CTTok: Voxel-abulary for Autoregressive 3D CT Volume Generation with Large Language Models

Jiayi Wang[1], Hadrien Reynaud[1], Mischa Dombrowski[1], Xiaoliang Wang[1], Ibrahim Ethem Hamamci[2], Suprosanna Shit[2], Sezgin Er[2,3], Bjoern Menze[2], and Bernhard Kainz[1,4]

[1] FAU Erlangen-Nürnberg, Erlangen, DE, jiayi.w.wang@fau.de
[2] Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland
[3] Istanbul Medipol University, Istanbul, Turkey
[4] Department of Computing, Imperial College London, London, UK

Abstract. Existing text-to-imaging generation methods inherit diffusion architectures from natural image synthesis, relying on explicit cross-attention to condition generation on text at every step. We hypothesize that this design is unnecessarily complex for medical tomographic volumes: unlike natural scenes, human anatomy exhibits strong structural regularity, with organs occupying predictable locations and varying far less across individuals than open-ended visual content. Given a text-aligned tokenization, sequential modeling alone should therefore suffice for text-conditional anatomical generation, without dedicated cross-attention layers. We propose CTTok, a discrete autoregressive approach that extends the vocabulary of a pre-trained language model with anatomical tokens representing 3D CT patches. Text conditioning is handled by the language model's existing causal attention over these anatomically grounded tokens. Because the synthesized volume is itself constructed from such tokens, any mild boundary artifacts from patch-level discretization carry no semantic ambiguity and can be removed by a single-pass, unconditional GAN refinement with no access to the original prompt, since all text-to-anatomy correspondence is already resolved at the token level. CTTok outperforms state-of-the-art diffusion and flow-matching baselines in diversity, image quality, and text-image alignment, at a fraction of the training and inference compute. Source code and model weights: https://github.com/WongJiayi/CTTok.

Keywords: 3D Medical Image Generation · Vision-Language Models · Autoregressive Generation · CT Synthesis

1 Introduction

While Vision-Language models for interpreting medical scan volumes have matured considerably, the reverse task of generating anatomically plausible volumes that faithfully match a textual description remains an open challenge.

Text-conditioned image synthesis addresses three practical needs. First, *data augmentation*: training downstream AI models on combined real and synthetic

volumes has been shown to improve classification performance by up to 8% over real data alone [9]. Second, *privacy-preserving data sharing*: synthetic volumes contain no patient-identifying information, bypassing HIPAA and GDPR constraints that currently bottleneck multi-institutional research [7,4]. Third, *rare-pathology simulation*: text conditioning enables on-demand generation of volumes depicting underrepresented abnormalities (*e.g.*, sub-centimetre nodules, rare tumours), directly addressing long-tail data scarcity in medical AI training [17].

Recent advances in medical imaging have increasingly adopted state-of-the-art methods from natural image processing. Large language models (LLMs) have been widely used for radiology report generation [15], while the success of generative models such as Sora [1] has promoted the use of diffusion models [11,22,20] and flow matching techniques [18] for medical image generation. For text-to-volume generation specifically, MedSyn [30] achieves high-fidelity synthesis using a U-Net architecture and GenerateCT [9] proposes a two-stage diffusion pipeline: the first stage generates low-resolution volumes with good semantic alignment, while the second stage increases resolution and the number of slices by combining text embeddings with visual embeddings during diffusion steps. Following these diffusion-based approaches, recent works have explored alternative architectures for improved efficiency. CT-Flow [28] treats CT volumes as video sequences instead of using 3D structures, adapting video flow matching models to the CT domain and generating high-fidelity volumes slice-by-slice through a single-stage process. MAISI [8] and MAISIv2 [37] use ControlNet to generate high-resolution volumes conditioned on segmentation masks, enabling users to edit organs interactively. Two key challenges persist: prohibitive generation time, since all these methods require iterative denoising or flow integration, and weak alignment between generated anatomy and the pathologies specified in the conditioning text.

Discrete tokenisation offers a fundamentally different path [34,33,10,6,32]. Prior work has demonstrated VQ-VAE-based autoregressive modeling of 3D brain MRI [25]. However, these approaches rely on *learned latent codebooks* in which every token is coupled through the decoder: altering a single token alters overlapping local neighborhoods due to convolutional coupling, preventing strict spatial disentanglement. We propose to depart from this paradigm by constructing a *frozen* codebook of anatomically interpretable atlas blocks via random-hyperplane hashing, where each token independently encodes a fixed spatial region rather than being coupled through a shared decoder. This independence enables a pre-trained LLM to learn correspondences between free-text report tokens and volume tokens through standard causal attention, without requiring dedicated cross-attention layers or iterative denoising, which is a property we verify empirically in Table 5.

In this work, we propose **CTTok**: a fast and controllable text-to-CT generation approach. Our approach consists of three main components. First, we construct a **Voxel-abulary**, *i.e.*, a visual codebook containing over 250k high-frequency $16\times16\times16$ blocks extracted from real CT volumes ($256\times256\times D$ slices) from the CT-Rate dataset [9], which includes more than 40k volumes. Second, we fine-tune Qwen-1.7B [31] to generate 1D vision token sequences from anatomi-

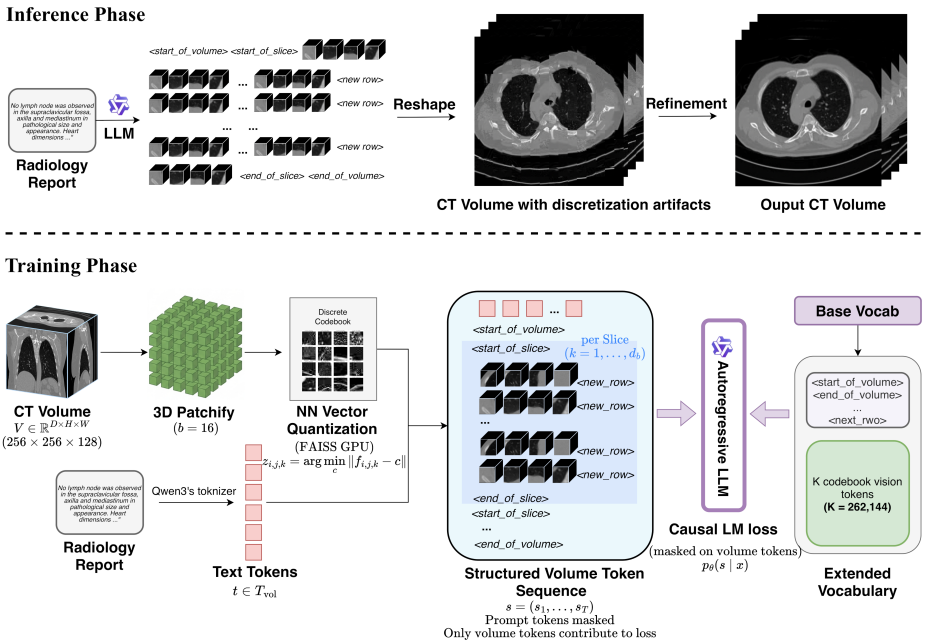

Fig. 1: **Structured Autoregressive CTTok Pipeline for Text-to-3D CT Generation.** We tokenize 3D CT volumes and radiology reports for joint autoregressive modeling. During inference, the LLM predicts vision tokens from text, which are then reshaped and refined to eliminate artifacts and produce high-fidelity 3D outputs.

cal text prompts. Using special tokens such as `<vision_begin>`, `<slice_begin>`, and `<new_row>`, the model learns to produce sequences that can be reconstructed into 3D volumes. A single-pass 2.5D GAN [12] removes residual blocking artifacts from the discrete reconstruction. The main technical contributions are:

1. A novel anatomical codebook of $16 \times 16 \times 16$ CT patches, constructed via random hyperplane hashing over a pretrained vision encoder's feature space, capturing high-frequency structural patterns across more than 40k volumes. Each token independently encodes a fixed spatial region, enabling direct correspondences between text and anatomy.
2. CTTok, a discrete autoregressive framework that reformulates text-to-CT synthesis as language modeling. Text conditioning is captured by the LLM's causal attention over a mixed text-vision sequence, removing the need for dedicated cross-attention. We verify empirically that text-to-anatomy correspondence is established at the token level, enabling a single-pass 2.5D GAN to refine residual artifacts without any text input.

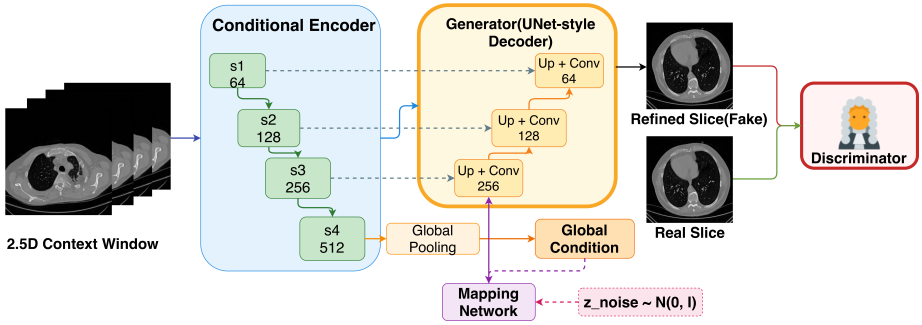

Fig. 2: **Training pipeline of the 2.5D conditional GAN.** A 9-slice axial context window conditions a UNet-style generator via multi-scale skip connections. A global condition vector from the deepest encoder features modulates the mapping network. No text input is used at any stage of this task.

3. We adapt FID, FVD[26], and IRS [3] with a CT-native feature extractor and validate that these metrics track established Inception-based metrics from StyleGAN-V [24,35].

2 Method

CTTok operates in three phases: (1) codebook construction from CT volumes, (2) training an autoregressive LLM to generate discrete volume tokens, and (3) refinement of reconstructed volumes using a conditional GAN.

Codebook Construction: We construct a codebook of $K = 262{,}144$ representative 3D patches from CT volumes. We extract non-overlapping cubic patches $p_{i,j,k} \in \{0, \ldots, 255\}^{16\times16\times16}$ from volumes padded to multiples of 16.Each patch is encoded by a vision encoder $E(\cdot)$, pretrained with a SigLIP contrastive loss [36] between text and volume, such that $f_{i,j,k} = E(p_{i,j,k}) \in \mathbb{R}^{1024}$. To efficiently scale codebook construction to many tens of thousands of volumes and several hundred thousand buckets without the memory bottleneck of K-Means clustering, we apply random hyperplane hashing. Given a random projection matrix $R \in \mathbb{R}^{1024\times B}$, we compute hash codes:

$$h_n = \sum_{m=1}^{B} \mathbb{I}[(f_n^\top R)_m > 0] \cdot 2^{m-1}.$$

We use $B = 18$ random hyperplanes, yielding $K = 2^{18} = 262{,}144$ hash buckets, which is the largest value that maintains training efficiency (see Tab. 2). One patch per bucket is randomly sampled to form the codebook $\mathcal{C} = \{(c_k, a_k)\}_{k=1}^{K}$, where c_k is the feature vector and a_k is the corresponding atlas block.

Given a textual description x, we model a 3D CT volume V as a sequence of discrete tokens generated autoregressively by a causal LLM. The causal LLM

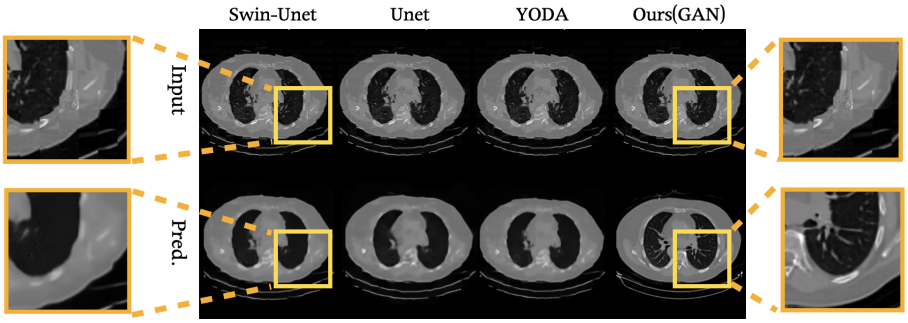

Fig. 3: **Refinement results comparison.** Top: raw discrete reconstruction with blocking artifacts. Bottom: GAN-refined output. Zoomed regions show that our method produces significantly sharper anatomical details than U-Net [23], Swin-UNet [2], and YODA [21].

model factorizes the joint distribution over volume tokens as a product of conditionals: $p(V \mid x) = \prod_{t=1}^{T} p_\theta(s_t \mid s_{<t}, x)$, where s_t denotes the t-th token in the sequence and θ are the model parameters. Text conditioning on x is handled through the model's standard causal attention over the concatenated text-vision sequence, requiring no dedicated cross-attention layers.

Volume Tokenization: Each CT volume $\mathbf{I} \in \{0, \ldots, 255\}^{D \times H \times W}$ is padded to multiples of $b = 16$ and partitioned into $N = (H/b)(W/b)(D/b)$ cubic patches $\mathbf{p}_n \in \{0, \ldots, 255\}^{16 \times 16 \times 16}$ in z-major order. We normalize patches to $[-1, 1]$ and extract features: $\mathbf{f}_n = E(\mathbf{p}_n) \in \mathbb{R}^{1024}$. Discrete tokens are assigned via nearest-neighbor quantization over the codebook:

$$z_n = \arg\min_k \|\mathbf{f}_n - \mathbf{c}_k\|_2^2, \quad z_n \in \{0, \ldots, K-1\}.$$

Retrieval is accelerated using FAISS GPU FlatL2 indexing [5].

Structured Serialization: The discrete volume $\mathbf{Z} \in \{0, \ldots, K-1\}^{(D/b) \times (H/b) \times (W/b)}$ is serialized with special tokens marking volume, slice, and row boundaries. Given text x, we construct: $x \parallel$ ``Size:[h_b, w_b, d_b]'' $\parallel$ structure markers $\parallel$ voxel tokens. Voxel tokens are offset by v_{offset} to avoid collision with the base vocabulary: $\hat{z}_n = z_n + v_{\text{offset}}$.

Training: We fine-tune the causal LLM using a standard next-token prediction loss restricted to volume tokens:

$$\mathcal{L} = -\sum_{t \in \mathcal{T}_{\text{vol}}} \log p_\theta(s_t \mid s_{<t}),$$

so the model learns to predict anatomy while the text prefix is treated as given context. We use bf16 mixed precision with Muon [14] for Transformer weights and Adam [16] for all remaining parameters.

Table 1: Quantitative comparison with state-of-the-art conditional generation methods. Conditioning: MaisiV2 [37] (segmentation masks), GenerateCT [9] (metadata), others (medical reports from CT-RATE [9]). Average inference time per volume on A100. CTTok evaluated at 10k steps, temperature 0.65, top-K selection by CLIP score.
†MaisiV2 achieves perfect IRS due to explicit segmentation mask conditioning.

		CT-Feature-Extractor			Inception			
Model	**Cond.**	IRS ↑	FID_{train} ↓	FID_{valid} ↓	FID ↓	FVD_{16} ↓	CLIP Score ↑	Time ↓
MaisiV2 [37]	Masks	100.0%†	$14.95_{\pm 5.99}$	–	116.75	973.43	–	34*s*
MedSyn [30]	Report	14.39%	$14.74_{\pm 6.22}$	$14.70_{\pm 6.21}$	104.07	1354.37	$28.29_{\pm 2.23}$	100*s*
GenerateCT [9]	Report	9.47%	$21.15_{\pm 5.96}$	$21.19_{\pm 5.94}$	**39.12**	1556.14	$10.03_{\pm 1.41}$	184*s*
CTFlow [28]	Report	16.56%	$27.12_{\pm 6.69}$	$27.25_{\pm 6.68}$	119.05	1342.10	$33.75_{\pm 1.52}$	99*s*
CTTok (ours) top@1	Report	**18.48%**	$\mathbf{9.02_{\pm 6.23}}$	$\mathbf{8.91_{\pm 6.21}}$	42.68	905.31	$29.78_{\pm 1.99}$	**2.2s**
CTTok (ours) top@2	Report	16.25%	$9.42_{\pm 6.16}$	$9.30_{\pm 6.14}$	44.11	**842.38**	$\mathbf{43.11_{\pm 2.22}}$	4.4*s*

Inference: At inference, we autoregressively sample volume tokens conditioned on text. Tokens are parsed using structural markers and assembled into a 3D volume by indexing the precomputed pixel atlas $A \in \{0, \ldots, 255\}^{K \times 16 \times 16 \times 16}$.
2.5D Refinement via Conditional GAN: The discrete reconstruction may exhibit blocking artifacts due to the fixed patch size. We compared several refinement architectures, including U-Net [23], Swin-UNet [2], and YODA [21], all of which produced blurry results (Fig. 3). We therefore employ a 2.5D conditional GAN (R3GAN [12]) that refines the volume slice-by-slice. For each axial slice z, we extract a temporal window of $K = 9$ slices: $\mathbf{w}_z = \mathbf{V}_{rec}[z-4 : z+4] \in \mathbb{R}^{9 \times H \times W}$, which serves as condition for refining the center slice.

A conditional encoder E_{cond} extracts multi-scale features $\{s_1, s_2, s_3, s_4\}$ from $\mathbf{w}_z$. These features are injected into generator G via skip connections. A global condition vector $\mathbf{c}_{vec}$ built from global pooling s_4 modulates the mapping network. The refined slice is: $\hat{\mathbf{I}}_z = G(\mathbf{z}_{noise}, \mathbf{c}_{vec}, \{s_l\}_{l=1}^{4})$, where $\mathbf{z}_{noise} \sim \mathcal{N}(0, I)$.

In Fig. 2 the discriminator D assesses realism given context $\mathbf{w}_z$. The generator loss combines adversarial and L_1 reconstruction terms:

$$\mathcal{L}_G = \mathcal{L}_{adv}(G, D) + 100 \|\hat{\mathbf{I}}_z - \mathbf{I}_{gt,z}\|_1.$$

We use relativistic adversarial loss [13] with R1 gradient penalties for stability. At inference, we process the volume in a sliding-window manner.

3 Experiments and Results

Dataset: We use the CT-RATE [9,27] dataset, which aggregates scans from multiple vendors and protocols with slice thicknesses ranging from 0.5mm to 5mm (10× variation). The training set contains over 40k complete 3D scans and the validation set includes over 3k scans, making it the largest publicly available CT dataset. Each volume is paired with an aligned radiology report and TotalSegmentator [29] segmentation masks. We perform text-to-volume generation at 256×256×128 resolution conditioned on the reports.

Table 2: Performance across codebook sizes ($K = 2^B$) as mean$_{\pm\text{std}}$ over 1000 slices. Inference time is measured per slice on a single A100. $B = 18$ is selected to maintain high image quality while ensuring training efficiency.

2^B	MSE ↓	MAE ↓	PSNR ↑	SSIM ↑	Time (s) ↓
13	$381.45_{\pm 135.82}$	$9.36_{\pm 2.40}$	$22.57_{\pm 1.48}$	$0.94_{\pm 0.02}$	**0.41**
14	$321.65_{\pm 110.93}$	$8.65_{\pm 2.21}$	$23.29_{\pm 1.44}$	$0.95_{\pm 0.02}$	0.67
15	$290.75_{\pm 98.44}$	$8.14_{\pm 2.12}$	$23.73_{\pm 1.45}$	$0.96_{\pm 0.05}$	1.17
16	$262.44_{\pm 90.20}$	$7.62_{\pm 2.05}$	$24.15_{\pm 1.45}$	$0.96_{\pm 0.05}$	2.24
18	$200.79_{\pm 73.16}$	$6.25_{\pm 1.83}$	$25.34_{\pm 1.55}$	$0.97_{\pm 0.01}$	8.29
19	$194.80_{\pm 71.37}$	$6.17_{\pm 1.82}$	$25.51_{\pm 1.56}$	$0.97_{\pm 0.01}$	13.08

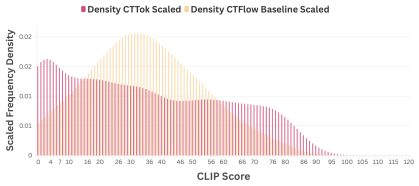

Fig. 4: CLIP score distribution for CTTok (temp. 0.6 top@1) and CT-Flow. Both methods show comparable text–image alignment (∼34.0).

Implementation: Training proceeds in two phases (Fig. 1): LLM fine-tuning for 2k steps (384 GPU hours on H200/H100) and GAN refinement (100 GPU hours on A100). We use mixed optimization with Muon [14] (lr=0.02) for Transformer weights and Adam (lr=0.005, $\beta = (0.9, 0.95)$) for other parameters. Inference takes 2.2s on an A100 with 4.7GB memory. Evaluation metrics include: IRS [3] and FID for diversity and quality, Inception FID/FVD [35] for validation and inter-slice continuity, and CLIP scores with original and fine-tuned CT-CLIP [9] for text alignment.

Table 3: Sampling temperature ablation: IRS, quality, and text alignment.

Temp.	IRS ↑	FID_{valid} ↓	CLIP Score ↑
0.6	16.5%	$10.75_{\pm 6.20}$	$\mathbf{33.69}_{\pm 2.13}$
0.7	20.1%	$7.86_{\pm 6.52}$	$27.50_{\pm 1.93}$
0.8	22.3%	$\mathbf{7.44}_{\pm \mathbf{6.32}}$	$19.86_{\pm 2.41}$
1.0	**26.1**%	$7.88_{\pm 6.52}$	$12.53_{\pm 1.72}$
1.2	25.7%	$11.33_{\pm 6.79}$	$13.65_{\pm 1.38}$

Table 4: Top-K ablation. We generate K samples and select the one with the highest CLIP score for evaluation.

top-K	IRS ↑	FID_{valid} ↓	CLIP Score ↑
1	**18.48**%	$\mathbf{8.91}_{\pm 6.21}$	$29.78_{\pm 1.99}$
2	16.25%	$9.30_{\pm 6.14}$	$43.11_{\pm 2.22}$
3	15.11%	$9.43_{\pm 6.10}$	$50.06_{\pm 2.33}$
5	13.54%	$9.64_{\pm 6.02}$	$\mathbf{59.11}_{\pm \mathbf{2.30}}$

Diversity vs Alignment: As shown in Table 1, diversity (IRS) and alignment (CLIP) exhibit an inverse relationship: our model demonstrates superior diversity using top-1 selection, achieving 11.6% improvement over the most diverse text-conditioned baseline (CT-Flow). MaisiV2 achieves perfect IRS (1.0) due to explicit segmentation mask conditioning, which supports the validity of the metric, and its near-zero FID_{valid} is expected given this strong structural guidance. Using the original CT-CLIP backbone, all methods show relatively low CLIP scores. The high variance observed in CLIP scores reflects the original CT-CLIP backbone's limited sensitivity to fine-grained textual differences across methods; adopting the vocabulary-adapted CT-Vocab [9] backbone resolves this, yielding stable and discriminative scores across all models. Our model achieves the highest CLIP score of 43.11 using top-2 selection, compared to CT-Flow's 34.5, representing a 24.9% improvement. Notably, CT-Flow directly uses CT-CLIP's

Table 5: Disentangling LLM and GAN contributions. We compare generation quality before (Raw) and after (Denoised) the GAN refinement.

	CLIP ↑	$\text{FID}_{\text{train}}$ ↓	$\text{FID}_{\text{valid}}$ ↓	FID ↓	FVD_{16} ↓
Raw (pre-GAN)	**48.46**$_{\pm 0.99}$	**15.96**$_{\pm 6.04}$	**15.66**$_{\pm 6.02}$	155.82$_{\pm 0.84}$	**1510.20**$_{\pm 22.66}$
Denoised (post-GAN)	47.26$_{\pm 0.81}$	16.15$_{\pm 5.91}$	15.91$_{\pm 5.90}$	**95.53**$_{\pm 0.68}$	1652.28$_{\pm 23.53}$

text encoder for preprocessing, providing an inherent advantage in CLIP calculation, whereas our model employs a different tokenizer yet demonstrates strong performance by leveraging the LLM's text understanding capabilities. Additionally, Fig. 4 shows that our model generates more samples with scores exceeding 60, though this is not reflected in the average.

Image Quality & Inter-slice Continuity: All variants of our model achieve lower FID_{train} than baselines, indicating superior image quality. Specifically, we improve upon the previous best (MedSyn) by approximately 50%. Using the StyleGAN-V Inception backbone, our models achieve the lowest and second-lowest FID and FVD scores, demonstrating both high image quality and natural inter-slice continuity. Additionally, as shown in Fig. 1, our CT-specific FID follows the same trend as the baseline StyleGAN-V FID across training steps, validating the reliability of our proposed metric.

Ablation Study: Table 3 shows that higher temperature trades alignment for diversity, with the optimal setting depending on training stage. Table 4 confirms the diversity-alignment trade-off: higher CLIP scores yield lower IRS. Notably, top-5 selection achieves CLIP scores up to 59 while maintaining consistent FID, demonstrating effective quality-alignment balance. We adopt top-2 selection as our default operating point because it offers the best balance: a 45% gain in CLIP alignment over top-1 (43.11 vs. 29.78) at a modest 2.2% reduction in IRS diversity (16.25% vs. 18.48%), while keeping FID stable. Table 5 addresses whether text conditioning is performed by the LLM or introduced by the unconditional GAN refiner: CLIP score and medical-domain FID remain stable across refinement, confirming that text–anatomy alignment is established at the LLM stage, with the GAN serving solely as a perceptual refiner for inter-patch artifacts.

Discussion: CTTok achieves the best FID (9.30), highest text alignment (43.11), and competitive diversity (16.25%) while generating volumes in 2.2 s, supporting our hypothesis that causal attention over anatomically grounded tokens is more efficient for text-to-CT synthesis without cross-attention or iterative denoising. The diversity-alignment trade-off in the top-K ablation (Table 4) is inherent to CLIP-based sample selection, not specific to our method; the unfiltered model (top-1) already matches baseline diversity.

4 Conclusion

CTTok reformulates text-to-CT synthesis as autoregressive language modeling over a frozen atlas codebook of spatially disentangled tokens, replacing cross-attention and iterative sampling with a single forward pass and lightweight GAN

refinement. Because each token independently encodes a fixed anatomical region, text conditioning is captured by the LLM's causal attention, achieving state-of-the-art image quality and text alignment at a fraction of the compute cost of diffusion and flow-matching alternatives. Future work will explore end-to-end codebook refinement while preserving full spatial disentanglement, extension to additional anatomical regions, and downstream clinical evaluation.
Limitations: Our evidence for the LLM's role in anatomical generation is empirical rather than mechanistic; direct probing of attention behavior, downstream clinical validation, and scaling to higher resolutions beyond the current $256^2 \times 128$ remain practical directions for future work.

Acknowledgments. We acknowledge HPC resources from NHR@FAU (projects b143dc, b180dc), funded by federal and Bavarian state authorities and Gerhard Wellein and his team's HPC approach. NHR@FAU hardware is partially funded by DFG 440719683. Additional support was received from ERC projects MIA-NORMAL 101083647 and CHARMS 101246053, DFG 513220538 and 512819079, and the state of Bavaria (HTA and the Bavarian Foundation Model Initiative). We further acknowledge resources provided by the Isambard-AI National AI Research Resource (AIRR), operated by the University of Bristol and funded by DSIT via UKRI and STFC [ST/AIRR/I-A-I/1023] [19]. We thank the Helmut Horten Foundation for generous support, which made this work possible, and the Istanbul Medipol University for invaluable support and provision of data. This research was supported in part by the Intramural Research Program of the National Library of Medicine (NLM), National Institutes of Health (NIH), and used the computational resources of the NIH high-performance computing Biowulf cluster. We used coding agents and LLMs from Anthropic, OpenAI, Google for coding, experiment orchestration, cluster monitoring, and as a writing aid.

Disclosure of Interests. The authors have no relevant competing interests.

References

1. Brooks, T., Peebles, B., Holmes, C., et al.: Video generation models as world simulators. OpenAI Blog **1**(8), 1 (2024), https://openai.com/index/video-generation-models-as-world-simulators/
2. Cao, H., Wang, Y., Chen, J.o.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) ECCV 2022 Workshops. pp. 205–218. Springer Nature Switzerland, Cham (2023)
3. Dombrowski, M., Zhang, W., Cechnicka, S., et al.: Image generation diversity issues and how to tame them. In: CVPR'25. pp. 3029–3039 (2025)
4. Dombrowski, M., Kainz, B.: Can diffusion models generalize? privacy and fairness trade-offs for medical data sharing. In: MIDL'25 (2025)
5. Douze, M., Guzhva, A., Deng, C., et al.: The faiss library. IEEE Transactions on Big Data (2025)
6. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: CVPR'21. pp. 12873–12883 (2021)
7. Giuffrè, M., Shung, D.L.: Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. npj Digital Medicine **6**(1), 186 (2023)

8. Guo, P., Zhao, C., Yang, D., et al.: MAISI: Medical ai for synthetic imaging. arXiv:2409.11169 (2025)
9. Hamamci, I.E., Er, S., Wang, C., et al.: Generalist foundation models from a multimodal dataset for 3d computed tomography. Nature Biomedical Engineering (2026). https://doi.org/10.1038/s41551-025-01599-y, https://doi.org/10.1038/s41551-025-01599-y
10. He, J., Yu, Q., Liu, Q., Chen, L.C.: Flowtok: Flowing seamlessly across text and image tokens. arXiv:2503.10772 (2025)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS'20 **33**, 6840–6851 (2020)
12. Huang, N., Gokaslan, A., Kuleshov, V., Tompkin, J.: The GAN is dead; long live the GAN! a modern GAN baseline. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 37, pp. 44177–44215 (2024)
13. Jolicoeur-Martineau, A.: The relativistic discriminator: a key element missing from standard GAN. In: ICLR'19 (2019)
14. Jordan, K., Jin, Y., Boza, V., et al.: Muon: An optimizer for hidden layers in neural networks (2024), https://kellerjordan.github.io/posts/muon/
15. Kao, J.P., Kao, H.T.: Large language models in radiology: A technical and clinical perspective. European Journal of Radiology Artificial Intelligence **2**, 100021 (2025)
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014)
17. Ktena, I., Wiles, O., Albuquerque, I., et al.: Generative models improve fairness of medical classifiers under distribution shifts. Nature Medicine **30**(4), 1166–1173 (2024)
18. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: ICLR'23 (2022)
19. McIntosh-Smith, S., Alam, S.R., Woods, C.: Isambard-AI: a leadership class supercomputer optimised specifically for artificial intelligence. arXiv.2410.11199 (2024)
20. Pinaya, W.H.L., Tudosiu, P.D., Dafflon, J., et al.: Brain imaging generation with latent diffusion models. In: Mukhopadhyay, A., Oksuz, I., Engelhardt, S., Zhu, D., Yuan, Y. (eds.) Deep Generative Models. pp. 117–126. Springer Nature Switzerland, Cham (2022)
21. Rassmann, S., Kügler, D., Ewert, C., Reuter, M.: Regression is all you need for medical image translation. IEEE Transactions on Medical Imaging **45**(5), 2156–2172 (2026). https://doi.org/10.1109/TMI.2025.3650412
22. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR'22. pp. 10684–10695 (2022)
23. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. pp. 234–241. Springer International Publishing, Cham (2015)
24. Skorokhodov, I., Tulyakov, S., Elhoseiny, M.: StyleGAN-V: a continuous video generator with the price, image quality and perks of StyleGAN2. arXiv:2112.14683 (2021)
25. Tudosiu, P.D., Pinaya, W.H.L., Graham, M.S.o.: Morphology-preserving autoregressive 3d generative modelling of the brain. In: Zhao, C., Svoboda, D., Wolterink, J.M., Escobar, M. (eds.) Simulation and Synthesis in Medical Imaging. pp. 66–78. Springer International Publishing, Cham (2022)
26. Unterthiner, T., Van Steenkiste, S., Kurach, K.o.: Towards accurate generative models of video: A new metric & challenges. arXiv:1812.01717 (2018)
27. VLM3D Organizers: VLM3D challenge: Vision-language modeling in 3D medical imaging. https://vlm3dchallenge.com/ (2025), ICCV 2025 Workshop

28. Wang, J., Reynaud, H., Erick, F.X., Kainz, B.: CTFlow: Video-inspired latent flow matching for 3d ct synthesis. arXiv:2508.12900 (2025)
29. Wasserthal, J., Breit, H.C., Meyer, M.T., et al.: Totalsegmentator: Robust segmentation of 104 anatomic structures in CT images. Radiology: Artificial Intelligence **5**(5), e230024 (2023)
30. Xu, Y., Sun, L., Peng, W., et al.: Medsyn: Text-guided anatomy-aware synthesis of high-fidelity 3d ct images. IEEE Transactions on Medical Imaging **43**(8), 2456–2470 (2024). https://doi.org/10.1109/TMI.2024.3415032
31. Yang, A., Li, A., Yang, B., et al.: Qwen3 technical report. arXiv:2505.09388 (2025)
32. Yu, L., Lezama, J., Gundavarapu, N.B., et al.: SPAE: Semantic pyramid AutoEncoder for multimodal generation with frozen LLMs. In: NeurIPS. vol. 36, pp. 52692–52704 (2023)
33. Yu, Q., He, J., Deng, X., Shen, X., Chen, L.C.: Randomized autoregressive visual generation. arXiv:2411.00776 (2024)
34. Yu, Q., Weber, M., Deng, X., et al.: An image is worth 32 tokens for reconstruction and generation. NeurIPS **37**, 128940–128966 (2024)
35. Yu, S., Tack, J., Mo, S., et al.: Generating videos with dynamics-aware implicit generative adversarial networks. In: ICLR - arXiv:2202.10571 (2022)
36. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: ICCV. pp. 11975–11986 (2023)
37. Zhao, C., Guo, P., Yang, D., et al.: MAISI-v2: Accelerated 3d high-resolution medical image synthesis with rectified flow and region-specific contrastive loss. arXiv:2508.05772 (2025)