNA16, 2,088/50/100 for PANORAMA, 60/10/30 for PENGWIN, 3,919/50/100 for LUNA25, 8,257/50/100 for AbdomenAtlas, and 4,531/50/100 for RSNA2023. Following the fixed-size sub-volume protocol used by DeepSparse, the 5,195 AbdomenAtlas1.0Mini volumes yield 8,407 reconstruction samples. The complete preprocessing and projection-generation pipeline, including the voxel spacing, HU normalization, block construction, cone-beam geometry, and projection quantization, is described in AppendixA.

Implementation. AnyCT-Ω initializes its multi-view encoder with pretrained VGGT-Ω weights and jointly fine-tunes the encoder and CT-specific head. Each batch samples 6, 8, or 10 projections and 10,000 supervised query points. We optimize point-wise MSE with AdamW for 30,000 steps on eight GPUs, using per-GPU batch size 4, learning rate 3×10^{-5} , backbone multiplier 0.04, and 1,000 warm-up steps. Evaluation fixes the requested view count and performs chunked voxel queries.

Method	CT-RATE (Chest)		LUNA25 (Chest)		LUNA16 (Chest)		PANORAMA (Abdomen)		RSNA2023 (Abdomen)		AbdomenAtlas (Abdomen)		PENGWIN (Pelvis)	
6-View														
DIF-Net	22.67	57.92	24.32	67.57	25.55	73.19	23.64	64.03	22.06	53.08	21.40	51.33	23.21	64.81
DIF-Gaussian	25.39	74.30	27.77	83.02	28.35	86.18	26.10	80.33	25.21	73.76	24.69	70.10	23.95	73.71
C^2RV	—	—	—	—	29.23	87.47	25.96	81.16	—	—	—	—	25.35	78.96
DeepSparse	**27.59**	**81.86**	**29.24**	**86.70**	**30.22**	**89.96**	**27.53**	**85.34**	**27.17**	**81.58**	**26.70**	**75.37**	**27.53**	**86.41**
AnyCT-Ω(Ours)	27.22	80.03	29.09	86.28	29.83	88.81	27.10	83.50	26.21	77.70	25.32	72.60	27.39	84.67
8-View														
DIF-Net	23.11	59.73	24.98	68.08	26.09	76.96	24.21	75.06	22.62	57.35	22.20	56.75	23.91	70.12
DIF-Gaussian	25.89	75.70	28.39	84.24	29.05	87.24	26.69	81.68	25.83	76.22	25.25	71.70	24.38	74.30
C^2RV	—	—	—	—	29.95	88.46	26.56	82.49	—	—	—	—	25.92	80.60
DeepSparse	**28.26**	**83.25**	**30.17**	**88.04**	**31.14**	**90.76**	**28.31**	**86.32**	**27.96**	**81.79**	**27.71**	**79.36**	**28.52**	**88.66**
AnyCT-Ω(Ours)	28.16	81.75	30.16	87.58	30.94	89.92	28.04	85.72	27.12	80.60	26.16	74.79	28.31	86.85
10-View														
DIF-Net	23.53	61.81	25.48	73.46	26.69	78.56	26.73	77.84	23.30	59.23	22.94	58.66	24.32	67.49
DIF-Gaussian	26.73	78.34	28.85	84.94	29.65	87.92	27.35	83.09	26.31	77.26	25.64	72.75	24.98	75.43
C^2RV	—	—	—	—	30.70	89.16	27.23	82.97	—	—	—	—	26.54	81.54
DeepSparse	**29.09**	**84.88**	30.61	88.57	**31.86**	**91.41**	**29.42**	**88.36**	**28.66**	**84.47**	**28.39**	**80.40**	29.03	**90.27**
AnyCT-Ω(Ours)	28.79	83.21	**30.84**	**88.61**	31.52	90.71	28.63	86.88	27.81	82.24	26.75	76.56	**29.10**	88.14

Table 1: Per-dataset evaluation across seven CT benchmarks under 6, 8, and 10 view settings. DIF-Net, DIF-Gaussian, C^2RV, and DeepSparse are trained separately on each individual dataset, whereas AnyCT-Ω is trained once on the mixed AnyCT-70K dataset and evaluated directly on every source dataset. Each entry reports PSNR (dB) and SSIM ($\times 10^{-2}$). The best and second-best values are highlighted in bold and underlined, respectively.

Baseline schedules and metrics. DIF-Net and DIF-Gaussian use 400 epochs on LUNA16, LUNA25, AbdomenAtlas, PENGWIN, PANORAMA, and RSNA2023; 60 epochs on CT-RATE; and 30 epochs on AnyCT-70K. DeepSparse uses two stages with 400 + 400, 60 + 60, and 30 + 30 epochs for the same settings, respectively. C^2RV is not publicly released, so unavailable experiments are reported as dashes. AnyCT-Ω uses the 30,000-step schedule above, approximately 12 epochs under our mixed-data sampler. We report volume-level PSNR in dB and SSIM × 100, averaged over test volumes. Table 1 includes all completed entries. A dash denotes an unavailable result, not zero. Comparisons use the same projection geometry, split, preprocessing, and metric implementation.

4.2 Main Results

Quantitative reconstruction. Table 1 evaluates cross-source transfer on individual datasets. In this experiment, the baselines are trained separately for each source, whereas AnyCT-Ω is trained once on the mixed AnyCT-70K corpus and evaluated without source-specific fine-tuning. AnyCT-Ω remains competitive across the seven datasets and achieves the best PSNR on LUNA25 and PENGWIN under the 10-view setting. These results indicate that a single model trained on heterogeneous sources can transfer effectively across anatomical domains, although source-specific models remain stronger on several individual benchmarks. Table 2 provides the controlled comparison in which all methods are trained and evaluated on the same heterogeneous AnyCT-70K mixture. AnyCT-Ω achieves the best performance under all three view settings. Compared with DeepSparse, the strongest completed baseline in this protocol, AnyCT-Ω improves PSNR by 1.50, 1.72, and 1.64 dB at 6, 8, and 10 views, respectively. The corresponding SSIM improvements are 3.76, 3.63, and 3.06 points. These results support the effectiveness of the proposed geometry-guided adaptation and content-routed residual specialization under heterogeneous joint training.

Qualitative reconstruction. Fig. 2 presents 10-view qualitative comparisons on LUNA16, PANORAMA, and PENGWIN, representing thoracic, abdominal, and pelvic anatomy, respectively. For each dataset, the columns show the ground truth (GT), DIF-Net, DIF-Gaussian, DeepSparse, and AnyCT-Ω, while the three rows correspond to matched axial, coronal, and sagittal views. Across the three anatomical regions, DIF-Net exhibits substantial blurring and streak-like artifacts under sparse-view acquisition. DIF-Gaussian recovers more global anatomical structures, but residual artifacts and fine-detail loss remain visible, particularly near high-contrast boundaries and small osseous structures. DeepSparse and AnyCT-Ω produce reconstructions that are visually more consistent with the GT and preserving sharper anatomical boundaries across different regions. These observations are consistent

Method	AnyCT-70K					
	6-View		8-View		10-View	
DIF-Net	22.45	61.39	23.39	63.28	23.63	61.84
DIF-Gaussian	25.83	77.05	26.41	78.78	26.93	79.18
DeepSparse	<u>25.96</u>	<u>77.88</u>	<u>26.70</u>	<u>79.93</u>	<u>27.42</u>	<u>81.81</u>
AnyCT-Ω(Ours)	**27.46**	**81.64**	**28.42**	**83.56**	**29.06**	**84.87**

Table 2: Comparison under heterogeneous joint training on AnyCT-70K. All methods are trained and evaluated on the same mixed seven-source dataset under 6, 8, and 10 view settings. Each entry reports PSNR (dB) and SSIM ($\times 10^{-2}$). The best and second-best values are highlighted in bold and underlined, respectively.

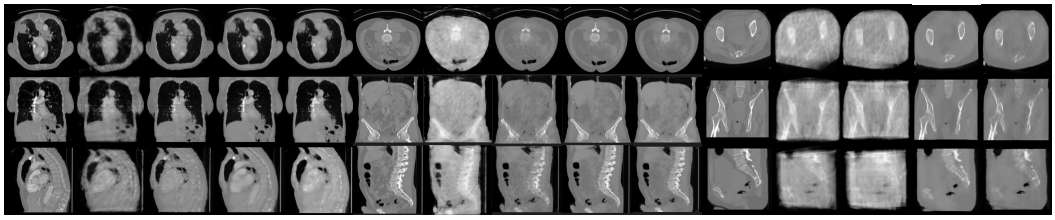


Figure 2: Qualitative comparison of 10-view CT reconstruction on LUNA16, PANORAMA, and PENGWIN. For each dataset, columns from left to right show GT, DIF-Net, DIF-Gaussian, DeepSparse, and AnyCT-Ω. Rows show matched axial, coronal, and sagittal slices.

with the quantitative results in Tables 1 and 2. Fig. 3 further visualizes the reconstruction error of AnyCT-Ω on representative cases from the three anatomical regions. The top and middle rows show the GT and reconstructed slices, respectively, while the bottom row shows the absolute reconstruction error in HU. The largest errors are mainly concentrated near high-contrast tissue interfaces, fine bone structures, and small anatomical boundaries. All error maps use the same color range from 0 to 150 HU. Additional HU error maps are provided in Appendix B.

4.3 Ablation Study

Data scaling across heterogeneous sources. To evaluate the effect of cross-organ data scaling, we compare single-source and four-source training under the 10-view setting. As shown in Table 3, four-source training improves PSNR on all four datasets, with gains of 0.35, 2.44, 0.53, and 0.32 dB on LUNA16, CT-RATE, PENGWIN, and PANORAMA, respectively. SSIM also improves on LUNA16, CT-RATE, and PANORAMA. Although PENGWIN shows a small SSIM decrease of 0.35 points, its PSNR still improves by 0.53 dB. These results indicate that thoracic, abdominal, and pelvic data provide complementary reconstruction priors and enable positive transfer across anatomical domains.

Source scaling and residual MoE. We further isolate the effect of increasing the number of training sources and adding the content-routed residual MoE on LUNA16. The models are trained for 30,000 steps. As shown in Table 4, increasing the training mixture from one to four and then seven sources raises PSNR from 30.65 to 31.00 and 31.35 dB, respectively. Adding the residual MoE further improves the result to 31.52 dB. Thus, the complete source-scaling protocol yields a gain of 0.70 dB, while the MoE contributes an additional gain of 0.17 dB under the seven-source setting. Additional batch-size, fine-tuning, token-fusion, and HU-error analyses are provided in Appendix B.

5 Discussion

The ablation results demonstrate positive transfer across anatomical regions. As shown in Table 3, jointly training on LUNA16, CT-RATE, PANORAMA, and PENGWIN improves PSNR on all four datasets compared with single-source training. This indicates that thoracic, abdominal, and pelvic data can provide complementary reconstruction priors. More experimental results can be viewed in the appendix.

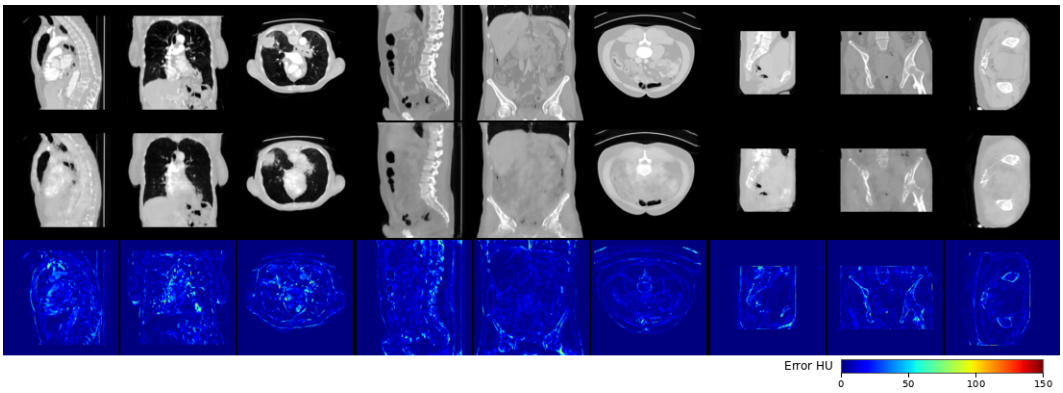


Figure 3: Absolute-error visualization of AnyCT-Ω across representative thoracic, abdominal, and pelvic CT cases. The top and middle rows show the GT and reconstructed slices, respectively; the bottom row shows absolute errors in HU. All error maps share the same color scale from 0 to 150 HU.

Table 3: Data-scaling ablation under the 10-view setting. Each entry reports PSNR (dB) and SSIM ($\times 10^{-2}$).

Dataset	Single source		Four-source mixture		Δ	
LUNA16	30.65	89.67	31.00	89.85	+0.35	+0.18
CT-RATE	26.05	75.93	28.49	82.29	+2.44	+6.36
PENGWIN	27.12	84.86	27.65	84.51	+0.53	−0.35
PANORAMA	28.16	85.48	28.48	86.48	+0.32	+1.00

This cross-organ benefit is particularly evident in the mixed AnyCT-70K evaluation. Competing methods generally degrade when trained jointly on heterogeneous sources, whereas AnyCT-Ω achieves the best performance under all view settings and reaches 29.06 dB / 84.87 SSIM at 10 views. In addition, AnyCT-Ω reaches the reported mixed-data performance after 30,000 optimization steps under our training protocol. This result suggests that AnyCT-Ω can reach competitive mixed-data performance with a relatively short optimization schedule.

6 Conclusion

We presented AnyCT-Ω, a unified framework for heterogeneous sparse-view CT reconstruction, together with the AnyCT-70K benchmark constructed from seven public CT sources. The method adapts pretrained cross-view representations to projection-based CT through known-geometry feature sampling and point-wise attenuation prediction. A content-routed residual mixture-of-experts head further provides input-dependent specialization without relying on dataset identifiers or dataset-specific prediction heads. Experiments under a controlled heterogeneous joint-training protocol show that AnyCT-Ω consistently outperforms the evaluated baselines on AnyCT-70K at 6, 8, and 10 views. The data-scaling experiments also show positive transfer as the number of training sources increases. These findings suggest that cross-source diversity can be beneficial when the model combines a shared representation with content-adaptive residual specialization.

Table 4: Ablation of training-source scaling and the content-routed residual MoE on 10-view.

Single	4 Combined	7 Combined	MoE	LUNA16		CT-RATE		PANORAMA		PENGWIN	
✓				30.65	89.67	26.05	75.93	28.16	85.48	27.12	84.86
	✓			31.00	89.85	28.49	82.29	27.65	84.51	28.48	86.48
		✓		31.35	90.48	28.62	82.76	27.85	85.08	28.99	87.76
		✓	✓	**31.52**	**90.71**	**28.79**	**83.21**	**28.63**	**86.88**	**29.10**	**88.14**

AI USE STATEMENT

Artificial intelligence tools were used solely to assist with mathematical reasoning and the presentation of selected mathematical formulations. The research idea, method design, experimental setup, implementation, experimental results, analysis, and references were independently developed, conducted, and verified by the authors. The authors take full responsibility for the correctness and originality of this submission.

REFERENCES

Jonas Adler and Ozan Oktem. Learned primal-dual reconstruction. *IEEE Transactions on Medical Imaging*, 37(6):1322–1332, 2018. ISSN 1558-254X. doi: 10.1109/tmi.2018.2799231. URL http://dx.doi.org/10.1109/TMI.2018.2799231.

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.

Anders H Andersen and Avinash C Kak. Simultaneous algebraic reconstruction technique (sart): a superior implementation of the art algorithm. *Ultrasonic imaging*, 6(1):81–94, 1984.

Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Lee A Feldkamp, Lloyd C Davis, and James W Kress. Practical cone-beam algorithm. *Journal of the Optical Society of America A*, 1(6):612–619, 1984.

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.

Sebastiaan Hermans, Zixuan Hu, Robyn L Ball, Hui Ming Lin, Luciano M Prevedello, Ferco H Berger, Ibrahim Yusuf, Jeffrey D Rudie, Maryam Vazirabad, Adam E Flanders, et al. Rsna 2023 abdominal trauma ai challenge: Review and outcomes. *Radiology: Artificial Intelligence*, 7(1): e240334, 2024.

Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *ACM Transactions on Graphics (TOG)*, 44(6):1–16, 2025.

Kyong Hwan Jin, Michael T McCann, Emmanuel Froustey, and Michael Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE transactions on image processing*, 26(9): 4509–4522, 2017.

Hiroyuki Kudo, Taizo Suzuki, and Essam A Rashed. Image reconstruction for sparse-view ct and interior ct—introduction to compressed sensing and differentiated backprojection. *Quantitative imaging in medicine and surgery*, 3(3):147, 2013.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pp. 71–91. Springer, 2024.

Yiqun Lin, Zhongjin Luo, Wei Zhao, and Xiaomeng Li. Learning deep intensity field for extremely sparse-view cbct reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 13–23. Springer, 2023.

Yiqun Lin, Hualiang Wang, Jixiang Chen, and Xiaomeng Li. Learning 3d gaussians for extremely sparse-view cone-beam ct reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 425–435. Springer, 2024a.

Yiqun Lin, Jiewen Yang, Hualiang Wang, Xinpeng Ding, Wei Zhao, and Xiaomeng Li. C 2 rv: cross-regional and cross-view learning for sparse-view cbct reconstruction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11205–11214. IEEE, 2024b.

Yiqun Lin, Jixiang Chen, Hualiang Wang, Jiewen Yang, Jiarong Guo, Yi Zhang, and Xiaomeng Li. Deepsparse: A foundation model for sparse-view cbct reconstruction. *IEEE Transactions on Medical Imaging*, 2026.

Yanzhen Liu, Sutuke Yibulayimu, Gang Zhu, Chao Shi, Chendi Liang, Chunpeng Zhao, Xinbao Wu, Yudi Sang, and Yu Wang. Automatic pelvic fracture segmentation: a deep learning approach and benchmark dataset. *Frontiers in Medicine*, 12:1511487, 2025.

Dré Peeters, Bogdan Obreja, Noa Antonissen, Zaigham Saghir, Ugo Pastorino, Mario Silva, Geertruida H de Bock, Hester Gietema, Fergus Gleeson, Marjolein A Heuvelmans, et al. Benchmarking of ai and radiologists for indeterminate lung nodule malignancy risk estimation at screening ct: The luna25 challenge. *Radiology: Artificial Intelligence*, 8(5):e260179, 2026.

Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. In *International Conference on Learning Representations*, volume 2024, pp. 28435–28445, 2024.

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.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12159–12168. IEEE, 2021.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in neural information processing systems*, 34:8583–8595, 2021.

Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas Van Den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical image analysis*, 42:1–13, 2017.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Malcolm Slaney and AC Kak. *Principles of computerized tomographic imaging*. IEEE press, 1988.

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.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. Vggt-ω. *arXiv preprint arXiv:2605.15195*, 2026.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709. IEEE, 2024.

Ruyi Zha, Yanhao Zhang, and Hongdong Li. Naf: neural attenuation fields for sparse-view cbct reconstruction. In *International conference on medical image computing and computer-assisted intervention*, pp. 442–452. Springer, 2022.

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 *International conference on medical image computing and computer-assisted intervention*, pp. 384–393. Springer, 2019.

A Supplementary Experiments and Analysis

The supplementary experiments use the same loss and evaluation protocols as the main paper. Unless otherwise stated, ablation experiments use 10-view input projections and are evaluated on LUNA16. In addition to component-level ablations, we conduct two cross-dataset transfer experiments. First, we train a model jointly on LUNA16 and CT-RATE, and evaluate it on seven datasets to assess generalization to held-out sources. Second, we pretrain the model on CT-RATE and continue training it on LUNA16 to evaluate whether cross-dataset initialization benefits downstream reconstruction. The corresponding results are reported in Tables 8 and 9.

All experiments were conducted using eight NVIDIA H20 GPUs, with 97,876 MB of GPU memory per device. The reconstructed volume has a resolution of 256^3, and $N_q = 10,000$ query points are sampled for each training sample. The number of input views is selected from $V \in \{6, 8, 10\}$. The VGGT-Ω backbone and the CT-specific modules are jointly fine-tuned. For the main AnyCT-Ω mixed-data experiments, training is performed for 30,000 optimization steps, corresponding to approximately 12 epochs under our data sampler. Reconstruction quality is measured using PSNR and SSIM, where SSIM is reported in units of 10^{-2}.

A.1 Data Preprocessing and Projection Generation

All CT volumes are converted to a common voxel coordinate system before projection generation. The volumes are resampled to an isotropic voxel spacing of $1.6\,\mathrm{mm}$ along all three axes using third-order interpolation. Each resampled volume is then center-cropped or zero-padded to a spatial resolution of $256 \times 256 \times 256$.

Let $h(\mathbf{x})$ denote the HU value at voxel position $\mathbf{x}$. We clip the HU values to $[-1000, 1024]$ and linearly normalize them to $[0, 1]$:

$$\tilde{\mu}(\mathbf{x}) = \frac{\mathrm{clip}(h(\mathbf{x}), -1000, 1024) + 1000}{2024}. \tag{12}$$

The normalized volumes are stored as 8-bit arrays using $\mathrm{round}(255\tilde{\mu})$ and are converted back to floating-point values in $[0, 1]$ during training and evaluation.

For efficient data loading and point sampling, each 256^3 volume is associated with $4 \times 4 \times 4$ block index groups. Each group contains 64^3 voxel coordinates, corresponding to the configured block size of $64 \times 64 \times 64$. During training, one block is randomly selected and $N_q = 10,000$ query points are sampled from this block without replacement. The corresponding normalized attenuation values are used as point-wise supervision.

We generate synthetic projection images using a cone-beam forward projector with the geometry specified in the preprocessing configuration. The source-to-detector distance is set to $\mathrm{DSD} = 1200\,\mathrm{mm}$ and the source-to-origin distance is set to $\mathrm{DSO} = 800\,\mathrm{mm}$. The detector contains 256×256 pixels, with a detector pixel size of $2.4\,\mathrm{mm} \times 2.4\,\mathrm{mm}$. The projected volume has a resolution of 256^3 and a voxel size of $1.6\,\mathrm{mm} \times 1.6\,\mathrm{mm} \times 1.6\,\mathrm{mm}$. Both the object and detector offsets are set to zero. We use cone-beam mode with a forward-projection accuracy of 0.5 voxel per sampling interval. No additional projection filter or synthetic noise model is applied. For each volume, $K = 50$ projections are generated over a 180° angular range starting from 0°. The projection angles are uniformly sampled as

$$\theta_k = \frac{180^\circ}{K} k, \qquad k = 0, \ldots, K-1, \tag{13}$$

which gives an angular interval of 3.6°. Let $\mathcal{A}_{\mathrm{cone}}$ denote the cone-beam forward projector and $\mathcal{G}$ denote the acquisition geometry. The generated projection at angle θ_k is written as

$$P_k = \mathcal{A}_{\mathrm{cone}}\left(\tilde{\mu}; \mathcal{G}, \theta_k\right). \tag{14}$$

The detector-axis convention and detector-image orientation are kept consistent with the DIF-Gaussian preprocessing pipeline.

To reduce storage cost, let

$$p_{\max} = \max_{k,u,v} P_k(u, v) \tag{15}$$

Table 5: Effect of batch size on 10-view LUNA16 reconstruction.

Batch size	PSNR (dB)	SSIM ($\times 10^{-2}$)
1	27.41	85.03
4	30.48	89.16
32	**30.65**	**89.67**

denote the maximum projection value of one volume. The projections are quantized and stored as 8-bit arrays:

$$\bar{P}_k(u,v) = \mathrm{round}\left(255\frac{P_k(u,v)}{p_{\max}}\right) \in \{0,\dots,255\}. \tag{16}$$

The quantized projections, $p_{\max}$, and their corresponding angles are stored together. During data loading, the projections are converted back to floating-point values and rescaled according to the DIF-Gaussian convention:

$$P_k^{\mathrm{input}} = \frac{\bar{P}_k}{255}\frac{p_{\max}}{0.2}. \tag{17}$$

At training time, we select $V \in \{6,8,10\}$ views from the $K = 50$ stored projections. The nominal view indices are obtained by approximately uniform sampling:

$$\mathcal{I}_{V,o} = \left\{\left\lfloor \frac{jK}{V} \right\rfloor + o\right\}_{j=0}^{V-1}, \tag{18}$$

where o is a valid random offset during training. During evaluation, the number of views is fixed and the view offset is deterministic.

For geometry-guided feature sampling, a normalized query point $\mathbf{x} = (x_1, x_2, x_3) \in [0,1]^3$ is first converted to physical coordinates:

$$\mathbf{q} = \left((x_1 - \tfrac{1}{2})L_x,\ (x_2 - \tfrac{1}{2})L_y,\ (\tfrac{1}{2} - x_3)L_z\right), \tag{19}$$

where $L_x = L_y = L_z = 256 \times 1.6\,\mathrm{mm}$. After rotation by $-\theta_k$ around the axial direction, we obtain

$$\mathbf{q}_k = R_z(-\theta_k)\mathbf{q}. \tag{20}$$

The point is then projected onto the detector using

$$s_k = \frac{\mathrm{DSD}}{\mathrm{DSO} - q_{k,x}}, \qquad u_k = \frac{2q_{k,z}s_k}{N_u d_u}, \qquad v_k = \frac{2q_{k,y}s_k}{N_v d_v}, \tag{21}$$

where $(N_u, N_v) = (256, 256)$ and $(d_u, d_v) = (2.4, 2.4)\,\mathrm{mm}$. The resulting normalized detector coordinates are used to bilinearly sample the dense feature maps and intermediate token maps.

B Additional Ablation, Transfer, and Visualization Results

B.1 Additional Experimental Details

Effect of batch size. Table 5 evaluates the influence of batch size on 10-view LUNA16 reconstruction. Increasing the batch size improves both PSNR and SSIM, with the best result obtained at batch size 32.

Effect of backbone fine-tuning. Table 6 compares different strategies for adapting the pretrained VGGT-Ω backbone. Freezing the backbone gives the weakest result, while LoRA fine-tuning provides a clear improvement. Full fine-tuning achieves the best PSNR and SSIM, indicating that adapting the complete representation is beneficial for CT reconstruction.

Table 6: Effect of backbone fine-tuning strategy on 10-view LUNA16 reconstruction.

Fine-tuning strategy	PSNR (dB)	SSIM ($\times 10^{-2}$)
Frozen backbone	27.48	85.36
LoRA fine-tuning	29.94	88.75
Full fine-tuning	**30.65**	**89.67**

Table 7: Effect of token fusion mode on 10-view LUNA16 reconstruction.

Token fusion mode	PSNR (dB)	SSIM ($\times 10^{-2}$)
`last`	28.52	79.37
`mean`	30.51	88.95
`concat`	**30.65**	**89.67**

Effect of token fusion mode. The `token_fusion_mode` determines how intermediate layers of the VGGT-Ω aggregator are selected and fused. The `last` setting uses only the final layer. The `mean` setting averages the features from layers 4, 11, 17, and 23. The `concat` setting concatenates these four layer features and is used in the main experiments. As shown in Table 7, multi-level feature fusion substantially outperforms using only the final layer.

Cross-dataset generalization from two-source training. To evaluate whether the learned representation can transfer to datasets that are not used for training, we jointly train AnyCT-Ω on LUNA16 and CT-RATE only. The best checkpoint is then evaluated on all seven datasets without dataset-specific fine-tuning. As shown in Table 8, the model transfers to the five held-out sources and achieves 30.24/87.74 on LUNA25, 26.70/83.67 on PANORAMA, 27.05/85.28 on PENGWIN, 26.03/78.11 on RSNA2023, and 23.15/67.50 on AbdomenAtlas in terms of PSNR/SSIM. The performance varies across anatomical regions, with LUNA25 being the strongest held-out source and AbdomenAtlas being the most challenging. These results provide evidence that the learned representation is not restricted to the two training distributions, while robustness to scanner-specific acquisition shifts remains unverified because all projections are generated using the same standardized synthetic geometry.

Sequential cross-dataset pretraining and fine-tuning. We further evaluate whether pretraining on CT-RATE provides a useful initialization for LUNA16 reconstruction. Specifically, we first train the model on CT-RATE and then continue training from the best CT-RATE checkpoint on LUNA16. As reported in Table 9, this sequential training strategy achieves 31.66 dB PSNR and 90.78 SSIM on the LUNA16 test set. Compared with the full fine-tuning result in Table 6, which obtains 30.65 dB PSNR and 89.67 SSIM under the same 10-view setting, cross-dataset pretraining improves PSNR by 1.01 dB and SSIM by 1.11 points. This result suggests that CT-RATE can provide complementary reconstruction knowledge for LUNA16. Since this experiment evaluates one sequential training configuration, it should be interpreted as evidence of transfer rather than a complete causal ablation of pretraining.

B.2 Supplementary HU Error Visualization

The component-level ablations above analyze the effects of batch size, backbone fine-tuning, and token fusion. The two transfer experiments further show that the learned representation can generalize

Table 8: Cross-dataset generalization of a model jointly trained on LUNA16 and CT-RATE. The model is evaluated without dataset-specific fine-tuning on the two training sources and five held-out CT datasets.

Dataset	LUNA16	CT-RATE	LUNA25	PANORAMA	PENGWIN	RSNA2023	AbdomenAtlas
PSNR	30.98	28.53	30.24	26.70	27.05	26.03	23.15
SSIM ($\times 10^{-2}$)	89.90	82.29	87.74	83.67	85.28	78.11	67.50

Table 9: Sequential cross-dataset pretraining and fine-tuning on 10-view LUNA16 reconstruction. The model is first trained on CT-RATE and then fine-tuned on LUNA16.

Training strategy	PSNR (dB)	SSIM ($\times 10^{-2}$)
CT-RATE pretraining $\rightarrow$ LUNA16 fine-tuning	31.66	90.78

from a small number of training sources to held-out datasets, while sequential CT-RATE pretraining can improve subsequent LUNA16 reconstruction. We next visualize the remaining reconstruction errors using absolute HU error maps.

Figure 4 presents supplementary qualitative results on CT-RATE, RSNA2023, and LUNA25. The error maps highlight reconstruction discrepancies around anatomical boundaries and high-contrast structures. The color bar reports the absolute reconstruction error in Hounsfield units (HU).

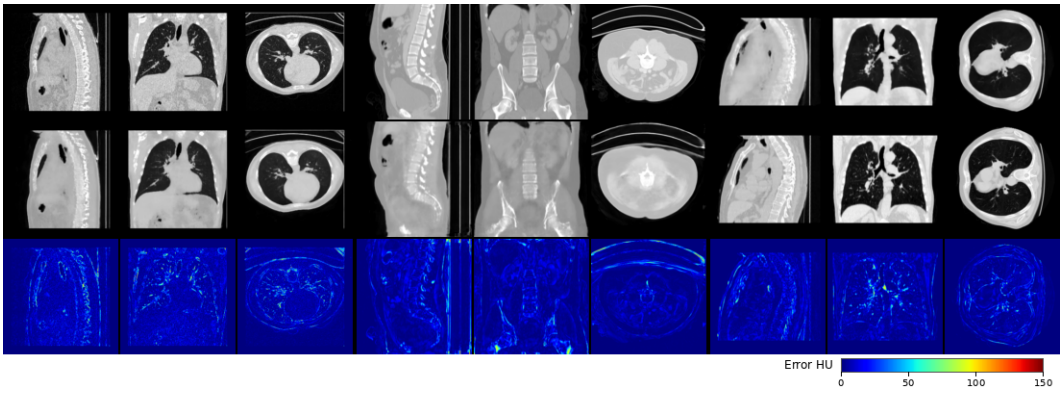


Figure 4: Supplementary HU error visualization. From left to right, the three image groups correspond to CT-RATE, RSNA2023, and LUNA25. The error maps show the absolute reconstruction error in Hounsfield units (HU).

C LIMITATIONS

Although AnyCT-Ω demonstrates positive transfer across multiple CT sources, several limitations remain. First, AnyCT-70K is not perfectly balanced across sources, so pooled metrics may be dominated by the largest datasets. Second, the two-source transfer experiment provides evidence of generalization to held-out datasets, but it does not cover all possible source combinations or acquisition conditions. Third, the projection data are generated using a standardized cone-beam geometry; therefore, the current experiments do not fully evaluate robustness to scanner-specific geometry, realistic noise, patient motion, metal artifacts, or other acquisition shifts. Fourth, the sequential CT-RATE pretraining experiment does not include all matched controls needed to isolate the causal contribution of pretraining from differences in optimization history. Finally, most ablation studies are conducted on 10-view LUNA16, and wall-clock time, inference memory, and scaling behavior across different hardware configurations remain to be systematically evaluated.