

1 A Technical Appendices and Supplementary Material

2 In this supplementary document, we provide additional insights and supporting materials for our main
3 paper. We begin by outlining the key limitations of our work and discussing the broader societal and
4 clinical implications of the BTB3D framework, particularly with regard to its capabilities in radiology
5 report generation and text-conditioned 3D CT volume synthesis. We then present further experimental
6 details that expand upon the methodology described in the main text. Additional qualitative and
7 quantitative results are also included to further validate our findings. All code, pretrained model
8 weights, and instructions to reproduce our experiments will be made openly available in our GitHub
9 repository to promote transparency, reproducibility, and further research in the field.

10 A.1 Limitations and Broader Impacts

11 **Limitations.** While our BTB3D framework significantly improves tokenization and decoding for
12 3D medical vision-language modeling (especially for 3D chest CT scans), several limitations remain.
13 First, our current framework does not explicitly model clinical reasoning or uncertainty, both of which
14 are crucial for real-world deployment in clinical settings. Incorporating modules for uncertainty
15 estimation or causal inference remains an open challenge for future studies.

16 Second, the scope of our experiments is limited to 3D chest CT scans, primarily due to the lack of
17 large-scale, publicly available paired datasets (with reports) for other anatomical regions or modalities
18 (e.g., MRI, PET). While BTB3D is designed to generalize across 2D and 3D inputs, we have not yet
19 validated its transferability to other clinical domains. Future efforts should extend the architecture to
20 whole-body imaging and explore zero-shot or few-shot generalization across organs.

21 Third, although BTB3D supports both 2D and 3D training modes, we used only 2D slices extracted
22 from 3D volumes in CT-RATE to remain consistent with baselines and ensure fair comparisons. As
23 such, we have not demonstrated its transfer capabilities (e.g., transfer from pretrained 2D models or
24 joint training with 2D paired datasets), which limits the evaluation of BTB3D’s capabilities.

25 Fourth, statistical analysis is limited due to the high computational cost of 3D training. We did not
26 perform repeated runs or report confidence intervals. Thus, while our results demonstrate strong and
27 consistent performance across multiple benchmarks, statistical significance remains to be validated.

28 Lastly, although we evaluate both clinical and generative metrics, BTB3D has not yet been assessed
29 in real clinical workflows. Validation through expert reader studies or prospective trials is necessary
30 to fully establish its utility, reliability, and safety in decision-making environments.

31 **Broader impacts.** BTB3D offers promising benefits for both clinical and research communities.
32 High-quality radiology report generation from 3D CT scans can reduce reporting delays, alleviate
33 radiologist burnout, and improve documentation consistency, particularly in high-volume settings
34 such as emergency departments or large-scale screening programs. Moreover, BTB3D’s ability
35 to synthesize realistic, anatomically plausible CT scans from text opens new avenues for training,
36 simulation, and rare disease modeling. In educational settings, synthetic CT scans can be used to
37 create diverse training sets for radiology students and enhance learning outcomes through exposure
38 to rare or atypical cases. In research, BTB3D may support data augmentation, helping mitigate class
39 imbalance in supervised learning tasks and enabling pretraining in low-resource environments.

40 However, the generative capabilities of the BTB3D framework also present potential risks. Synthetic
41 scans, if not properly labeled or constrained, may be misused in regulatory or insurance contexts, or
42 exploited for fraudulent purposes. Additionally, there is a non-trivial risk of inadvertently replicating
43 identifiable patient anatomy, even when training data is anonymized. While our work uses only pub-
44 licly available, fully de-identified datasets (e.g., CT-RATE), developers must ensure strict compliance
45 with data governance and anonymization protocols when deploying similar models.

46 Finally, the substantial computational cost of training such models may exacerbate disparities in
47 access to advanced medical AI. Ensuring open-source availability and supporting low-resource
48 inference are critical to democratizing BTB3D’s benefits. Continued oversight and interdisciplinary
49 dialogue will be essential to ensuring the responsible deployment of generative models in healthcare.

50 A.2 Additional Experimental Details

51 **Three-stage training performance.** To further evaluate the impact of our progressive training
52 strategy, we qualitatively assess the reconstruction performance of BTB3D across the three stages of
53 training in Figure 1. The figure presents axial, coronal, and sagittal slices reconstructed by our two
54 model variants ($16\times 16\times 8$ and $8\times 8\times 8$) at each stage and compares them to the ground truth.

55 In *Stage 1*, the model is trained solely on small, non-overlapping 3D volumes and 2D slices. This
56 initialization allows the model to learn basic volumetric structure and inter-slice continuity but suffers
57 from block artifacts and poor spatial coherence, especially in regions with high anatomical complexity.
58 As seen in the leftmost columns of Figure 1, both variants at this stage produce noisy and blurry
59 reconstructions with limited detail and sharpness. *Stage 2* introduces overlapping windows during
60 training, which significantly improves anatomical consistency across slices and helps reduce the
61 block discontinuities learned in Stage 1. This stage yields the largest qualitative leap in fidelity, with
62 more well-defined boundaries of lung lobes, airways, and soft tissue structures. The improvements
63 are most noticeable in the axial and sagittal views, where inter-slice alignment and smoothness are
64 critical for clinical interpretability. *Stage 3* refines the decoder with high-resolution patches while
65 keeping the encoder part frozen. This final stage enhances local detail, texture realism, and structural
66 sharpness without sacrificing the global consistency achieved in Stage 2. Notably, lung fissures,
67 pleural contours, and fine vascular structures become visibly clearer, indicating that the decoder has
68 learned to reconstruct complex anatomical regions with higher fidelity and resolution.

69 Comparing the two BTB3D variants, the $8\times 8\times 8$ model consistently achieves superior anatomical
70 fidelity and inter-slice coherence in reconstruction, as expected due to its lower compression rate and
71 higher capacity. In contrast, the $16\times 16\times 8$ variant (while more compressed) proves more effective for
72 language-driven tasks such as report generation, where coarse global structure suffices and memory
73 efficiency is critical. Our three-stage training pipeline plays a pivotal role for both models: it first
74 captures global structure, then progressively refines local detail, enabling accurate and high-resolution
75 3D reconstructions from compact tokens. This staged optimization bridges the gap between token
76 efficiency and clinical utility, facilitating both high-fidelity text-conditional CT generation and precise
77 report generation, underscoring BTB3D’s versatility in multimodal 3D medical image understanding.

78 **Radiology report generation from 3D chest CT.** We comprehensively evaluate BTB3D’s capabil-
79 ity in generating accurate and clinically coherent radiology reports from volumetric chest CT scans.
80 As described in the main paper, each CT volume is first compressed into a compact sequence of
81 frequency-aware 3D tokens by our encoder, and these tokens are then passed to a pretrained LLM
82 (LLaMA 3.1 8B) for report generation via a linear projection layer. To ensure fairness, we use
83 the official weights for CT2Rep and CT-CHAT, both trained on the same CT-RATE dataset. Since
84 Merlin was originally trained on a private dataset and neither its weights nor training data are publicly
85 available, we retrained Merlin on CT-RATE using its official codebase and default hyperparameters.

86 Our evaluation includes both internal (CT-RATE test set) and external (RAD-ChestCT) benchmarks.
87 The results in Table 1 show that BTB3D (particularly the $16\times 16\times 8$ variant) outperforms prior
88 methods in average abnormality level F1 scores, achieving higher clinical precision in most of the
89 findings. This trend holds in Table 2, where our BTB3D framework demonstrates strong out-of-
90 distribution generalization, with the $16\times 16\times 8$ variant yielding a 46% relative F1 improvement over
91 CT-CHAT (the previous state-of-the-art method). This highlights the robustness of our tokenization
92 and training pipeline, even when evaluated on unseen institutional distributions.

93 Figure 2 offers a qualitative comparison across models, illustrating report generation for the same
94 CT scan. BTB3D’s reports more faithfully reproduce key clinical details from the ground truth,
95 particularly with the higher compression ($16\times 16\times 8$) variant, supporting the notion that coarser
96 representations, while less suitable for pixel-level synthesis, may better capture global semantics for
97 language modeling. Meanwhile, the $8\times 8\times 8$ variant provides more spatially detailed reconstructions
98 but slightly underperforms on text generation metrics, suggesting a trade-off between compression
99 depth and semantic abstraction. In all comparisons, BTB3D demonstrates a compelling advantage by
100 combining precise volumetric tokenization with a scalable training strategy, ultimately allowing both
101 fine-grained anatomical reconstruction and clinically relevant report generation.

Text-conditional 3D chest CT generation. To evaluate the generative capabilities of BTB3D, we benchmark its performance on synthesizing realistic and anatomically coherent 3D chest CT volumes from free-text clinical prompts. As shown in Figure 3, BTB3D generates sharper, more anatomically faithful volumes compared to MedSyn and GenerateCT, with improved inter-slice consistency and alignment to prompt semantics. For fair comparison, we use the official pretrained weights for both MedSyn and GenerateCT, which were trained on the same modality (3D chest CT), ensuring that differences in performance stem from architectural and training innovations.

Quantitative results reported in the main paper (Table 4) show that our BTB3D framework with the lower compression variant ($8 \times 8 \times 8$) achieves the best overall performance across all generative metrics. Specifically, it reduces the mean Fréchet Inception Distance (FID) from 9.51 (GenerateCT) and 12.59 (MedSyn) to just 2.24, a 76.5% improvement, demonstrating superior fidelity. We compute FID using the FIDMetric from the MONAI library [1], leveraging a RadImageNet-pretrained ResNet50 backbone [2], which is better suited for grayscale radiology images than traditional Inception networks. Following standard practice established by GenerateCT and MAISI [3], FID is calculated on the central 40% of slices (in each anatomical plane) across 100 randomly selected volumes per method, reducing boundary noise and focusing on clinically relevant regions.

In terms of temporal and anatomical realism, Fréchet Video Distance (FVD) is computed using both CT-Net (specialized for 3D chest CTs) [4] and I3D (trained on RGB videos). BTB3D again significantly outperforms prior work, halving the FVD compared to GenerateCT. To assess semantic alignment between text prompts and generated volumes, we compute CLIP scores using the CLIPScore implementation from Torchmetrics. We follow GenerateCT’s protocol: axial slices are resized to 224×224 and converted to pseudo-RGB by repeating the single intensity channel. Using clip-vit-base-patch16, we observe that BTB3D achieves the highest text-image alignment score (24.27), suggesting it captures fine semantic cues better than prior methods.

Interestingly, we note a compression-quality trade-off: while the $8 \times 8 \times 8$ variant excels in fine-grained reconstruction and text-conditional volume synthesis, the more compact $16 \times 16 \times 8$ variant still surpasses existing baselines and may be preferable in memory-constrained or latency-sensitive settings. Together, these results confirm that BTB3D’s volumetric tokenization and three-stage training pipeline offer a significant leap forward in text-conditional 3D medical image generation, bridging the gap between semantic understanding and pixel-level anatomical coherence.

References

- [1] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.
- [2] Xueyan Mei, Zelong Liu, Philip Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E. Link, Thomas Yang, Ying Wang, Hayit Greenspan, Timothy Deyer, Zahi A. Fayad, and Yang Yang. Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 0(ja):e210315, 0.
- [3] Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, et al. Maisi: Medical ai for synthetic imaging. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4430–4441. IEEE, 2025.
- [4] Rachel Lea Draelos, David Dov, Maciej A Mazurowski, Joseph Y Lo, Ricardo Henao, Geoffrey D Rubin, and Lawrence Carin. Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. *Medical image analysis*, 67:101857, 2021.

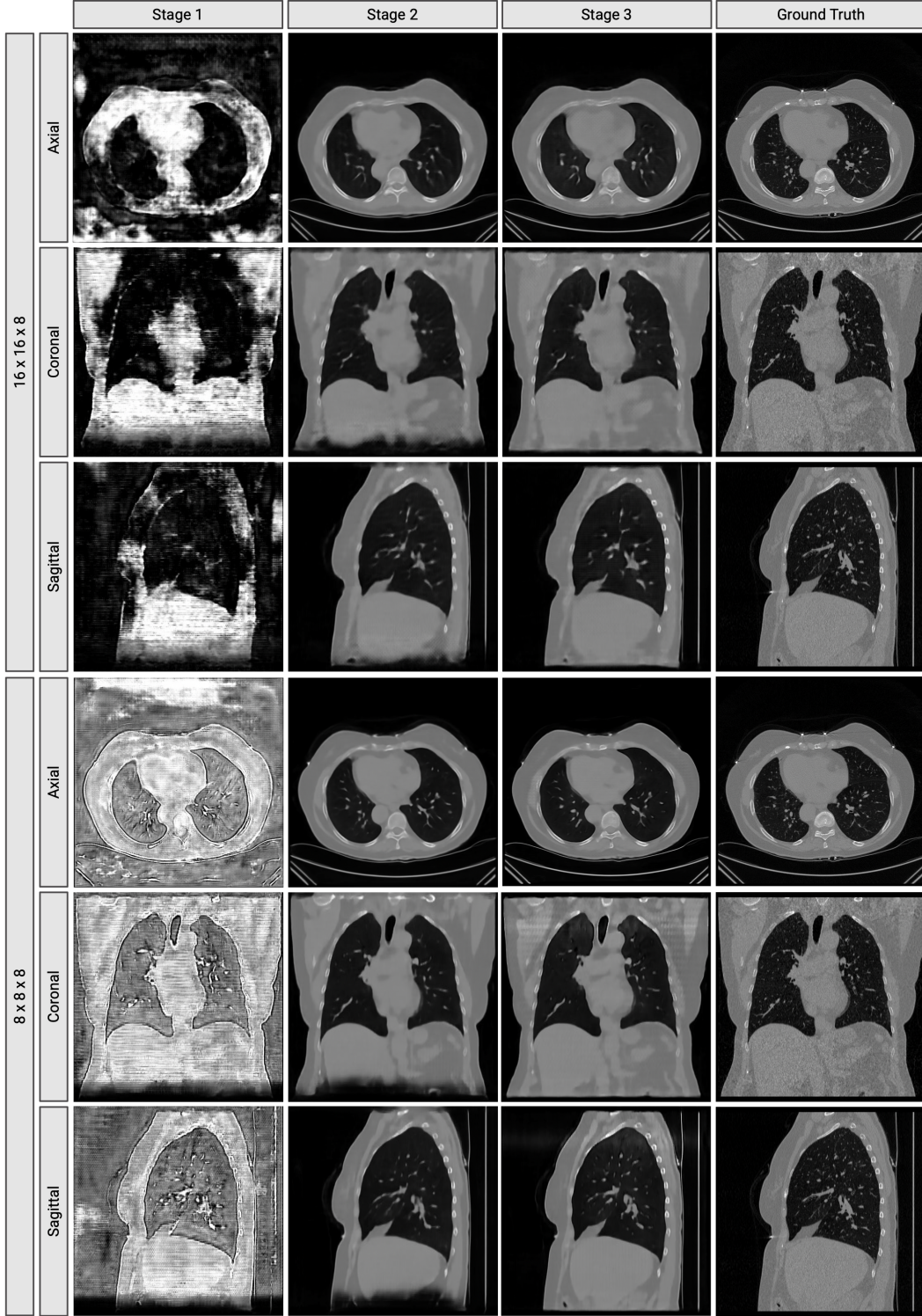


Figure 1: Qualitative reconstruction results across axial, coronal, and sagittal planes for two BTB3D variants: $16 \times 16 \times 8$ (top) and $8 \times 8 \times 8$ (bottom). The figure shows outputs after Stage 1 (short-volume training), Stage 2 (overlapping-window training), and Stage 3 (decoder refinement), compared to the ground truth. The progressive improvements highlight the effectiveness of our three-stage training strategy. Stage 2 yields the largest gain in anatomical fidelity and inter-slice consistency, while Stage 3 further sharpens structural details such as lung fissures and vascular boundaries.

Table 1: Abnormality-based F1 scores on CT-RATE along with the prevalence ratio of each abnormality in the test set. BTB3D consistently achieves the highest clinical accuracy across most categories, with the higher-compression variant ($16\times 16\times 8$) demonstrating superior overall performance.

Abnormality	Ratio	Ours-16	Ours-8	CT-CHAT	Merlin	CT2Rep
Medical material	0.103	0.142	0.120	0.006	0.057	0.000
Arterial wall calcification	0.285	0.414	0.273	0.451	0.262	0.322
Cardiomegaly	0.107	0.305	0.207	0.123	0.176	0.013
Pericardial effusion	0.074	0.095	0.095	0.009	0.060	0.000
Coronary artery wall calc.	0.252	0.403	0.260	0.412	0.235	0.335
Hiatal hernia	0.137	0.164	0.118	0.207	0.110	0.074
Lymphadenopathy	0.260	0.358	0.209	0.069	0.227	0.013
Emphysema	0.197	0.196	0.155	0.391	0.216	0.198
Atelectasis	0.235	0.269	0.242	0.341	0.199	0.323
Lung nodule	0.448	0.427	0.397	0.443	0.290	0.029
Lung opacity	0.390	0.408	0.382	0.266	0.312	0.557
Pulmonary fibrotic sequela	0.273	0.318	0.211	0.069	0.117	0.104
Pleural effusion	0.124	0.308	0.199	0.173	0.183	0.341
Mosaic attenuation pattern	0.083	0.183	0.094	0.064	0.076	0.198
Peribronchial thickening	0.117	0.125	0.043	0.000	0.054	0.099
Consolidation	0.191	0.259	0.185	0.120	0.174	0.236
Bronchiectasis	0.109	0.126	0.094	0.091	0.075	0.013
Interlobular septal thick.	0.082	0.135	0.087	0.075	0.065	0.032
Mean	0.193	0.258	0.187	0.184	0.160	0.160

Table 2: Abnormality-wise F1 scores on the RAD-ChestCT dataset (external test set), along with the prevalence ratio of each abnormality. BTB3D demonstrates strong generalization performance, particularly with the $16\times 16\times 8$ variant, achieving the highest F1 score across most categories.

Abnormality	Ratio	Ours-16	Ours-8	CT-CHAT	Merlin	CT2Rep
Medical material	0.327	0.235	0.154	0.000	0.107	0.000
Calcification	0.706	0.671	0.406	0.567	0.359	0.434
Cardiomegaly	0.109	0.187	0.151	0.181	0.031	0.010
Pericardial effusion	0.155	0.193	0.101	0.007	0.047	0.089
Hiatal hernia	0.117	0.149	0.087	0.149	0.136	0.173
Lymphadenopathy	0.165	0.252	0.227	0.121	0.176	0.000
Emphysema	0.273	0.201	0.172	0.412	0.242	0.142
Atelectasis	0.298	0.350	0.265	0.387	0.194	0.153
Lung nodule	0.802	0.424	0.408	0.721	0.495	0.068
Lung opacity	0.539	0.542	0.389	0.140	0.394	0.560
Pulmonary fibrotic sequela	0.132	0.204	0.139	0.026	0.174	0.157
Pleural effusion	0.200	0.290	0.199	0.043	0.135	0.032
Peribronchial thickening	0.085	0.117	0.064	0.000	0.095	0.039
Consolidation	0.139	0.217	0.145	0.108	0.149	0.264
Bronchiectasis	0.158	0.128	0.084	0.034	0.122	0.000
Interlobular septal thick.	0.069	0.091	0.075	0.008	0.057	0.000
Mean	0.267	0.266	0.192	0.182	0.182	0.133

	<p>Findings: Trachea and both main bronchi are open. No occlusive pathology was detected in the trachea and both main bronchi. There are linear atelectasis in the right lung middle lobe and medial upper lobe anterior segment, and left lung upper lobe lingular segment. Minimal emphysematous changes were observed in both lungs. There is a millimetric nodule in the upper lobe of the right lung. No mass or infiltrative lesion was detected in both lungs. As far as can be observed: Heart contour and size are normal. No pleural or pericardial effusion was detected. The widths of the mediastinal main vascular structures are normal. No pathologically enlarged lymph nodes were detected in the mediastinum and hilar regions. No pathological wall thickness increase was observed in the esophagus within the sections. No upper abdominal free fluid-collection was detected in the sections. No enlarged lymph nodes in pathological dimensions were detected. In the upper abdominal organs within the sections, there is no mass with distinguishable borders as far as it can be observed within the borders of non-enhanced CT. Thoracic vertebral corpus heights, alignments and densities are normal. There are osteophytes in the vertebral corpus corners.</p> <p>Impression: Minimal emphysematous changes in both lungs. Linear atelectasis in both lungs. Millimetric nodule in the upper lobe of the right lung.</p>	Ground Truth
	<p>Findings: Trachea, both main bronchi are open. Mediastinal main vascular structures, heart contour, size are normal. Thoracic aorta diameter is normal. Pericardial effusion-thickening was not observed. Thoracic esophageal calibration was normal and no significant tumoral wall thickening was detected. No enlarged lymph nodes in prevascular, pre-paratracheal, subcarinal or bilateral hilar-axillary pathological dimensions were detected. When examined in the lung parenchyma window, Minimal paraseptal emphysematous changes are observed in the right lung lower lobe superior segment posterolateral. Linear atelectasis changes are observed in the right lung middle and upper lobe anterior segment. There is a nodule with a diameter of 3 mm in the upper lobe of the right lung. There was no finding compatible with pleural effusion, pneumothorax or pneumonia in both lungs. Upper abdominal organs included in the sections are normal. No space-occupying lesion was detected in the liver that entered the cross-sectional area. Bilateral adrenal glands were normal and no space-occupying lesion was detected. Mild degenerative changes are observed in the bone structure entering the examination area. Vertebral corpus heights are preserved.</p> <p>Impression: Minimal paraseptal emphysematous changes in the right lung lower lobe superior segment posterolateral, linear subpleural atelectasis changes in the right lung middle and upper lobe anterior segment, a nonspecific nodule in the upper lobe of the right lung; no findings in favor of pneumonia were detected.</p>	Ours (16x16x8)
	<p>Findings: Trachea and both main bronchi are open. No occlusive pathology was detected in the trachea and both main bronchi. The left lung is in the columnar fashion and there are calcific nodules in the interlobular fissure on the left. The sequelae were evaluated in favor of changes. There are linear atelectasis in both lungs. Emphysematous changes were observed in both lungs. No mass or infiltrative lesion was detected in both lungs. Mediastinal structures cannot be evaluated optimally because contrast material is not given. As far as can be observed: Heart contour and size are normal. There is minimal pericardial effusion. The widths of the mediastinal main vascular structures are normal. There are no pathologically enlarged lymph nodes in the mediastinum and hilar regions. There is no pathological wall thickness increase in the esophagus within the sections. No upper abdominal free fluid-collection was detected in the sections. No enlarged lymph nodes in pathological dimensions were detected. No fractures or lytic-destructive lesions were observed in the bone structures within the sections.</p> <p>Impression: Emphysematous changes in both lungs. Locally linear atelectasis in both lungs. Minimal pericardial effusion.</p>	Ours (8x8x8)
	<p>Findings: Trachea and both main bronchi are open. No occlusive pathology was detected in the trachea and both main bronchi. There are minimal emphysematous changes in both lungs. There are millimetric nonspecific nodules in both lungs. No mass or infiltrative lesion was detected in both lungs. Mediastinal structures cannot be evaluated optimally because contrast material is not given. As far as can be observed: Heart contour and size are normal. No pleural or pericardial effusion was detected. The widths of the mediastinal main vascular structures are normal. No pathologically enlarged lymph nodes were detected in the mediastinum and hilar regions. No pathological wall thickness increase was observed in the esophagus within the sections. No upper abdominal free fluid-collection was detected in the sections. No enlarged lymph nodes in pathological dimensions were detected. In the upper abdominal organs within the sections, there is no mass with distinguishable borders as far as it can be observed within the borders of non-enhanced CT. Thoracic vertebral corpus heights, alignments and densities are normal. Intervertebral disc distances are preserved. The neural foramina are open. No lytic-destructive lesions were detected in the bone structures within the sections.</p> <p>Impression: Minimal emphysematous changes in both lungs. Millimetric nodules in both lungs.</p>	CT-CHAT
	<p>Findings: Trachea, both main bronchi are open. Mediastinal main vascular structures, heart contour, size are normal. Thoracic aorta diameter is normal. Pericardial effusion-thickening was not observed. Thoracic esophagus calibration was normal and no significant tumoral wall thickening was detected. No enlarged lymph nodes in prevascular, pre-paratracheal, subcarinal or bilateral hilar-axillary pathological dimensions were detected. When examined in the lung parenchyma window; Aeration of both lung parenchyma is normal and no nodular or infiltrative lesion is detected in the lung parenchyma. Pleural effusion-thickening was not detected. Upper abdominal organs included in the sections are normal. No space-occupying lesion was detected in the liver that entered the cross-sectional area. Bilateral adrenal glands were normal and no space-occupying lesion was detected. Bone structures in the study area are natural. Vertebral corpus heights are preserved.</p> <p>Impression: Thoracic CT examination within normal limits.</p>	Merlin
	<p>Findings: Trachea and both main bronchi are open. No occlusive pathology was detected in the trachea and both main bronchi. There is minimal bronchiectasis in the central parts of both lungs. Occasionally, linear atelectasis was observed in both lungs. There are emphysematous changes in both lungs. In the right lung, there are millimetric nodules with ground-glass areas around some of them. When evaluated together with the patient's primary disease, these appearances were primarily evaluated in favor of metastases. It is recommended that the patient be evaluated together with previous examinations, if any. There is no mass or infiltrative lesion in both lungs. Mediastinal structures cannot be evaluated optimally because contrast material is not given. As far as can be observed, the heart is larger than normal. There is no pleural or pericardial effusion. There are millimetric atheroma plaques in the aorta. No pathologically enlarged lymph nodes were detected in the mediastinum and hilar regions. There is no pathological wall thickness increase in the esophagus within the sections. No upper abdominal free fluid collection was detected in the sections. No enlarged lymph nodes in pathological dimensions were detected. There are sometimes millimetric hypodense lesions in the bone structures within the sections. Although the described appearances cannot be characterized because they are very small, it was thought that the presence of primary disease could indicate metastases of these appearances. Further investigation is recommended.</p>	CT2Rep

Figure 2: Example of radiology report generation for the same 3D chest CT scan using our BTB3D method (both 16×16×8 and 8×8×8 variants) compared to baseline models (CT-CHAT, Merlin, CT2Rep) and the ground truth report. Key phrases from the ground truth are highlighted and matched across model outputs using consistent colors to indicate alignment. Our BTB3D framework, especially the higher compression rate variant, produces more detailed, clinically relevant, and accurate radiology reports, showing superior coverage of anatomical structures and abnormalities.

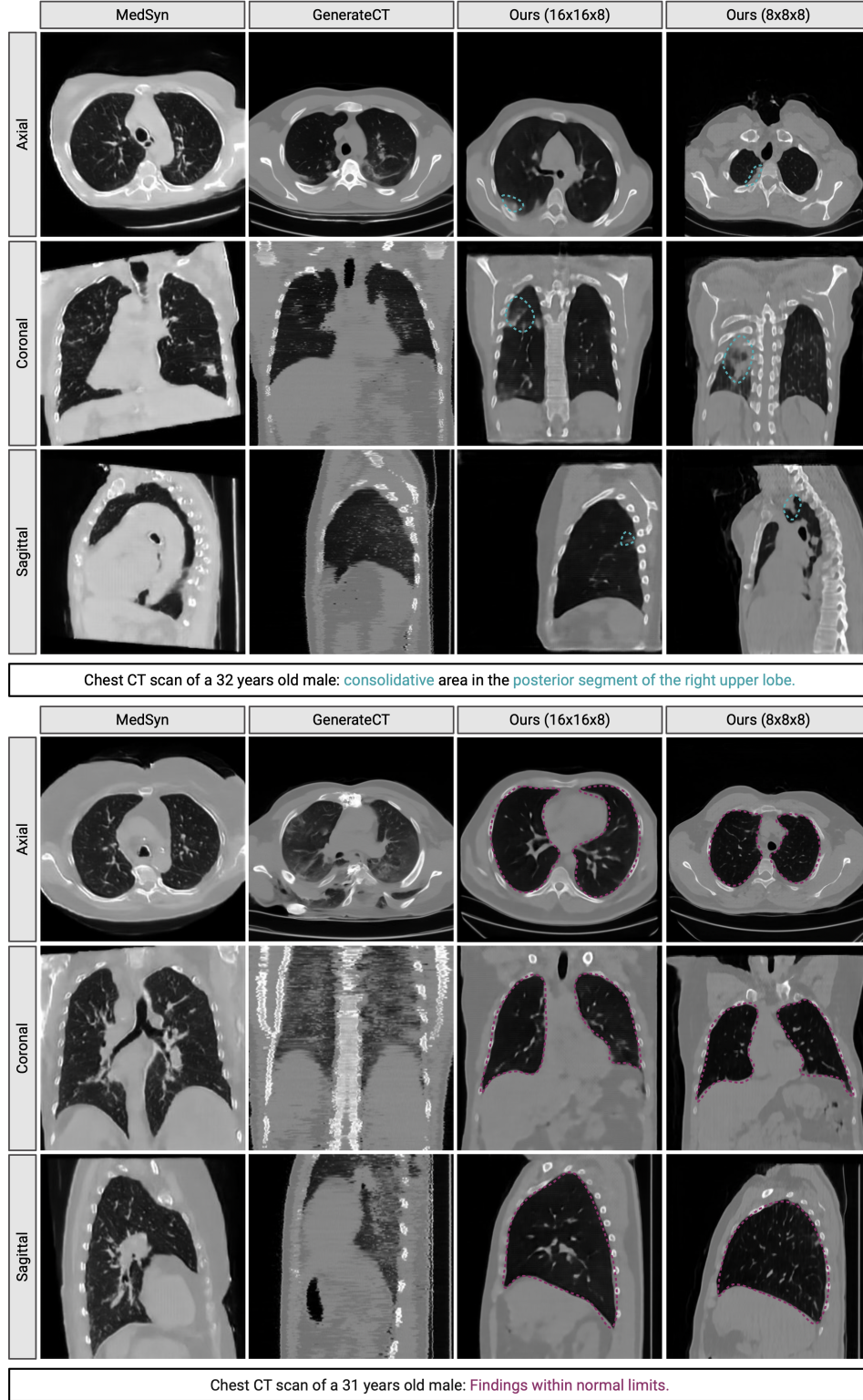


Figure 3: Text-conditioned CT generation results for two clinical prompts using MedSyn, GenerateCT, and our BTB3D models. For each case, we show one representative slice per anatomical plane. The lower-compression variant of BTB3D produces the most consistent volumes, demonstrating superior alignment with the prompt. Prompts and corresponding anatomical regions are highlighted using color overlays. Ground-truth volumes are omitted, following standard practice in generative tasks.