Retrieval-Augmented Anatomical Guidance for Text-to-CT Generation

Daniele Molino[1*], Camillo Maria Caruso[1], Paolo Soda[1,2], and Valerio Guarrasi[1]

[1] Unit of Artificial Intelligence and Computer Systems, Department of Engineering, Università Campus Bio-Medico di Roma, Rome, Italy
daniele.molino@unicampus.it, camillomaria.caruso@unicampus.it, valerio.guarrasi@unicampus.it, p.soda@unicampus.it

[2] Department of Diagnostics and Intervention, Biomedical Engineering and Radiation Physics, Umeå University, Umeå, Sweden
paolo.soda@umu.se

Abstract. Text-conditioned generative models for volumetric medical imaging provide semantic control but lack explicit anatomical guidance, often resulting in outputs that are spatially ambiguous or anatomically inconsistent. In contrast, structure-driven methods ensure strong anatomical consistency but typically assume access to ground-truth annotations, which are unavailable when the target image is to be synthesized. We propose a retrieval-augmented approach for Text-to-CT generation that integrates semantic and anatomical information under a realistic inference setting. Given a radiology report, our method retrieves a semantically related clinical case using a 3D vision-language encoder and leverages its associated anatomical annotation as a structural proxy. This proxy is injected into a text-conditioned latent diffusion model via a ControlNet branch, providing coarse anatomical guidance while maintaining semantic flexibility. Experiments on the CT-RATE dataset show that retrieval-augmented generation improves image fidelity and clinical consistency compared to text-only baselines, while additionally enabling explicit spatial controllability, a capability inherently absent in such approaches. Further analysis highlights the importance of retrieval quality, with semantically aligned proxies yielding consistent gains across all evaluation axes. This work introduces a principled and scalable mechanism to bridge semantic conditioning and anatomical plausibility in volumetric medical image synthesis. Code is available at https://github.com/arco-group/RAGText2CT.

Keywords: 3D Medical Image Synthesis · Text-to-CT · Generative Models · Retrieval-Augmented Generation · Anatomical Guidance

1 Introduction

Artificial Intelligence (AI) is increasingly integrated into medical imaging workflows [1,18], yet its large-scale deployment remains constrained by limited an-

* Corresponding author.

notated data, privacy regulations, and high labeling cost [3,8,27]. In this context, Generative AI offers a promising solution for synthesizing realistic medical data to support data augmentation, simulation, and privacy-aware learning [5,7,14,26]. Among imaging modalities, Computed Tomography (CT) plays a central clinical role by providing high-resolution volumetric representations [20]. However, volumetric generation poses significant challenges in terms of computational scalability and global anatomical coherence. A prominent research direction in this field is *text-conditioned* generation, where radiology reports guide the synthesis process. Radiological narratives provide high-level semantic descriptions derived from expert interpretation, enabling clinically aligned controllability. Nevertheless, reports are inherently underspecified with respect to spatial structure: they describe pathological findings but do not encode explicit anatomical constraints and omit large portions of normal anatomy. As a result, text-only conditioning may produce semantically plausible yet spatially ambiguous or anatomically inconsistent outputs. Early text-to-CT approaches, such as GenerateCT [11] and MedSyn [29], rely on multi-stage pipelines combining low-resolution synthesis with super-resolution refinement, which may introduce inter-slice inconsistencies. More recently, the growing interest in report-conditioned 3D CT generation, further catalyzed by the introduction of dedicated benchmarking efforts [12], has led to diffusion-based architectures that can operate directly in compact volumetric latent spaces [22,2]. In parallel, *structure-driven* methods, such as MAISI [9], condition synthesis on explicit spatial inputs, e.g., segmentation masks. While this paradigm enables precise anatomical control, it lacks semantic expressiveness and relies on ground-truth annotations at inference time, which are unavailable when the volume itself must be synthesized. This contrast reveals a key limitation of existing methods: text-based conditioning provides semantic flexibility without explicit spatial control, whereas mask-based conditioning ensures anatomical precision at the cost of semantic richness and realistic inference assumptions. Bridging these paradigms therefore requires leveraging anatomical information without observing or segmenting the target volume itself. To this end, we here introduce a Retrieval-Augmented Generation (RAG) [4,15,17] formulation extended to a multimodal volumetric setting that permits us to reinterpret anatomical structure as a *retrievable latent proxy* rather than a direct conditioning input. Indeed, rather than assuming explicit access to structural annotations, we treat anatomical information as a latent source that can be approximated by retrieving relevant related examples from previously observed data. In practice, given an input report, a pretrained 3D vision-language encoder retrieves semantically related clinical cases from a reference corpus. The associated anatomical annotations are employed as coarse structural proxies that provide informative spatial constraints. The retrieved proxy is then injected into a text-conditioned latent diffusion model via a ControlNet [30] branch, guiding synthesis toward anatomically coherent solutions while preserving semantic variability induced by the report. Intuitively, the retrieved anatomy serves as a spatial scaffold rather than a precise template, allowing the generative process to adapt fine-grained structure while enforcing global anatomical coherence.

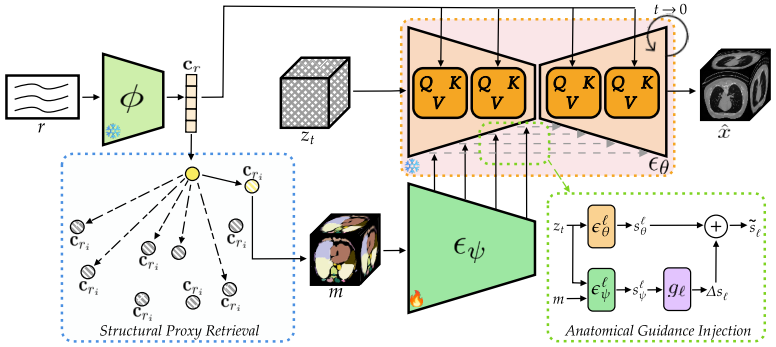

Fig. 1: The proposed retrieval-augmented Text-to-CT generation framework.

In summary, the contributions of this work are threefold:

- We propose a retrieval-augmented framework for report-conditioned 3D CT synthesis that treats anatomical structure as a latent, retrievable proxy.
- We introduce a multimodal integration strategy that injects retrieved anatomical proxies into a text-conditioned latent diffusion model via ControlNet, enabling anatomical guidance without annotations at inference time.
- We provide extensive quantitative and qualitative evaluations of image fidelity, clinical consistency, and spatial controllability, and analyze the impact of retrieval quality on generation performance.

2 Methods

Problem Formulation We consider the task of generating a CT volume $\hat{x}$ conditioned on a radiology report r, assuming no access to the corresponding volume x or structural annotations at inference time. We therefore model anatomical structure as a latent source of information, named as *retrieved structural proxy* and denoted as m in the following, that can be approximated through retrieval. It is worth noting that in our formulation m is not assumed to match the target anatomy, but it acts as a plausible spatial scaffold biasing the generative process toward anatomically coherent solutions. Therefore, given the condition provided by the semantic (r) and coarse anatomical (m) guidance, the generation process is defined as:

$$\hat{x} \sim p_\theta(x \mid r, m)$$

where p_θ denotes a generative model. The overall retrieval-augmented generation framework is illustrated in Figure 1.

Text-to-CT Generative Backbone We adopt a latent diffusion model for volumetric CT synthesis [22,25] operating in a compressed latent space obtained through a variational autoencoder (VAE) [16]. Diffusion is performed in the

latent domain, and the final CT volume is reconstructed via the VAE decoder. Text conditioning is implemented through embeddings extracted from a vision-language model based on the CLIP paradigm [24], extended with a 3D image encoder. This formulation enables the alignment of radiology reports and CT volumes in a shared semantic embedding space, providing effective grounding of clinical language into volumetric representations [21,23]. The resulting report embeddings are used to guide generation according to the semantic content of the input text. In this work, the diffusion backbone and text-conditioning mechanism are kept fixed, and we focus on augmenting the model with an anatomically informed conditioning pathway that does not alter the pretrained generative architecture.

Retrieval-Augmented Structural Proxy To approximate structural information in the absence of explicit anatomical inputs, we introduce a retrieval-based mechanism to obtain a structural proxy. Given a report r, we retrieve a semantically related case from a reference corpus using a pretrained 3D vision-language encoder [22] pre-trained on CT-RATE, denoted as $\phi(\cdot)$. From the retrieved case, we extract the associated anatomical annotation (e.g., a segmentation mask), which is used as the structural proxy m. Formally, let $\mathbf{c}_r = \phi(r) \in \mathbf{R}^d$ be the embedding of the input report, given a reference set of reports $\{r_i\}_{i=1}^N$ with embeddings $\{\mathbf{c}_{r_i} = \phi(r_i)\}_{i=1}^N$, the structural proxy is obtained as:

$$m = \mathcal{M}\Big(\arg\max_i \ \mathrm{sim}(\mathbf{c}_r, \mathbf{c}_{r_i})\Big)$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity and $\mathcal{M}(\cdot)$ is a deterministic operator that returns the anatomical annotation associated with the retrieved case. Semantic similarity in the shared embedding space is hypothesized to correlate with coarse pathological and anatomical patterns, making the retrieved annotation a suitable, albeit noisy, structural proxy. We retrieve the nearest neighbor to provide a deterministic and unambiguous signal and further analyze the effect of retrieval quality on generation performance in Section 4.

Anatomical Guidance via ControlNet To integrate the retrieved structural proxy into the generative process, we inject anatomical guidance through a dedicated control branch that preserves the pretrained generative flow, thereby maintaining report-driven semantic variability while enforcing global anatomical consistency via the retrieved proxy. To this end, we adopt a ControlNet-based conditioning mechanism [30], where we introduce a trainable control branch ϵ_ψ alongside the frozen diffusion backbone ϵ_θ, mirroring its encoder architecture and weights. Let z_t denote the noisy latent at diffusion step t, $\mathbf{c}_r$ the report embedding, and m the structural proxy. During the forward pass of the frozen backbone, the encoder produces multi-scale skip features $\{s_\theta^\ell\}_{\ell=0}^L$ together with a bottleneck representation b_θ,

$$(\{s_\theta^\ell\}, b_\theta) = E_\theta(z_t, t, \mathbf{c}_r)$$

which are used to predict the noise component in the decoder. In parallel, the ControlNet branch processes the same noisy latent and semantic conditioning, augmented with the structural proxy m, yielding corresponding control features

$$(\{s_\psi^\ell\}, b_\psi) = E_\psi(z_t, t, \mathbf{c}_r, m)$$

The control features are not injected directly into the backbone activations. Instead, we map s_ψ^ℓ and b_ψ through zero-initialized convolutions to produce residual corrections

$$\Delta s_\ell = \gamma\, g_\ell(s_\psi^\ell) \qquad \Delta b = \gamma\, g_{\text{mid}}(b_\psi)$$

where $g_\ell(\cdot)$ and $g_{\text{mid}}(\cdot)$ denote zero-initialized projections and γ is a scalar conditioning scale. We then inject the residuals into the skip connections and the bottleneck of the frozen backbone,

$$\tilde{s}_\ell = s_\theta^\ell + \Delta s_\ell \qquad \tilde{b} = b_\theta + \Delta b$$

and the final noise prediction is obtained by decoding the modulated features

$$\hat{\epsilon} = D_\theta(\{\tilde{s}_\ell\}, \tilde{b}, t, \mathbf{c}_r)$$

Zero-initialization ensures $\Delta s_\ell \approx 0$ and $\Delta b \approx 0$ at the start of training, recovering the original pretrained generator. During training, the diffusion backbone and the vision-language encoder are kept frozen, and only the parameters of E_ψ and the zero-initialized projection layers $g_\ell(\cdot)$ and $g_{\text{mid}}(\cdot)$ are optimized conditioning on ground-truth annotations via the rectified flow objective [19]. At inference time, the proxy m is obtained exclusively through retrieval, enabling anatomically informed generation without access to explicit annotations.

3 Experimental Configurations

Datasets and Data Preparation Experiments are conducted on the CT-RATE dataset [10], comprising paired 3D chest CT volumes and radiology reports covering 18 thoracic pathologies. We follow the official train/test split provided by the dataset, comprising 27,514 training volumes and 1,818 test volumes. All CT volumes are resampled to $0.75 \times 0.75 \times 1.5$ mm and resized to $512 \times 512 \times 128$ voxels. Intensity values are converted to Hounsfield Units, clipped to $[-1000, 1000]$, and normalized to $[0, 1]$. Anatomical segmentation masks extracted using TotalSegmentator [28] are available for all volumes in the repository. Retrieval is performed exclusively over the training set; test set reports are never included in the retrieval index and test set masks are never provided as conditioning input, ensuring no information leakage at evaluation time.

Compared Methods We compare our approach against both text-conditioned and structure-conditioned generative models. In the former group, we include GenerateCT [11] and MedSyn [29], representative multi-stage pipelines, as well as diffusion-based report-conditioned models Text-to-CT [22] and Report2CT [2].

These baselines enable isolating the effect of anatomical guidance. While in the latter, we include MAISI [9], which conditions directly on ground-truth segmentation masks. Since MAISI does not incorporate textual conditioning, it is not directly comparable in terms of semantic controllability. Instead, it serves as an upper bound, illustrating the level of anatomical consistency achievable when exact annotations are available at inference time. To ensure fair comparison, all baselines were retrained on CT-RATE following their official repositories, guaranteeing a uniform training distribution, spatial resolution, and preprocessing pipeline. Additionally, to assess the role of retrieval quality, we perform an ablation study evaluating three strategies for selecting the structural proxy: semantically nearest, semantically farthest, and random retrieval in the shared vision-language embedding space. All configurations share the same generative backbone, differing exclusively in the proxy selection criterion at inference time. This ablation isolates the impact of semantic alignment between the report and the retrieved structural proxy.

Evaluation Metrics We evaluate generative performance along three complementary axes: image fidelity, clinical consistency, and spatial controllability. *Image Fidelity.* Visual realism is assessed using FID score [13]. We report both slice-based FID computed on axial, coronal, and sagittal views using a 2D medical backbone, and volumetric FID using a 3D medical backbone.

Clinical Consistency. Clinical plausibility is evaluated using CT-Net [6], a 3D CNN trained on real CTs and kept frozen during evaluation. Classification performance on synthetic CTs measures alignment with the conditioning reports.

Spatial Controllability. Spatial controllability is evaluated using Dice score and 95^{th} percentile Hausdorff Distance (HD95). Segmentation masks are predicted from generated volumes using TotalSegmentator [28] and compared against the corresponding reference masks. Dice quantifies volumetric overlap, while HD95 captures boundary-level geometric deviations, providing a complementary assessment of structural adherence to the provided mask.

Implementation Details All components are implemented in PyTorch. Control features are injected at all downsampling and bottleneck layers. The ControlNet conditioning scale is fixed to 1.0 for all experiments, selected via a sweep over $[0.5, 2.0]$ balancing structural adherence and semantic flexibility, while text conditioning is applied using classifier-free guidance with a guidance scale of 5.0. Sampling is performed using rectified flow with 30 steps. The ControlNet branch is trained for 50 epochs using AdamW with learning rate of 1×10^{-5} and weight decay 1×10^{-2}. Training uses a batch size of 4, distributed across two NVIDIA A100 GPUs, and is performed with mixed-precision arithmetic.

4 Results

This section presents a comprehensive evaluation of the proposed retrieval-augmented Text-to-CT approach, with the goal of isolating the contribution of the retrieved structural proxy across three complementary evaluation axes: image fidelity, clinical consistency, and spatial controllability. * in tables denotes

Table 1: FID scores computed using 2.5D and 3D feature extractors.

Method	FID 2.5D ↓				FID 3D ↓
	Axial	Coronal	Sagittal	Avg.	
GenerateCT [11]	6.701*	11.793*	9.694*	9.396*	0.166*
MedSyn [29]	8.789*	8.900*	8.717*	8.802*	0.169*
Report2CT [2]	1.345*	2.294*	1.855*	1.775*	0.043*
Text-to-CT [22]	0.500*	0.519*	0.561*	0.527*	0.015*
MAISI [9]	0.462*	0.568*	0.503*	0.511*	0.012*
RAG-Farthest	0.298*	0.399*	0.340*	0.346*	0.010*
RAG-Random	0.271	0.343*	0.341*	0.311*	0.005
RAG-Nearest	0.275	0.317	0.318	0.303	0.004

Table 2: Clinical consistency evaluation of text-conditioned models using CT-Net.

Model	AUC ↑		Precision ↑	
	Macro	Weighted	Macro	Weighted
GenerateCT [11]	0.581*	0.568*	0.285*	0.361*
MedSyn [29]	0.560*	0.547*	0.282*	0.362*
Report2CT [2]	0.720*	0.705*	0.436*	0.508*
Text-to-CT [22]	0.745*	0.731*	0.477*	0.549*
RAG-Farthest	0.712*	0.716*	0.429*	0.520*
RAG-Random	0.729*	0.728*	0.460*	0.545*
RAG-Nearest	0.787	0.776	0.535	0.606

Table 3: Dice score and HD95 computed between reference masks and masks predicted from generated CTs.

Model	DICE ↑	HD95 ↓
Real	0.847	1.872
MAISI [9]	0.792	2.811
RAG-Farthest	0.697*	5.447*
RAG-Random	0.710*	5.391*
RAG-Nearest	0.772	3.072

statistically significantly inferior performance compared to the proposed method according to a paired Wilcoxon signed-rank test across 5 independent runs with different random seeds ($p < 0.05$).

Image Fidelity Table 1 reports FID results in both 2.5D and 3D settings. Across all configurations, introducing a structural proxy improves image fidelity with respect to text-only generation. Notably, RAG variants also outperform MAISI; we attribute this to the absence of semantic conditioning in MAISI, which produces anatomically consistent volumes that are nonetheless misaligned with the semantic distribution of test reports, which FID directly penalizes. All retrieval-augmented variants achieve lower FID scores than competing methods, indicating improved global anatomical coherence. Semantically nearest retrieval yields the most stable improvements, while random and farthest retrieval result in slightly degraded FID scores. This suggests that retrieval quality has a modest effect on low-level appearance statistics, while playing a more prominent role in higher-level semantic and anatomical alignment.

Clinical Consistency Table 2 reports classification performance on synthetic CT volumes generated by the different methods. For reference, CT-Net achieves an AUC of 0.824 when evaluated on real CT volumes, providing an upper-bound performance estimate. MAISI is not included in this evaluation, as it conditions solely on segmentation masks and does not incorporate report-based seman-

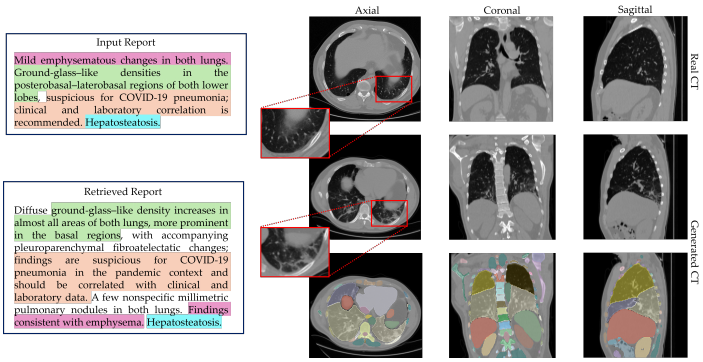

Fig. 2: Qualitative example of retrieval-augmented generation. Left: input report and retrieved report, with overlapping clinical concepts highlighted. Right: real CT and the generated CT with and without the retrieved anatomical proxy m.

tic guidance. Retrieval-augmented generation improves diagnostic performance with respect to text-only baselines, indicating better preservation of clinically meaningful patterns. Semantically nearest retrieval achieves the strongest performance, highlighting the importance of semantic alignment between the report and the retrieved structural proxy. Random and farthest retrieval degrade performance, confirming the sensitivity of clinical realism to retrieval quality.

Spatial Controllability Table 3 reports Dice and HD95, measuring the degree to which each method adheres to the provided structural conditioning signal. This evaluation is restricted to methods incorporating structural conditioning, as comparing text-only methods against a reference mask would only measure accidental overlap. Specifically, for RAG variants, generated volumes are segmented and compared against the retrieved proxy, measuring spatial controllability as adherence to the provided scaffold, rather than anatomical correctness with respect to unknown ground-truth. MAISI conditions directly on the target volume's own ground-truth mask, giving it direct structural access to the case being synthesized and making it an oracle upper bound rather than a directly comparable approach. Our method requires no mask at inference; the structural proxy is obtained exclusively via retrieval from the input report. Real CT volumes are reported to establish a reference for the segmentation pipeline. RAG-Nearest approaches MAISI's structural adherence while preserving semantic flexibility. Notably, a model trivially copying the proxy would score high here but degrade on Table 2, confirming that the evaluation axes are complementary to assess generation quality.

Qualitative Comparison Figure 2 illustrates a representative example. The retrieved case exhibits semantic overlap with the input report, and the generated CT adheres to the spatial prior provided by the proxy mask. In particular, the retrieved anatomical scaffold constrains the global thoracic layout, reducing spatial ambiguity while allowing local variations consistent with the report.

5 Conclusion

We introduced a retrieval-augmented framework for Text-to-CT generation that bridges semantic and anatomical conditioning without requiring annotations at inference time. Anatomy is modeled as a retrievable proxy in a shared 3D vision-language embedding space, improving coherence while preserving report-driven variability. Experiments on CT-RATE demonstrate consistent gains in image fidelity, clinical consistency, and spatial controllability. Future work will investigate pathology-specific evaluation and longitudinal scenarios, leveraging temporally related priors to model disease progression.

Acknowledgments. Daniele Molino is a Ph.D. student enrolled in the National Ph.D. in Artificial Intelligence, XL cycle, course on Health and Life Sciences, organized by Università Campus Bio-Medico di Roma. This work was partially founded by: i) Università Campus Bio-Medico di Roma under the program "University Strategic Projects" within the project "AI-powered Digital Twin for next-generation lung cancEr cAre (IDEA)"; ii) PNRR MUR project PE0000013-FAIR. Resources are provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) and the Swedish National Infrastructure for Computing (SNIC) at Alvis @ C3SE, partially funded by the Swedish Research Council through grant agreements no. 2022-06725 and no. 2018-05973.

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

References

1. Alyasseri, Z.A.A., et al.: Review on COVID-19 diagnosis models based on machine learning and deep learning approaches. Expert systems **39**(3), e12759 (2022)
2. Amirrajab, S., et al.: Radiology Report Conditional 3D CT Generation with Multi Encoder Latent diffusion Model. arXiv preprint arXiv:2509.14780 (2025)
3. Bansal, M.A., Sharma, D.R., Kathuria, D.M.: A systematic review on data scarcity problem in deep learning: solution and applications. ACM Computing Surveys (Csur) **54**(10s), 1–29 (2022)
4. Chen, J., et al.: Benchmarking large language models in retrieval-augmented generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 17754–17762 (2024)
5. Chlap, P., et al.: A review of medical image data augmentation techniques for deep learning applications. Journal of medical imaging and radiation oncology **65**(5), 545–563 (2021)
6. Draelos, R.L., et al.: Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. Medical image analysis **67**, 101857 (2021)
7. Frangi, A.F., Tsaftaris, S.A., Prince, J.L.: Simulation and synthesis in medical imaging. IEEE transactions on medical imaging **37**(3), 673–679 (2018)
8. Guo, P., et al.: Multi-institutional collaborations for improving deep learning-based magnetic resonance image reconstruction using federated learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2423–2432 (2021)

9. Guo, P., et al.: Maisi: Medical ai for synthetic imaging. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4430–4441. IEEE (2025)
10. Hamamci, I.E., et al.: Developing generalist foundation models from a multimodal dataset for 3d computed tomography. arXiv preprint arXiv:2403.17834 (2024)
11. Hamamci, I.E., et al.: Generatect: Text-conditional generation of 3d chest ct volumes. In: European Conference on Computer Vision. pp. 126–143. Springer (2024)
12. Hamamci, I.E., et al.: Challenge for Vision-Language Modeling in 3D Medical Imaging (VLM3D) (Mar 2025). https://doi.org/10.5281/zenodo.15052708
13. Heusel, M., et al.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
14. Kazerouni, A., et al.: Diffusion models in medical imaging: A comprehensive survey. Medical image analysis **88**, 102846 (2023)
15. Khandelwal, U., et al.: Nearest neighbor machine translation. arXiv preprint arXiv:2010.00710 (2020)
16. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
17. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems **33**, 9459–9474 (2020)
18. Litjens, G., et al.: Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Scientific reports **6**(1), 26286 (2016)
19. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
20. Mazonakis, M., Damilakis, J.: Computed tomography: What and how does it measure? European journal of radiology **85**(8), 1499–1504 (2016)
21. Molino, D., et al.: Any-to-Any Vision-Language Model for Multimodal X-ray Imaging and Radiological Report Generation. In: 2025 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2025)
22. Molino, D., et al.: Text-to-CT Generation via 3D Latent Diffusion Model with Contrastive Vision-Language Pretraining. arXiv preprint arXiv:2506.00633 (2025)
23. Molino, D., et al.: XGeM: A multi-prompt foundation model for multimodal medical data generation. Computerized Medical Imaging and Graphics p. 102718 (2026)
24. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
25. Rombach, R., et al.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
26. Singh, N.K., Raza, K.: Medical image generation using generative adversarial networks: A review. Health informatics: A computational perspective in healthcare pp. 77–96 (2021)
27. Tajbakhsh, N., et al.: Guest editorial annotation-efficient deep learning: the holy grail of medical imaging. IEEE transactions on medical imaging **40**(10), 2526–2533 (2021)
28. Wasserthal, J., et al.: TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. Radiology: Artificial Intelligence **5**(5), e230024 (2023)
29. Xu, Y., et al.: Medsyn: Text-guided anatomy-aware synthesis of high-fidelity 3d ct images. IEEE Transactions on Medical Imaging (2024)
30. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)