PromptMedCT: A Training-Free LLM-Guided Framework for Volumetric CT Data Realization with Quantifiable Disease Manifestations and Multi-Tier Annotations

Muhammad Owais [1](✉), Muhammad Shafay [1], Divya Velayudhan [1], Muhammad Zubair [2], Naveed Syed [3], Naoufel Werghi [1], Quatullah Rustum [4], and Irfan Hussain [1](✉)

[1] Khalifa University Center for Autonomous Robotic Systems (KU-CARS), Khalifa University, Abu Dhabi, United Arab Emirates
{muhammad.owais,irfan.hussain}@ku.ac.ae
[2] School of Computer Science, University of Galway, H91 TK33, Galway, Ireland
[3] Cleveland Clinic Abu Dhabi, Abu Dhabi, United Arab Emirates
[4] Sheikh Shakhbout Medical City, Abu Dhabi, United Arab Emirates

Abstract. Explainable foundation models in radiology are fundamentally limited by the lack of datasets with structured multi-tier annotations (*i.e.,* descriptive prompts, infection/lobar masks, and bounding boxes), which are critical for learning clinically grounded temporal, causal, and diagnostic reasoning. While accurate data realization offers a potential solution, existing data generation approaches either rely on large-scale supervised training or generate raw data without the structured multi-tier annotations. To overcome these limitations, we propose *PromptMedCT*, the first LLM-guided, training-free disease realization framework that bridges clinical semantics with anatomical constraints to synthesize lung CT scans with multi-tier annotations. The proposed framework can be operated in two modes: a prompt mode for high-fidelity synthesis (as a domain-specific GPT) and a parametric mode for large-scale data generation with multi-tier annotations. Using the proposed *PromptMedCT* framework (with parametric mode), we construct a dataset comprising *10K 2D CT scans* with comprehensive multi-tier annotations. We validate the realism and effectiveness of our framework through a two-step validation strategy that combines comprehensive quantitative and qualitative evaluations. Quantitatively, we perform synthetic-to-real generalization by training a vision–language model on our generated data and testing it on real clinical datasets, achieving 59.22% DICE and 42.46% IoU. For further qualitative evaluation, we perform a medical expert review in which specialists assess a set of generated data and annotations, with 92% of scans rated realistic and 79.5% of the prompts accurately described. Finally, a qualitative comparison against five state-of-the-art domain-specific and general LLM-based generators shows that our method consistently outperforms them in visual realism and multi-tier annotation quality. Additional implementation details and data resources are available at GitHub.

Keywords: PromptMedCT · Medical Image Synthesis · Training-free.

1 Introduction

The rapid advancement of AI-driven medical imaging has transformed radiology, enabling automated disease staging and computer-aided diagnosis that increasingly approach expert-level interpretation [24,2,18]. Recent foundation models, including vision-language models (VLMs) and large language models (LLMs), exhibit strong multimodal reasoning capabilities driven by language-guided supervision [28,4,6]. However, their practical impact in radiology remains constrained by the scarcity of clinically structured image-text data. In real clinical workflows, such as report generation, lesion-level interpretation, and disease staging, models require rich multi-tier annotated datasets that capture disease evolution and support clinically grounded reasoning. The absence of such structured data resources limits the development of deployable systems with significant performance [7,25,1].

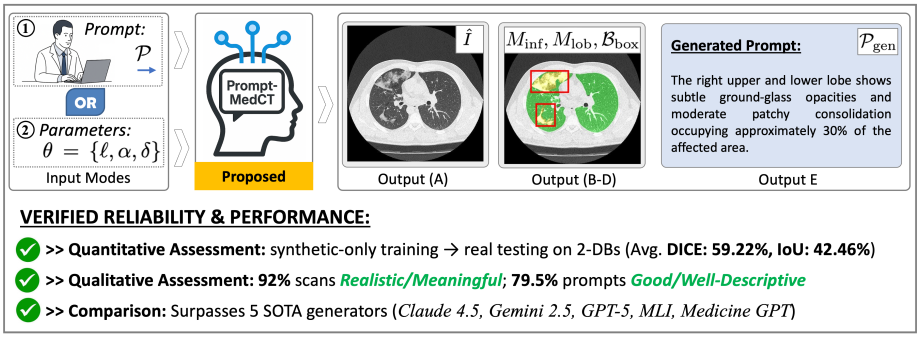

Fig. 1. Overview of *PromptMedCT*, a training-free LLM-guided diffusion framework that generates anatomically consistent lung CT scans with multi-tier annotations.

The manual generation of such datasets is expensive, time-consuming, and restricted by privacy regulations, significantly limiting large-scale data sharing [32,26]. These constraints are particularly critical in lung CT imaging, where accurate lesion localization and quantitative assessment directly influence prognosis, staging, and treatment planning. Synthetic data generation has therefore evolved as a promising alternative. Existing approaches include GAN-based synthesis [15], diffusion models [27,23], and training-free methods [30]. More recent text-conditioned diffusion frameworks enable report-guided 3D CT generation (MedSyn [29]), image-mask co-generation through cross-attention (MedSegFactory [19]), multi-conditioned pathology modeling (mDDPM [17]), free-text CT synthesis (Text2CT [12]), and report-conditioned latent diffusion (Report2CT [3]).

Despite advances in semantic alignment and anatomical realism, three critical limitations still persist: i) reliance on large-scale training data to learn generative priors, which limits scalability in privacy-constrained domains; ii) limited explicit parametric control over lesion characteristics such as shape, size, and radiodensity;

Table 1. Comparison of recent medical imaging data generation frameworks. ✓= supported; ✗= absent; TF: Training-Free. Inputs denote controllability (Parametric, Text), outputs denote annotation depth (Mask, BBox, Description, Modality).

Reference	TF	Input (Control)		Output			
		Param.	Text	Mask	BBox	Desc.	Modality
Yu *et al.*, *Biomed. Opt. Ex.*, 2021 [30]	✓	✗	✗	✗	✗	✗	2D Fundus (Retina)
Pinaya *et al.*, *MICCAI*, 2022 [23]	✗	✓	✗	✗	✗	✗	3D Brain MRI
Dorjsembe *et al.*, *IEEE-EMBC*, 2024 [9]	✗	✗	✗	✓	✗	✗	2D Endoscopic Polyp
Xu *et al.*, *IEEE-TMI*, 2024 [29]	✗	✗	✓	✓	✗	✗	3D Lung CT
Krishna *et al.*, *Phys. Med. Biol.*, 2025 [17]	✗	✓	✗	✓	✗	✗	3D Lung CT
Mao *et al.*, *ICCV*, 2025 [19]	✗	✗	✓	✓	✗	✗	2D/3D Med. Img. + Mask
Guo *et al.*, *arXiv*, 2025 [12]	✗	✗	✓	✗	✗	✗	3D Chest CT
Amirrajab *et al.*, *arXiv*, 2025 [3]	✗	✗	✓	✗	✗	✓	3D Chest CT
Nazir *et al.*, *ICCV*, 2025 [21]	✗	✗	✓	✓	✗	✗	2D Med. Img. (Inpainting)
Jiang *et al.*, *IEEE-TBE*, 2025 [13]	✗	✗	✗	✓	✗	✗	3D Thoracic CT
PromptMedCT (*Ours*)	✓	✓	✓	✓	✓	✓	**2D/3D Lung CT**

and iii) the absence of structured multi-tier annotations (including semantic masks, bounding boxes, and radiology-aligned prompts) required for supervised learning and vision-language alignment. To our knowledge, no prior framework unifies training-free generation, structured semantic prompting, and complete multi-tier annotation within a single controllable pipeline.

To address this gap, we introduce *PromptMedCT*, a unified, training-free, LLM-guided diffusion framework that integrates natural-language semantic prompting with explicit parametric control to generate anatomically consistent, clinically interpretable lung CT scans together with structured multi-tier annotations. Figure 1 illustrates the overall framework. Table 1 compares recent frameworks with our *PromptMedCT* in terms of controllability, annotation depth, and generative scope. Our key contributions are as follows:

1. We present *PromptMedCT*, the first LLM-driven, training-free framework to generate clinically meaningful, structurally accurate lung CT scans along with multi-tier annotations, effectively addressing data scarcity in medical radiology (see *Section* 2 and Supplementary Material: GitHub).
2. Using the proposed framework, we construct a dataset comprising *10K 2D CT scans* with comprehensive multi-tier annotations (*i.e.,* descriptive prompts, infection/lobar masks, and bounding boxes).
3. We employ a dual-stage validation strategy combining quantitative synthetic-to-real benchmarking and expert-driven qualitative evaluation, along with comparisons to generative and vision–language models (see Table 2, Fig. 3, and Fig. 4).

2 Proposed Method

We introduce *PromptMedCT*, a training-free framework for clinically controllable lung CT data synthesis with structured multi-tier annotations (Fig. 2). The pipeline consists of four stages: 1) *LLM-Guided Semantic Parsing*, which converts

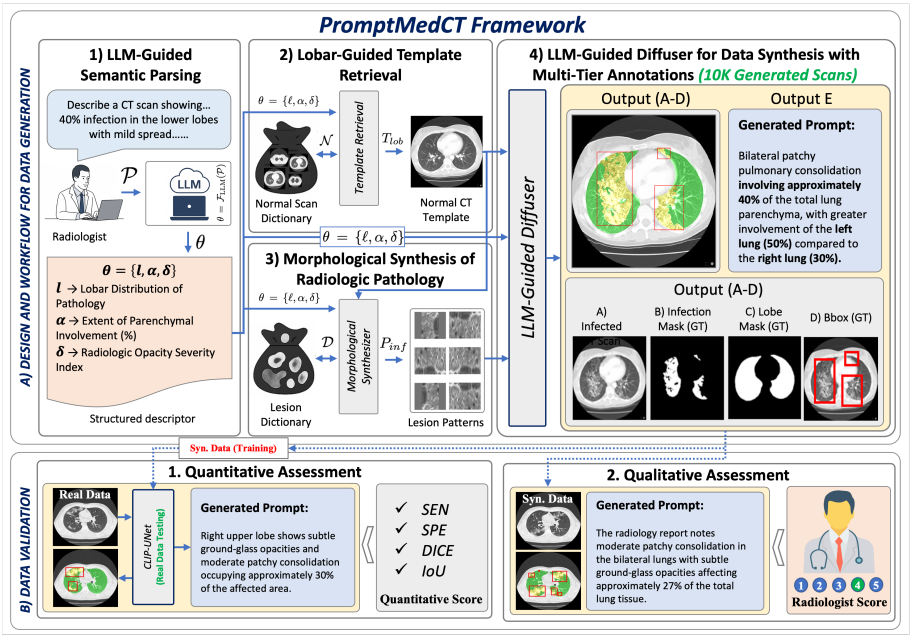

Fig. 2. Overall design and workflow of the *PromptMedCT* framework for synthesis of clinically controllable CT data with structured multi-tier annotations.

radiological prompts into structured disease descriptors; 2) *Lobar-Guided Template Retrieval*, selecting anatomically aligned healthy CT slices; 3) *Morphological Synthesis*, retrieving and deforming lesion masks from a visual dictionary; and 4) *LLM-Guided Diffuser*, embedding lesions into CT templates while generating multi-level annotations. We detail each stage below.

1) LLM-Guided Semantic Parsing: Given a radiological prompt $\mathcal{P}$, we extract a structured disease descriptor $\theta = \{\ell, \alpha, \delta\}$ capturing three key attributes: i) lobar distribution $\ell \in \{L^z, R^z, B^z\}$, where laterality (left/right/bilateral) is combined with zonal position $z \in \{u, m, l\}$ (upper/middle/lower); ii) fractional involvement $\alpha \in [0, 1]$, representing the proportion of affected lung; and iii) opacity severity $\delta \in \mathbb{R}^+$, modeling lesion intensity and spatial diffusion.

These attributes allow explicit control over lesion location, span, and visual severity during synthesis. To extract θ, we apply *LLaMA3* [11], adapted with radiologist-defined templates for domain-specific interpretation. The model maps the input prompt $\mathcal{P}$ (such as *"Describe a lung CT scan presenting 40% infection in the lower lobes with mild spread"*) to $\theta = \mathcal{F}_{\text{LLM}}(\mathcal{P})$, creating the semantic backbone that maps clinical language to parameterized imaging data generation.

2) Lobar-Guided Template Retrieval: Given the structured descriptor $\theta = \{\ell, \alpha, \delta\}$, we select an anatomically consistent healthy CT slice from a normal scan dictionary $\mathcal{N} = \{T_j\}_{j=1}^{M}$ (Fig. 2). The dictionary is pre-organized into lobar subsets $\mathcal{N}_\ell$ corresponding to the upper, middle, and lower zones. Template

selection is based on the availability of ℓ:

$$T_{\text{lob}} = \begin{cases} \mathcal{R}(\ell, \mathcal{N}), & \text{if } \ell \text{ is defined,} \\ \mathcal{U}(\mathcal{N}), & \text{otherwise,} \end{cases} \tag{1}$$

where $\mathcal{R}$ identifies the template T_{lob} that maximizes spatial consistency with the defined lobar region, and $\mathcal{U}$ represents uniform random sampling. This strategy preserves anatomical coherence when lobar cues are present, while also preserving variability for under-specified prompts.

3) Morphological Synthesis of Radiologic Pathology: With the given $\theta = \{\ell, \alpha, \delta\}$ and the extracted template T_{lob}, the morphological synthesizer generates anatomically consistent lesion patterns (Fig. 2). A simple segmentation model *LOBE-Net*$(\cdot)$ (based on DeepLabV3+ [8]) generates the lobar mask M_{lob} using the extracted CT template T_{lob} to define the target synthesis region (*i.e.,* $M_{\text{lob}} = \textit{LOBE-Net}(T_{\text{lob}})$). A lesion dictionary $\mathcal{D} = \{P_i\}_{i=1}^{N}$ contains prototype infection patterns. The synthesizer retrieves the prototype whose area best matches the desired coverage $\alpha|M_{\text{lob}}|$:

$$P_{\text{inf}} = \arg\min_{P_i \in \mathcal{D}} \left| A(P_i) - \alpha |M_{\text{lob}}| \right|, \tag{2}$$

where $A(P_i)$ is the area of the pixels of patch P_i. The selected prototype P_{inf} is geometrically matched (translation, rotation, scaling) to M_{lob}, and its opacity is modulated by δ to approximate radiologic severity. If a single prototype is insufficient, multiple patches are iteratively selected to approximate the target coverage:

$$\{P_{\text{inf}}^{(k)}\}_{k=1}^{n} = \arg\min_{\{P_i\} \subset \mathcal{D}} \left| \sum_{k=1}^{n} A(P_{\text{inf}}^{(k)}) - \alpha |M_{\text{lob}}| \right|. \tag{3}$$

This module operationalizes semantic descriptors into pixel-level pathology while preserving anatomical coherence and progression consistency.

4) LLM-Guided Diffuser for Data Synthesis with Multi-Tier Annotations: This stage forms the generative core of *PromptMedCT*. It integrates the semantic descriptor θ, anatomical template T_{lob}, and synthesized lesion prototype P_{inf}, collectively $\phi = \{\theta, T_{\text{lob}}, P_{\text{inf}}\}$. The output comprises a pathological CT slice $\hat{I}$ and structured annotations $\psi = \{M_{\text{inf}}, M_{\text{lob}}, \mathcal{B}_{\text{box}}, \mathcal{P}_{\text{gen}}\}$.

Within the lobar region M_{lob}, the lesion prototype is spatially deformed via a hybrid rigid-elastic warp $\tilde{M}(i,j) = [\mathcal{T}_{\text{warp}}(P_{\text{inf}}; \gamma, \alpha, \delta)](i,j)\, \mathbf{1}_{M_{\text{lob}}}(i,j)$, where γ controls geometric deformation. Optimal parameters are obtained by:

$$\gamma^* = \arg\min_{\gamma} \left| A(\tilde{M}) - \alpha |M_{\text{lob}}| \right| + \lambda \mathcal{R}(\gamma), \tag{4}$$

ensuring target coverage with smooth anatomical consistency. Radiologic attenuation and soft boundaries are simulated by diffusing the warped mask $G = \tilde{M} * \mathcal{K}_{\sigma(\delta)}$, with $\sigma(\delta) = \sigma_0 + k\delta$, and modulating contrast via $\beta(\delta) = \beta_0 + h\delta$. The synthesized CT image becomes $\hat{I}(i,j) = T_{\text{lob}}(i,j)\big(1 - \beta(\delta) G(i,j)\big)$. The

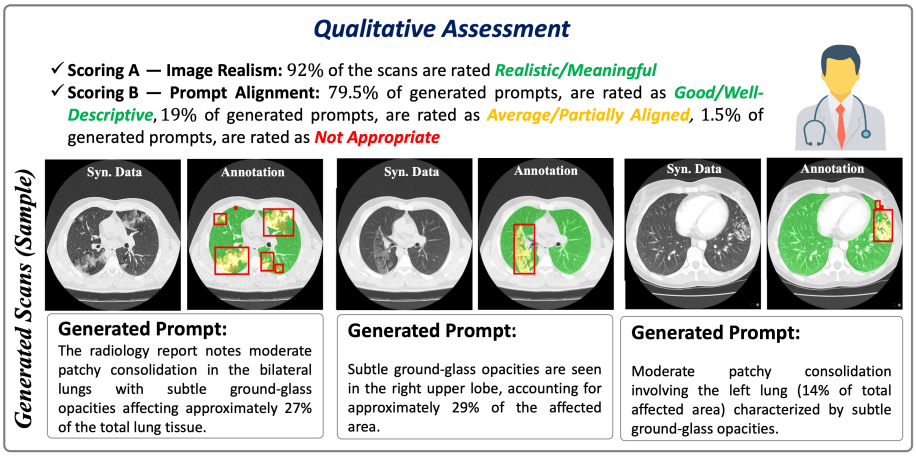

Fig. 3. Illustrative examples of synthetically generated 2D CT scans with corresponding multi-tier annotations along with expert-based evaluation scores.

infection mask is obtained by thresholding $M_{\text{inf}}(i,j) = \mathbf{1}[G(i,j) > \tau_s]$, and $\mathcal{B}_{\text{box}}$ is computed as the tight axis-aligned bounding box of M_{inf}. Finally, a radiologist-style report is generated as $\mathcal{P}_{\text{gen}} = \mathcal{F}_{\text{LLM}}(\theta, M_{\text{lob}}, M_{\text{inf}})$. Overall, the diffuser performs the controlled transformation $(\hat{I}, \psi) = \mathcal{F}_{\text{Diff}}(\phi)$, bridging semantic intent and anatomically consistent lung CT data realization with full pixel- and text-level supervision.

3 Results

1) Experimental Setup and Evaluation Protocol: We employ a dual-stage quantitative and qualitative evaluation to assess the realism and effectiveness of our data generation framework. At first, we construct a dataset of *10K 2D CT scans* with multi-tier annotations (DB-1) using the proposed *PromptMedCT* framework under parametric mode. For quantitative evaluation, we perform synthetic-to-real generalization by training on generated data (DB-1) and testing on two real-world benchmarks: MedSeg COVID-19 (DB-2) [20] (100 annotated CT slices) and COVID-19 CT Seg (DB-3) [14] (367 annotated slices), directly measuring cross-domain generalization. For qualitative evaluation, we conduct expert assessment to verify diagnostic plausibility and perform a comparison with state-of-the-art generators to assess visual realism, controllability, and annotation fidelity.

2) Quantitative Assessment: Table 2 presents synthetic-to-real benchmarking results. CLIP-based models trained exclusively on the *PromptMedCT* synthetic dataset (DB-1) demonstrate strong cross-domain generalization across two real-world CT benchmarks (DB-2 and DB-3), achieving an average of 59.22% DICE and 42.46% IoU. The maintained sensitivity, DICE, and IoU scores, together

Table 2. Quantitative synthetic-to-real generalization assessment using the generated *PromptMedCT* dataset for training and two real clinical datasets for testing.

Model (Backbone)	DB-2: MedSeg COVID-19 (Real)				DB-3: COVID-19 CT Seg (Real)			
	SEN (%)	SPE (%)	DICE (%)	IoU (%)	SEN (%)	SPE (%)	DICE (%)	IoU (%)
CLIP-UNet (ViT-B/32) [31]	57.85	95.90	54.38	37.34	49.22	98.01	35.13	21.31
CLIP-UNet (ViT-L/14) [31]	63.39	97.98	66.56	49.89	70.02	98.49	51.89	35.04

with high specificity (>95%), demonstrate that disease representations learned from the proposed synthetic data effectively transfer to real clinical lung CT scans. These findings indicate that *PromptMedCT* preserves clinically meaningful structural and radiological characteristics, including progression-aware lesion patterns, enabling downstream segmentation without exposure to real training data.

3) Qualitative Assessment: To assess the diagnostic plausibility of our synthetic CT data, we conducted a subjective evaluation independently reviewed by two medical professionals and verified by a senior radiologist using a standardized scoring protocol. The reviewed set spans diverse progression-aware infection patterns, including variability in lobar involvement, extent, and opacity severity (see Fig. 3). The assessment comprised two criteria:

- **Scoring A - Image Realism:** Scans were rated as *Realistic/Meaningful* or *Synthetic/Not Good* based on anatomical plausibility and radiological consistency.
- **Scoring B - Prompt Alignment:** The corresponding LLM-generated clinical description was rated on a 3-point scale [*1: Not Appropriate*, *2: Average/Partially Aligned*, or *3: Good/Well-Descriptive*] reflecting its agreement with visual findings.

The qualitative assessment followed a dual-axis scoring rubric that assessed both visual realism and prompt alignment. For *Image Realism*, 92% of evaluated data samples were rated as *Realistic/Meaningful*, indicating strong anatomical plausibility and diagnostic coherence across the generated CT scans. For *Prompt Quality*, 79.5% of LLM-generated descriptions were rated *Good/Well-Descriptive*, 19% *Average/Partially Aligned*, and only 1.5% *Not Appropriate*. These results confirm that *PromptMedCT* produces visually credible, progression-consistent CT scans together with semantically aligned, radiologically coherent textual descriptions.

Table 3. Radiologist-driven evaluation of SOTA methods and proposed *PromptMedCT*.

Model	Realism ↑	Annotation Quality ↑	Rank	Realism (%)	Time (sec)	Remarks
Claude 4.5 [5]	Low	Low	1	–	–	Inconsistent anatomy
Gemini 2.5 [10]	Low–Moderate	Low	2	–	–	Weak structure
GPT-5 [16]	Moderate	Low–Moderate	3	–	–	Limited alignment
MLI [30]	Moderate	Low	4	–	–	No pixel-level/text alignment
Medicine GPT [22]	High	Moderate	5 (Second-best)	60%	9–15	Partial annotations
PromptMedCT (Ours)	**Very High**	**High**	**6 (Best)**	**92% (+32%)**	**1–3 / 1–2**	**Structured multi-tier + explicit localization**

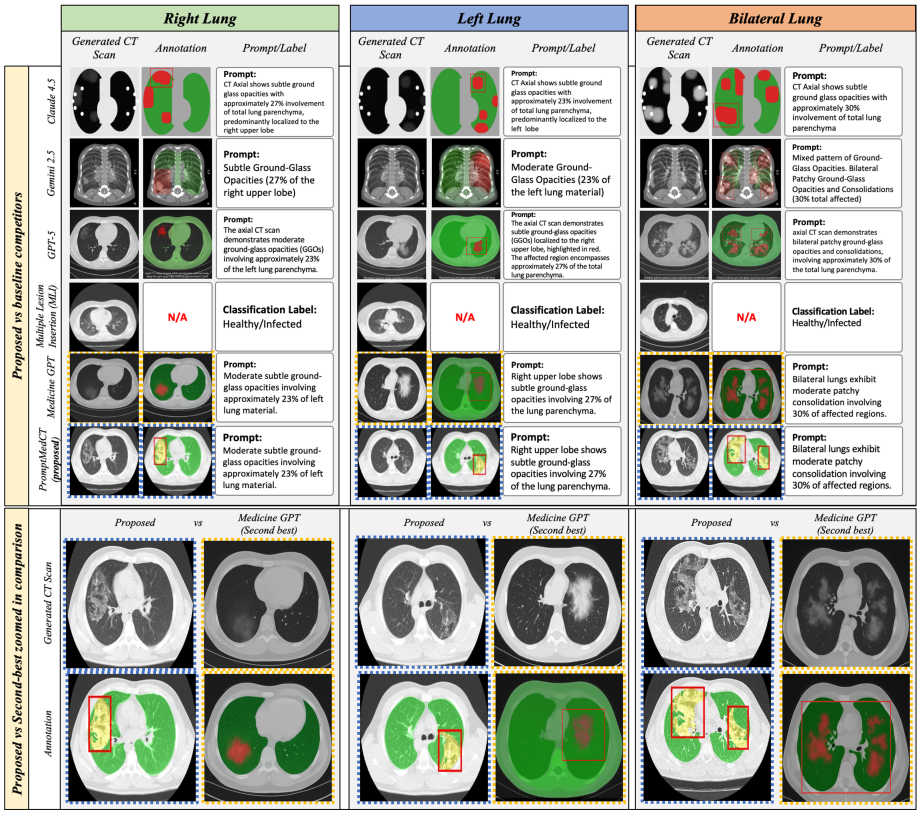

Fig. 4. Visual comparison of synthetic CT scans generated by Claude Sonnet [5], Gemini [10], GPT-5 [16], MLI [30], Medicine GPT [22], and our *PromptMedCT*, all conditioned on the same input prompt.

4) Comparison: Finally, Fig. 4 shows a qualitative comparison between the proposed *PromptMedCT* and representative SOTA general-purpose and medical VLMs (Claude 4.5 [5], Gemini 2.5 [10], GPT-5 [16], Medicine GPT [22]), as well as the training-free medical generator named Multiple Lesion Insertion (MLI) [30]. For evaluation, a set of representative samples was generated for these models under the same prompt conditions and assessed through expert-driven qualitative analysis to ensure a fair comparison (see Table 3 for more details). GAN- and diffusion-based medical synthesis models (as listed in Table 1) are excluded, as they require supervised training and are not directly comparable in a training-free framework. The expert review study highlights the following key observations: i) general-purpose LLMs struggled to generate anatomically consistent axial CT slices, often producing projection-like images or semantically inconsistent lesion patterns. ii) MLI [30] generated visually plausible scans but provided only class-level supervision without pixel-wise or text-level annotations. iii) GPT-5 and Medicine GPT [22] produced better outputs; however, structured multi-tier

annotations and precise pathology-text alignment were absent. iv) *PromptMedCT* generates anatomically coherent CT scans with explicit lesion localization and multi-tier outputs, outperforming Medicine GPT [22] (second-best) with 92% realism preference vs. 60% (+32%). It also reduces generation time from 9-15 sec/sample (in case of Medicine GPT [22]) to 1-3 sec (prompt mode) and 1-2 sec (parametric mode), enabling efficient and scalable structured data generation (Table 3).

4 Conclusion

We introduce *PromptMedCT*, a unified, training-free, LLM-guided diffusion framework that integrates natural-language semantic prompting with explicit parametric control to generate anatomically consistent, clinically interpretable lung CT scans together with structured multi-tier annotations. Using our proposed framework, we constructed a dataset of *10K 2D CT scans* with comprehensive multi-tier annotations and validated its realism and effectiveness through a two-step strategy combining quantitative and qualitative evaluations. Synthetic-to-real benchmarking demonstrates meaningful cross-domain generalization, while expert evaluation confirms the anatomical and diagnostic plausibility of the generated data.

Acknowledgments. This work was supported in part by the Khalifa University Center for Autonomous Robotic Systems (KU-CARS); and in part by the Advanced Research and Innovation Center (ARIC), which was jointly funded by Mubadala United Arab Emirates (UAE) Clusters and Khalifa University of Science and Technology. The authors also acknowledge the use of OpenAI's ChatGPT for language refinement and readability improvement during manuscript preparation.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., Sontag, D.: Large language models are few-shot clinical information extractors. In: Proceedings of the 2022 conference on empirical methods in natural language processing. pp. 1998–2022 (2022)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
3. Amirrajab, S., Salahuddin, Z., Kuang, S., Woodruff, H.C., Lambin, P.: Radiology report conditional 3d ct generation with multi encoder latent diffusion model. arXiv preprint arXiv:2509.14780 (2025)
4. Aneja, A., Dhull, A., Singh, A., Singh, K.K.: Large language model on multi-modal data. In: Multimodal Generative AI, pp. 63–78. Springer (2025)
5. Anthropic: Introducing the next generation of claude (2024), https://www.anthropic.com/news/claude-sonnet-4-5

6. Bayoudh, K., Knani, R., Hamdaoui, F., Mtibaa, A.: A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. The Visual Computer **38**(8), 2939–2970 (2022)
7. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: European conference on computer vision. pp. 1–21. Springer (2022)
8. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
9. Dorjsembe, Z., Pao, H.K., Xiao, F.: Polyp-ddpm: Diffusion-based semantic polyp synthesis for enhanced segmentation. In: 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 1–7. IEEE (2024)
10. Gemini Team Google: Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
11. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
12. Guo, P., Zhao, C., Yang, D., He, Y., Nath, V., Xu, Z., Bassi, P.R., Zhou, Z., Simon, B.D., Harmon, S.A., et al.: Text2ct: Towards 3d ct volume generation from free-text descriptions using diffusion model. arXiv preprint arXiv:2505.04522 (2025)
13. Jiang, Y., Lemaréchal, Y., Bafaro, J., Abi-Rjeile, J., Joubert, P., Després, P., Manem, V.: Lung-ddpm: Semantic layout-guided diffusion models for thoracic ct image synthesis. IEEE Transactions on Biomedical Engineering (2025)
14. Jun, M., Cheng, G., Yixin, W., Xingle, A., Jiantao, G., Ziqi, Y., Minqing, Z., Xin, L., Xueyuan, D., Shucheng, C., et al.: Covid-19 ct lung and infection segmentation dataset. https://zenodo.org/records/3757476 (2020)
15. Kazeminia, S., Baur, C., Kuijper, A., Van Ginneken, B., Navab, N., Albarqouni, S., Mukhopadhyay, A.: Gans for medical image analysis. Artificial intelligence in medicine **109**, 101938 (2020)
16. Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., et al.: Chatgpt: Jack of all trades, master of none. Information fusion **99**, 101861 (2023)
17. Krishna, A., Wang, G., Mueller, K.: Guided synthesis of annotated lung ct images with pathologies using a multi-conditioned denoising diffusion probabilistic model (mddpm). Physics in Medicine & Biology **70**(6), 065007 (2025)
18. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
19. Mao, J., Wang, Y., Tang, Y., Xu, D., Wang, K., Yang, Y., Zhou, Z., Zhou, Y.: Medsegfactory: Text-guided generation of medical image-mask pairs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21525–21535 (2025)
20. MedSeg, Jenssen, H., Sakinis, T.: Medseg covid dataset 1 (2021). https://doi.org/10.6084/m9.figshare.13521488.v2
21. Nazir, M., Aqeel, M., Setti, F.: Diffusion-based data augmentation for medical image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1330–1339 (2025)
22. OpenAI: Medicine gpt. https://chatgpt.com/g/g-rgWJmXVhn-medicine-gpt (2026), accessed: 2026-05-12

23. Pinaya, W.H., Tudosiu, P.D., Dafflon, J., Da Costa, P.F., Fernandez, V., Nachev, P., Ourselin, S., Cardoso, M.J.: Brain imaging generation with latent diffusion models. In: MICCAI workshop on deep generative models. pp. 117–126. Springer (2022)
24. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
25. Selivanov, A., Rogov, O.Y., Chesakov, D., Shelmanov, A., Fedulova, I., Dylov, D.V.: Medical image captioning via generative pretrained transformers. Scientific Reports **13**(1), 4171 (2023)
26. Soltanayev, S., Chun, S.Y.: Training deep learning based denoisers without ground truth data. Advances in neural information processing systems **31** (2018)
27. Wolleb, J., Bieder, F., Sandkühler, R., Cattin, P.C.: Diffusion models for medical anomaly detection. In: International Conference on Medical image computing and computer-assisted intervention. pp. 35–45. Springer (2022)
28. Wu, S., Fei, H., Qu, L., Ji, W., Chua, T.S.: Next-gpt: Any-to-any multimodal llm. In: Forty-first International Conference on Machine Learning (2024)
29. Xu, Y., Sun, L., Peng, W., Jia, S., Morrison, K., Perer, A., Zandifar, A., Visweswaran, S., Eslami, M., Batmanghelich, K.: Medsyn: text-guided anatomy-aware synthesis of high-fidelity 3-d ct images. IEEE Transactions on Medical Imaging **43**(10), 3648–3660 (2024)
30. Yu, Z., Yan, R., Yu, Y., Ma, X., Liu, X., Liu, J., Ren, Q., Lu, Y.: Multiple lesions insertion: boosting diabetic retinopathy screening through poisson editing. Biomedical Optics Express **12**(5), 2773–2789 (2021)
31. Zhao, H., Wang, B., Mistry, D., Wang, J., Dohopolski, M., Yang, D., Lu, W., Jiang, S., Nguyen, D.: Medical image segmentation assisted with clinical inputs via language encoder in a deep learning framework. Machine Learning: Science and Technology **6**(1), 015040 (2025)
32. Zhu, X., Vondrick, C., Fowlkes, C.C., Ramanan, D.: Do we need more training data? International Journal of Computer Vision **119**(1), 76–92 (2016)