

Text-to-CT Generation via 3D Latent Diffusion Model with Contrastive Vision-Language Pretraining

Daniele Molino^{a,*}, Camillo Maria Caruso^{a,*}, Filippo Ruffini^a, Paolo Soda^{a,b,**}, Valerio Guerrasi^a

^a*Unit of Artificial Intelligence and Computer Systems, Department of Engineering,
Università Campus Bio-Medico di Roma, Roma, Europe*

^b*Department of Diagnostics and Intervention, Biomedical Engineering and Radiation
Physics, Umeå University, Umeå, Sweden*

Abstract

Objective: While recent advances in text-conditioned generative models have enabled the synthesis of realistic medical images, progress has been largely confined to 2D modalities such as chest X-rays. Extending text-to-image generation to volumetric Computed Tomography (CT) remains a significant challenge, due to its high dimensionality, anatomical complexity, and the absence of robust frameworks that align vision-language data in 3D medical imaging. **Methods:** We introduce a novel architecture for Text-to-CT generation that combines a latent diffusion model with a 3D contrastive vision-language pretraining scheme. Our approach leverages a dual-encoder CLIP-style model trained on paired CT volumes and radiology reports to establish a shared embedding space, which serves as the conditioning input for generation. CT volumes are compressed into a low-dimensional latent space via a pretrained volumetric VAE, enabling efficient 3D denoising diffusion without requiring external super-resolution stages. **Results:** We evaluate our method on the CT-RATE dataset and conduct a comprehensive assessment of image fidelity, clinical relevance, and semantic alignment. Our model achieves competitive performance across all tasks, signif-

*Authors equally contributed to the work.

**Corresponding author.

Email addresses: daniele.molino@unicampus.it (Daniele Molino),
camillomaria.caruso@unicampus.it (Camillo Maria Caruso),
filippo.ruffini@unicampus.it (Filippo Ruffini), p.soda@unicampus.it,
paolo.soda@umu.se (Paolo Soda), valerio.guarrasi@unicampus.it (Valerio Guerrasi)

icantly outperforming prior baselines for text-to-CT generation. Moreover, we demonstrate that CT scans synthesized by our framework can effectively augment real data, improving downstream diagnostic performance. **Conclusion:** Our results show that modality-specific vision-language alignment is a key component for high-quality 3D medical image generation. By integrating contrastive pretraining and volumetric diffusion, our method offers a scalable and controllable solution for synthesizing clinically meaningful CT volumes from text, paving the way for new applications in data augmentation, medical education, and automated clinical simulation. Project page is at <https://github.com/cosbidev/Text2CT>.

Keywords: 3D Medical Image Synthesis, Vision-Language Alignment, Latent Diffusion Models, Contrastive Learning, Computed Tomography, Synthetic Medical Data

1. Introduction

Artificial Intelligence (AI) is rapidly transforming the medical domain, offering unprecedented opportunities for early diagnosis, treatment planning, and clinical decision support [1,2]. However, the full potential of AI in healthcare is often hindered by critical barriers, like scarcity of high-quality data, strict privacy regulations, and the high cost of expert annotations [3,4,5]. In response to these challenges, medical image generation has emerged as a compelling solution [6,7]. By synthesizing realistic data, generative models can help overcome data scarcity, support privacy-preserving applications, and enable the scalable training of AI systems [8,9]. In particular, text-conditioned generation offers fine-grained semantic control, allowing the synthesis of images that reflect detailed clinical findings. This paradigm has seen growing adoption in the context of 2D imaging modalities such as Chest X-rays (CXR). Models like RoentGen [10] and MedCoDi-M [11,12] have shown that diffusion-based approaches [13] conditioned on textual descriptions yield outputs that are visually realistic and semantically aligned with clinical reports. Yet, extending this capability to 3D volumes, e.g., Computed Tomography (CT) remains a largely open problem, mostly due to the high-resolution nature of the data, its volumetric structure, and the need for fine-grained anatomical consistency, challenges that remain largely underexplored by current generative models. CT is one of the most informative and widely used imaging modalities in clinical practice [14], offering high-resolution volumet-

ric views essential for diagnosing a broad range of conditions. However, this richness comes at the cost of increased data dimensionality and computational requirements, making the availability of large-scale CT datasets even more limited. The ability to synthesize CT volumes directly from textual descriptions would represent a major step forward, enabling the creation of diverse, anonymized datasets tailored to specific diagnostic scenarios. Although some initial efforts have been made in this direction [15,16], the field remains in its early stages, because the existing approaches often rely on general-purpose language models without explicitly aligning the encoded textual representations with the corresponding CT content. As a result, these models fall short of capturing the complex anatomical and pathological structures described in radiology reports, ultimately limiting the fidelity and controllability of the generated scans. Moreover, current methods [15,16] typically rely on multi-stage pipelines involving super-resolution modules to upscale low-resolution outputs. While this strategy reduces memory demands, it often introduces spatial inconsistencies and artifacts, compromising the anatomical coherence of the synthesized volumes. To overcome these limitations, we introduce a novel framework for Text-to-CT generation that leverages the power of latent diffusion models and vision-language pretraining. This enables a flexible and scalable solution for generating large-scale, anatomically faithful CT data tailored to specific diagnostic narratives.

The specific contributions of this work are:

- We introduce a three-dimensional CLIP [17] framework specifically trained to align radiology reports with CT volumes in a shared embedding space. This results in stronger text representations that are semantically and structurally grounded to imaging data. Beyond conditioning generation, we demonstrate the utility of these embeddings in a zero-shot setting.
- We present the first fully end-to-end Text-to-CT generation pipeline that produces high-resolution volumetric scans without requiring a separate super-resolution module, thus improving both spatial consistency and computational overhead.
- We conduct a comprehensive evaluation of the generated CT volumes, showing their superior capabilities in terms of quality, realism, clinical accuracy and potential utility in downstream tasks.

The remainder of this paper is organized as follows: Section 2 reviews related work on medical image synthesis and text-conditioned generation. Section 3 details our proposed methodology, including the vision-language alignment and generative components. In Section 4, we describe the dataset and preprocessing pipeline used in our experiments. We then present our experimental setup and evaluation metrics in Section 5, followed by a comprehensive analysis of results in Section 6. Finally, Section 7 concludes the paper and discusses future directions. Table 1 outlines the central challenge addressed in this work and clarifies how our approach advances the current state of Text-to-CT generation.

Problem or Issue	Current generative models focus on 2D modalities and fail to generate anatomically faithful and semantically controlled 3D CT volumes from clinical text.
What is Already Known	Diffusion models have enabled high-quality text-to-image generation in natural and 2D medical domains. Some 3D CT synthesis methods exist, but rely on multi-stage pipelines and weak textual conditioning, often leading to spatial artifacts and reduced clinical coherence.
What this Paper Adds	We introduce a novel Text-to-CT framework that combines a latent diffusion model with a contrastively trained 3D vision-language encoder. Our method directly synthesizes high-resolution chest CT volumes from radiology-style reports, achieving strong performance in both fidelity and clinical realism. We also show that synthetic scans improve downstream classification.
Who Would Benefit	Researchers in medical AI, especially those working on generative modeling, multimodal alignment, or data augmentation in 3D imaging, as well as practitioners seeking scalable solutions for clinical data simulation.

Table 1: Statement of significance on Text-to-CT generation using a 3D latent diffusion model with contrastive vision-language pretraining.

2. Related Works

The task of synthesizing high-resolution 3D medical images remains relatively underexplored compared to its 2D counterpart, due to the substantial increase in data dimensionality, memory requirements, and anatomical complexity. Initial efforts in this domain focused on unconditional generation, where models learn to synthesize volumetric scans from noise or latent priors, without external semantic conditioning. Among these, HA-GAN [18]

introduced a hierarchical training strategy to mitigate the computational burden of full-volume synthesis, generating coarse global structures alongside high-resolution sub-volumes. This approach enabled the generation of CT volumes up to $256 \times 256 \times 256$ voxels, outperforming earlier patch-based methods in terms of spatial consistency and visual realism. However, as a GAN-based model, it still suffered from mode collapse, limited diversity, and training instability [19,20].

A significant breakthrough came with the advent of diffusion probabilistic models [21], which have rapidly emerged as a more stable and effective alternative for generative modeling. In particular, Medical Diffusion [22] demonstrated that diffusion models outperform GANs in 3D medical image synthesis, achieving superior sample fidelity and diversity. Their architecture leverages a VQ-GAN [23] to compress volumetric data and applies denoising diffusion in this latent space, improving generation quality while reducing computational demands.

Building on this, more recent works have begun to explore text-conditioned generation of volumetric data, aiming to align radiology reports with 3D anatomical content. In order to produce 3D chest CT volumes guided by clinical descriptions, GenerateCT [15] proposes a cascaded architecture composed of three main components: (i) a causal vision transformer [24] that encodes 3D chest CT volumes into a sequence of tokens, (ii) a cross-modal transformer that aligns textual and visual embeddings via masked token prediction [25], and (iii) a text-conditioned diffusion model used for in-plane super-resolution [26]. The generation process is staged: a low-resolution volumetric representation is first autoregressively generated based on the text prompt, and then refined through a separate diffusion-based super-resolution module. While the framework demonstrates alignment between clinical prompts and the synthesized anatomy, its reliance on a 2D super-resolution module that processes each axial slice independently undermines the spatial consistency across the volume, an issue that becomes particularly evident when inspecting the generated CTs along the coronal and sagittal planes. Similarly, MedSyn [16] adopts a hierarchical generation scheme where low-resolution CT volumes are synthesized from free-text prompts and then refined via a second-stage three-dimensional super-resolution model. While this strategy ensures improved spatial coherence compared to slice-wise upsampling, it still introduces grid-like artifacts due to the decoupled nature of the generation and refinement stages. However, the use of a multi-stage pipeline in both GenerateCT and MedSyn is not merely architectural, it re-

flects a practical limitation: the high memory and compute requirements of direct high-resolution 3D generation make it infeasible without such intermediate steps, compelling these methods to rely on separate super-resolution stages to upscale the outputs.

Recently MAISI [27] has presented a unified framework that overcomes this limitation by enabling direct high-resolution 3D CT generation in a single pass. Rather than relying on external super-resolution modules, MAISI combines a volumetric compression network with a latent diffusion model, introducing tensor splitting parallelism (TSP) to efficiently achieve the generation of high-resolution volumes conditioned on segmentation masks. Although it does not support text conditioning, its architectural design significantly reduces computational overhead while preserving anatomical fidelity, making it the most scalable solution for synthetic CT generation.

Another major limitation of current text-to-CT approaches lies in the design of the text encoding pipeline, the component responsible for extracting semantic embeddings from clinical reports. In GenerateCT, the text encoder is based on a general-domain T5 [28] model, which lacks exposure to medical language and may struggle to capture radiologically meaningful patterns. MedSyn makes a first step toward domain adaptation by fine-tuning a biomedical BERT [29,30] on a subset of clinical reports using masked language modeling and paired-section prediction. However, neither method explicitly encourages the extracted textual embeddings to be semantically aligned with the corresponding CT volumes. This weak coupling between the linguistic and visual modalities may limit the model’s ability to generate anatomically coherent outputs conditioned on textual cues. Addressing this gap remains an open and promising direction for improving the controllability and clinical fidelity of text-to-CT generation.

3. Methods

To overcome the challenges posed by previous approaches, such as the lack of strong semantic alignment between radiology reports and volumetric anatomy, and the need for multi-stage super-resolution pipelines, we propose a unified generative architecture for synthesizing high-resolution 3D CT volumes from clinical descriptions, as illustrated in Figure 1. At the core of our method is a 3D CLIP module, shown in Figure 1.a, which is trained through contrastive learning to project CT volumes and their associated reports into a shared embedding space. This enables the model to extract text representa-

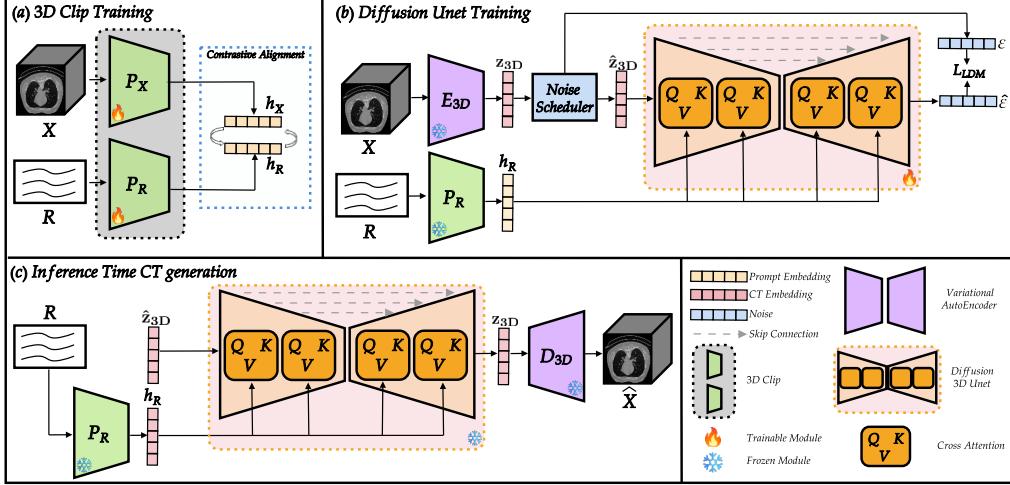


Figure 1: Overview of the proposed Text-to-CT generation framework: (a) **3D CLIP Training**: A contrastive learning setup aligns volumetric CT scans and radiology reports into a shared embedding space using a 3D Vision Transformer and a text encoder. (b) **Diffusion UNet Training**: A latent diffusion model is trained to denoise compressed CT representations, conditioned on textual embeddings. A pretrained VAE encoder compresses the volumes into latent vectors, which are noised and passed through a 3D U-Net, conditioned with report embeddings via cross-attention. (c) **Inference**: A synthetic latent code is generated from noise using the textual prompt, then decoded into a high-resolution CT volume via the VAE decoder.

tions that are semantically grounded in the anatomical content of the images, addressing the misalignment observed in state-of-the-art methods relying on generic language encoders. The extracted text embeddings are then used to condition the generation process through a latent diffusion model, which operates on a compact representation of the CT volume. This latent space is obtained via a volumetric VAE, depicted in Figure 1.b–c, which compresses the input scans into a lower-dimensional format while preserving structural fidelity. By avoiding the need to operate directly in voxel space, this design eliminates the dependence on external super-resolution stages and significantly reduces computational overhead. Finally, the latent diffusion model, implemented as a 3D U-Net with cross-attention mechanisms (Figure 1.b–c), iteratively denoises a random latent vector conditioned on the textual prompt. This allows the model to generate anatomically coherent CT scans that are semantically aligned with the input report, in a single forward pass. The combination of contrastive pretraining, latent-space generation, and vol-

umetric architectural design results in an efficient and end-to-end framework for controllable, high-resolution Text-to-CT synthesis. We describe each component in detail below.

3.1. Vision-Language Alignment via 3D CLIP

We adopt a contrastive learning strategy inspired by CLIP [17] to align CT volumes and radiology reports in a shared embedding space, enabling semantically grounded conditioning for generation, as depicted in Figure 1.a. However, while CLIP-based strategies have proven effective in natural image-text alignment and have been successfully adapted to the medical domain in the context of 2D modalities [12,31,32], their use in 3D medical imaging is still underexplored. To this end, we train a dual-encoder model comprising a 3D vision encoder $P_X(\cdot)$ and a text encoder $P_R(\cdot)$, which project input volumes \mathbf{X} and reports \mathbf{R} into embeddings $\mathbf{h}_\mathbf{X}$ and $\mathbf{h}_\mathbf{R}$, respectively.

The vision branch is implemented as a volumetric transformer, following [33], and is specifically designed to extract 3D features by jointly leveraging intra-slice and inter-slice information. This overcomes the limitations of 2D slice-based encoders, which fail to preserve spatial continuity in volumetric data [34,35]. The text encoder follows a masked self-attention architecture inspired by BERT, and is trained jointly with the vision encoder using the InfoNCE loss [36], defined as:

$$\mathcal{L}_{\mathbf{X}, \mathbf{R}} = -\log \frac{\exp(\mathbf{h}_\mathbf{X}^i{}^\top \cdot \mathbf{h}_\mathbf{R}^i / \tau)}{\exp(\mathbf{h}_\mathbf{X}^i{}^\top \cdot \mathbf{h}_\mathbf{R}^i / \tau) + \sum_{j \neq i} \exp(\mathbf{h}_\mathbf{X}^i{}^\top \cdot \mathbf{h}_\mathbf{R}^j / \tau)}, \quad (1)$$

where τ is a learnable temperature parameter, and the dot product is used to compute cosine similarity between normalized embeddings. The index i denotes a positive CT-report pair, while $j \neq i$ indicates negative pairs within the same batch. To ensure bidirectional alignment between modalities, we adopt a symmetric formulation:

$$\mathcal{L}_{\text{CLIP}} = \mathcal{L}_{\mathbf{X}, \mathbf{R}} + \mathcal{L}_{\mathbf{R}, \mathbf{X}}, \quad (2)$$

where $\mathcal{L}_{\mathbf{R}, \mathbf{X}}$ is computed analogously by treating the reports as queries and CT volumes as candidates. Once trained, the text encoder $P_R(\cdot)$ is frozen and used to extract conditioning embeddings for the generative model. This contrastive alignment ensures that clinical prompts capture both semantic content and spatial context, addressing the limitations of previous approaches

that relied on general-domain language encoders [15,16].

3.2. Latent Space Compression via VAE

To enable efficient 3D generation, we adopt a volumetric VAE [37] to compress full-resolution CT volumes into a lower-dimensional latent space (Figure 1.b). The encoder $E_{3D}(\cdot)$ maps the input scan \mathbf{X} to a latent representation $\mathbf{z}_{3D} = E_{3D}(\mathbf{X})$, which is decoded back by $D_{3D}(\cdot)$ to reconstruct the original volume $\hat{\mathbf{X}} = D_{3D}(\mathbf{z}_{3D})$. This representation bottleneck filters out noise and redundancies while retaining the essential anatomical structure. Operating in latent space significantly reduces memory and compute requirements, allowing the diffusion model to scale to high-resolution volumes without external super-resolution stages [13]. This is particularly relevant for volumetric data, where pixel-space generation incurs cubic complexity [38]. We use the 3D VAE architecture from MAISI [27], pretrained on a diverse set of CT and MRI scans to ensure generalizability across anatomical regions. The encoder produces latent tensors of size $\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4}$, enabling compact but expressive volumetric representations. Training is guided by a composite objective combining voxel-wise reconstruction loss $\mathcal{L}_{\text{recon}}$, perceptual similarity $\mathcal{L}_{\text{LPIPS}}$ [39], and an adversarial term \mathcal{L}_{adv} [19], to enforce both global fidelity and local realism. A KL divergence regularization \mathcal{L}_{KL} promotes smooth latent distributions [13]. To further reduce memory overhead, they integrate tensor splitting parallelism (TSP) [27], which partitions intermediate activations across spatial dimensions and distributes computation across devices, improving scalability during training and inference.

3.3. Text-Conditioned Diffusion in Latent Space

Our generative model leverages a latent diffusion process [13], which operates on compressed CT representations, significantly reducing computational cost while preserving semantic fidelity. Inspired by non-equilibrium thermodynamics [21,40], diffusion models learn to generate data by reversing a gradual corruption process applied to a known data distribution. Let \mathbf{z}_{3D} be the latent representation of a CT volume obtained via the pretrained VAE encoder. As shown in Figure 1.b, during training, we simulate a forward diffusion process where Gaussian noise is progressively added to \mathbf{z}_{3D} across T time steps, following a predefined variance schedule $\{\beta_t\}_{t=1}^T$. At each step, the noisy latent is obtained as:

$$\hat{\mathbf{z}}_{3D} = \alpha_t \mathbf{z}_{3D} + \sigma_t \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \quad (3)$$

where $\alpha_t = \prod_{s=1}^t (1 - \beta_s)$ and $\sigma_t = \sqrt{1 - \alpha_t^2}$.

The denoising network, parameterized as a 3D U-Net ϵ_θ , is trained to predict the noise component ϵ added at each step. The training objective follows the reparameterization proposed in [21], and is defined as:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{z, \epsilon, t} [\|\epsilon - \epsilon_\theta(\hat{\mathbf{z}}_{\mathbf{3D}}, t, \mathbf{h}_R)\|_2^2], \quad (4)$$

where $t \sim \mathcal{U}[1, T]$ is a uniformly sampled time step, and \mathbf{h}_R is the conditioning embedding extracted from the radiology report via the frozen text encoder P_R . To enable semantic guidance during generation, we inject the prompt embedding \mathbf{h}_R into the U-Net through cross-attention [24] at multiple layers and spatial resolutions, following the strategy introduced in Stable Diffusion [13]. This design allows the model to flexibly modulate the generation process based on the input text, ensuring that the synthesized volume aligns with the clinical content described in the report.

Classifier-free guidance [41] is integrated into both the training and inference phases to enhance the conditioning fidelity without requiring an external classifier. During training, the conditioning signal \mathbf{h}_R is randomly omitted with a fixed probability, allowing the model to jointly learn conditional and unconditional generation pathways.

At inference time, we apply guidance by interpolating between the predictions of the conditional and unconditional models:

$$\hat{\epsilon} = \epsilon_\theta(\hat{\mathbf{z}}_{\mathbf{3D}}, t) + w \cdot (\epsilon_\theta(\hat{\mathbf{z}}_{\mathbf{3D}}, t, \mathbf{h}_R) - \epsilon_\theta(\hat{\mathbf{z}}_{\mathbf{3D}}, t)) \quad (5)$$

where w is the guidance scale, controlling the strength of adherence to the conditioning text.

As shown in Figure 1.c, the inference generation process begins by sampling a latent vector $\hat{\mathbf{z}}_{\mathbf{3D}}$ from a standard Gaussian distribution. This noisy latent is then iteratively denoised using the learned reverse process, guided by the radiology prompt \mathbf{h}_R . After T denoising steps, we obtain a clean latent representation $\mathbf{z}_{\mathbf{3D}}$, which is finally decoded into a full-resolution CT volume using the pretrained VAE decoder. This approach allows the model to flexibly synthesize anatomically realistic CT scans that are semantically aligned with the clinical input description.

4. Materials

We base our experiments on the CT-RATE dataset [33], the largest publicly available collection of paired 3D chest CT volumes and radiology reports, spanning 18 pathological conditions¹. To ensure manageable training times while preserving diversity, we extracted a subset of approximately 30,000 volumes, which were inspected to identify and remove corrupted or mislabelled scans. In particular, a small number of non-thoracic volumes, such as head CTs, were found in the dataset despite its chest-focused scope, likely due to inconsistencies in metadata or acquisition logs. These outliers were excluded to ensure that the training set strictly reflects the anatomical domain of interest and to avoid introducing noise into the learning process. After filtering, we retained a total of 29,332 volumes, which were split into 27,514 for training and 1,818 for testing, following the official CT-RATE split to ensure comparability with prior work.

4.1. Preprocessing

To ensure a standardized and clinically meaningful evaluation, we applied a consistent preprocessing pipeline inspired by prior work [15,33]. All CT volumes were resampled to a uniform voxel spacing of $0.75 \times 0.75 \times 1.5$ mm, and then cropped or padded to a fixed spatial size of $512 \times 512 \times 128$ voxels. Intensity values were converted into Hounsfield Units (HU) using the rescale slope and intercept from the DICOM metadata, we subsequently clipped the HU values to the range $[-1000, +1000]$, which encompasses the relevant spectrum of anatomical structures in chest CT, from air-filled lung regions to dense osseous tissues [42,43]. This step not only removes outliers and scanner artifacts but also ensures compatibility with existing clinical standards, facilitating effective normalization. Subsequently, the resulting volumes were linearly normalized to the range $[0, 1]$, facilitating stable training of the generative model. This preprocessing yields uniform, anatomically aligned input volumes spanning a set of 18 diagnostic classes, suitable for both training and inference, facilitating reproducibility and comparability across experiments.

¹The 18 diagnostic classes are: Medical material, Arterial wall calcification, Cardiomegaly, Pericardial effusion, Coronary artery wall calcification, Hiatal hernia, Lymphadenopathy, Emphysema, Atelectasis, Lung nodule, Lung opacity, Pulmonary fibrotic sequela, Pleural effusion, Mosaic attenuation pattern, Peribronchial thickening, Consolidation, Bronchiectasis, and Interlobular septal thickening.

5. Experimental Configuration

This section, first details the implementation of each model component, including the 3D CLIP, the VAE, and the latent diffusion model. Second, it introduces the baselines considered for comparison and outlines the evaluation metrics used to assess image fidelity, clinical plausibility, and vision-language alignment. Third, it describes the additional experimental protocols used to analyze the impact of contrastive pretraining, classifier-free guidance, and the utility of synthetic data in downstream tasks.

5.1. Implementation Details

3D CLIP Architecture. Our 3D CLIP model comprises a dual-encoder framework: a vision encoder adapted from the CT-CLIP architecture [33], modified to ensure compatibility with the output space of our text encoder, which is based on a masked self-attention Transformer. The text encoder, inspired by BERT [30], employs masked self-attention mechanisms to capture contextual relationships within radiology reports. The encoders were jointly trained for 100 epochs, utilizing the AdamW optimizer [44] with a learning rate of 5×10^{-5} , weight decay of 1×10^{-4} , and epsilon of 1×10^{-8} . The batch size was set to 16 per GPU.

Variational Autoencoder. For latent compression, we adopt the VAE architecture proposed in MAISI [27], which was pretrained by the authors on a large corpus of CT and MRI scans from different datasets. The VAE was kept frozen during all diffusion training phases, and only used for latent encoding and decoding at training and inference time, respectively. To further improve training efficiency and reduce GPU memory usage, we precomputed all latent representations for the training set using the frozen VAE encoder. This allowed the diffusion model to operate directly on latent tensors without re-encoding the input volumes at each iteration, significantly accelerating the training process.

Latent Diffusion Model. The core generative model is a latent diffusion architecture based on a 3D U-Net with cross-attention conditioning. The model was trained for 200 epochs, employing the AdamW optimizer [44], with a learning rate of 1×10^{-4} and a batch size of 8 per GPU. We employed a Polynomial Learning Rate scheduler, progressively decaying the learning rate over the course of training. The model was trained using classifier-free

guidance [41], with a 10% probability of dropping the text condition during training.

All experiments were conducted on an high efficiency cluster equipped with 4 NVIDIA A100 GPUs with 80GB memory each.

5.2. Competitors

To benchmark the effectiveness of our proposed framework for Text-to-CT generation, we compare it against two state-of-the-art baselines that we select because they have publicly available code and pretrained weights: GenerateCT [15] and MedSyn [16]. As described in Section 2, both methods are designed to synthesize 3D CT volumes conditioned on radiology reports, and represent, to the best of our knowledge, the only reproducible pipelines currently available for this task.

5.3. Evaluation Metrics

Our evaluation protocol comprises both quantitative and qualitative assessments, aimed at evaluating fidelity, clinical realism, and semantic alignment.

5.3.1. Image Fidelity

To objectively evaluate the quality of the generated CT volumes, we employed the Fréchet Inception Distance (FID) [45], a widely used metric that measures the statistical similarity between real and synthetic data distributions. Given the 3D nature of CT data, we computed FID using two complementary approaches:

- 2.5D FID: We extracted slices along the axial, coronal, and sagittal planes from both real and generated volumes. These slices were processed using a 2D backbone pretrained on medical imaging tasks to capture relevant features. FID was then computed separately for each plane, and the results were averaged to obtain the final FID score.
- 3D FID: To fully capture the volumetric characteristics of CT data, we also employed a 3D backbone trained on medical imaging datasets to extract features from entire volumes. FID was then computed in the 3D feature space, providing a volumetric assessment of the synthetic samples quality.

5.3.2. Factual Correctness

While FID provides a useful statistical approximation of the similarity between real and generated data distributions, it fails to capture whether the generated images faithfully reflect the pathological findings described in the associated report, an aspect we refer to as Factual Correctness. In other words, low FID scores do not necessarily imply that the generated CT volumes encode realistic pathological patterns or anatomically coherent structures. To address this limitation, we adopted a classification-based evaluation strategy designed to assess whether the synthetic CT volumes preserve clinically meaningful features. Specifically, we employed CT-Net [46], a 3D convolutional neural network originally designed for diagnostic classification from full CT volumes. We reimplemented and retrained CT-Net on the training set of the CT-RATE dataset to tailor it to the anatomical and pathological distribution present in our dataset. Once trained, CT-Net was applied to the generated volumes corresponding to each report in the CT-RATE test set, in order to classify the presence or absence of the 18 pathological classes introduced in Section 4. Its predictions were then compared with the original ground truth labels associated with the conditioning reports. This setup allows us to quantify how effectively each generative model captures diagnostically relevant information, going beyond pixel-level fidelity to evaluate high-level consistency.

5.3.3. Vision-Language Alignment

To rigorously assess the quality of the joint embedding space learned by our 3D-CLIP model, we follow the evaluation protocol established in CT-CLIP [33], which includes three tasks: zero-shot multi-abnormality classification, report-to-volume retrieval, and volume-to-volume retrieval.

Zero-Shot Classification. We adopt the zero-shot multi-label classification setup proposed in [47], assessing our 3D-CLIP ability to recognize clinical conditions without any supervised fine-tuning. For each CT volume in the validation set, we construct a pair of binary textual prompts describing the presence or absence of each pathology (e.g., “This scan shows signs of pleural effusion” vs. “This scan does not show signs of pleural effusion”). We then compute the cosine similarity between the frozen vision embedding of the CT and the embeddings of the two prompts. These scores are normalized via a softmax function and used to assign the class label with the highest score. This process is repeated independently for each pathology, enabling

multi-label zero-shot prediction.

Volume-to-Volume Retrieval. To assess the structural coherence of the learned embedding space, we perform volume-to-volume retrieval. Given a query CT volume, we rank all other volumes in the validation set according to their cosine similarity to the query in the shared latent space. A volume is considered relevant if it shares at least one abnormality with the query, based on the intersection-over-union (IoU) of their binary pathology annotations. To quantify retrieval performance, we adopt the Mean Average Precision at K (MAP@K), a metric suited for task requiring fine-grained semantic alignment [48]: it measures how well the model prioritizes clinically relevant volumes among its top-K most similar results, thus reflecting not only the presence of correct matches but also their rank in the retrieval list.

Report-to-Volume Retrieval. This task evaluates the model’s ability to retrieve the most semantically appropriate CT volume given a free-text radiology report. Following the protocol in [33], we compute cosine similarities between the report embedding and the embeddings of all candidate volumes in the test set, ranking them by decreasing similarity. The underlying assumption is that well-aligned vision-language representations will bring truly corresponding image-report pairs closer in the shared embedding space. Retrieval performance is quantified using Recall@K, with $K \in \{5, 10, 50, 100\}$, a standard metric in text-to-image retrieval [49]. Recall@K measures the frequency with which the ground-truth volume appears among the top-K retrieved items, thereby capturing how effectively the model prioritizes clinically accurate matches in response to textual queries.

5.3.4. Ablation Study

To evaluate the impact of contrastive pretraining, we conducted an ablation in which the 3D-CLIP module was replaced with a text encoder pre-trained on a medical corpus but never aligned with CT scans. This modification removes the modality-specific grounding between textual and visual features. We compare this variant against our full model in terms of generation fidelity, using both 2.5D and 3D FID scores, as well as clinical plausibility, assessed via CT-Net classification performance.

5.3.5. Effect of Guidance Scale

We investigate the role of the classifier-free guidance scale w , defined in eq. 5 in shaping the trade-off between image fidelity and semantic alignment.

To this end, we generate CT volumes across a range of guidance values and evaluate their quality using both FID and clinical classification metrics. This analysis aims to better understand how the strength of the conditioning signal influences the generative process and its downstream utility.

5.3.6. Synthetic Data Utility

Synthetic data has emerged as a powerful tool for addressing data scarcity, class imbalance, and privacy constraints in medical imaging. By augmenting training sets with generated examples, it is possible to improve model generalization and robustness, particularly in scenarios where annotated data is limited or unevenly distributed [50]. We explore whether synthetic CT scans generated by our model can effectively support downstream clinical tasks. To this end, we train a diagnostic classifier (CT-Net) under three different data regimes: using only real CT volumes, only synthetic volumes, and a combination of both. The classifier is then evaluated exclusively on real test data, using standard metrics such as AUC and precision, to assess the added value of synthetic data in model training.

6. Results and Discussion

This section presents an in-depth analysis of the proposed method’s performance. Results are presented and discussed in the same order used to present the different evaluation metrics in Section 5.3.

6.1. Image Fidelity Evaluation

Method	FID 2.5D ↓				FID 3D ↓
	Axial	Coronal	Sagittal	Avg.	
GenerateCT [15]	6.701	11.793	9.694	9.396	0.166
MedSyn [16]	8.789	8.900	8.717	8.802	0.169
Ours	4.597	3.803	2.545	3.648	0.003

Table 2: FID scores computed using 2.5D (slice-based) and 3D feature extractors. The 2.5D scores are reported for axial, coronal, and sagittal slices, along with their average. Lower scores indicate greater similarity to real CT volumes. Best results for each column are highlighted in bold black, second-best in bold blue.

Table 2 reports the FID computed both in 2.5D, across axial, coronal, and sagittal views, and in 3D over the full volume. Our model outperforms

all competitors, achieving the lowest FID across all views, which indicates a closer match between the distribution of generated and real CT volumes. As expected and already discussed in Section 2, GenerateCT shows relatively good performance along the axial plane ($FID = 6.701$), but its scores increase significantly along the coronal and sagittal directions ($FID = 11.793$ and 9.694 , respectively). This behavior reflects the limitations of its 2D slice-wise super-resolution module, which fails to preserve spatial coherence across adjacent slices. In contrast, MedSyn applies 3D super-resolution, resulting in more uniform FID scores across the three planes. However, this architecture still introduces global artifacts, which are particularly visible in the form of grid-like patterns, as illustrated in Figure 2. Volumetric rendering of the scans, presented in the last row of the same figure, further confirms these limitations, where both GenerateCT and MedSyn show structural inconsistencies in the reconstructed bone structures. Notably, in the 3D FID evaluation, which captures the volumetric realism of entire scans, our method outperforms both baselines by a large margin, confirming the effectiveness of our end-to-end generation strategy.

6.2. Factual Correctness Evaluation

Model	AUC (Macro)	AUC (Weighted)	Precision (Macro)	Precision (Weighted)
Real	0.751	0.741	0.484	0.564
GenerateCT	0.581	0.568	0.285	0.361
MedSyn	0.560	0.547	0.282	0.362
Ours	0.745	0.731	0.477	0.549

Table 3: Evaluation of factual correctness using CT-Net [46], a 3D classification model trained on real CT-RATE scans. The table reports AUC and Precision (Macro and Weighted averages) computed across 18 disease classes, comparing real and synthetic test sets generated by different models. Higher values indicate better preservation of clinically meaningful features in the generated volumes. Best results for each column are highlighted in bold black, second-best in bold blue.

To assess the factual correctness of the generated CT volumes, we evaluated their diagnostic informativeness using CT-Net [46], a 3D convolutional neural network trained on the real CT-RATE training set. Performance was measured on real CTs and on synthetic test sets generated by each method. Table 3 reports AUC and Precision, computed as both macro and weighted averages over 18 disease classes. While macro averaging treats all classes equally, the weighted average accounts for class imbalance by adjusting each

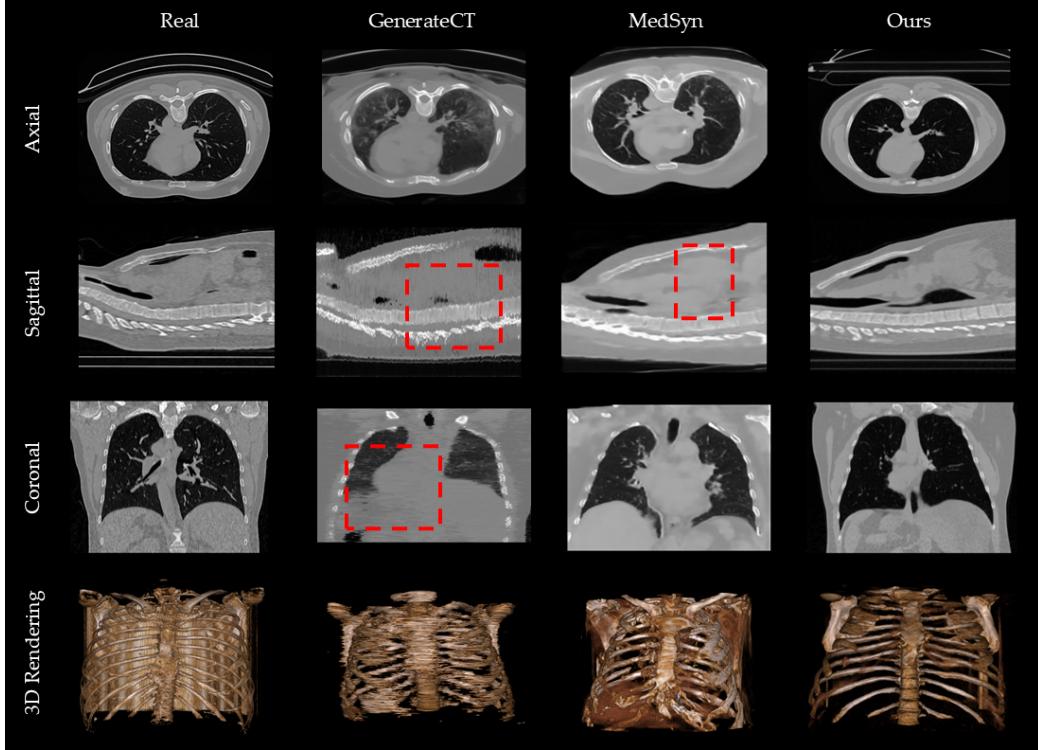


Figure 2: Qualitative comparison of real and synthetic CT volumes generated by competing methods using the same textual prompt. The top three rows show axial, sagittal, and coronal slices, respectively, while the bottom row displays corresponding 3D volume renderings. Red boxes highlight regions affected by visual artifacts introduced by super-resolution stages. GenerateCT [15] suffers from severe inter-slice discontinuities due to its 2D upsampling, especially visible in sagittal and coronal views. MedSyn [16] exhibits grid-like distortions from its 3D refinement stage. In contrast, our model produces anatomically coherent volumes across all planes and exhibits structural fidelity in the 3D rendering.

class contribution based on its prior. This distinction is important, as class imbalance not only affects the training of the CT-Net classifier, but also impacts the generative models, which tend to underperform on rare conditions seen less frequently during training. As a reference point, we also include in the first row of the table the performance of CT-Net on real test data. Our model nearly matches the real data upper bound in both AUC and precision, highlighting its ability to produce anatomically plausible and diagnostically meaningful CT scans. In contrast, GenerateCT and MedSyn show substantial degradation in classification performance, suggesting that

the pathologies in their generated volumes are either absent or poorly localized. This limitation is likely linked to their shallow text encoding strategies, which omit explicit alignment to the anatomical content of CT scans. In our case, the use of a dedicated 3D-CLIP module proves effective in bridging this gap, enabling the model to better encode and condition on semantically rich, clinically relevant features described in the radiology reports.

6.3. Evaluating Vision-Language Alignment

Methods	Zero-Shot			Volume-to-Volume (MAP@K)				Report-to-Volume (Recall@K)			
	Acc.	AUC	Pre.	MAP@1	MAP@5	MAP@10	MAP@50	Re@5	Re@10	Re@50	Re@100
CT-CLIP	0.668	0.731	0.323	0.886	0.683	0.572	0.489	0.029	0.050	0.180	0.287
Ours	0.633	0.676	0.353	0.987	0.848	0.801	0.758	0.041	0.072	0.229	0.368

Table 4: Comparison of our 3D-CLIP model and CT-CLIP [33] across three evaluation tasks: zero-shot multi-label classification, volume-to-volume retrieval (measured with MAP@K), and report-to-volume retrieval (measured with Recall@K). Best results for each column are highlighted in bold.

Table 4 summarizes the comparative evaluation of our 3D-CLIP model against CT-CLIP [33] across three vision-language tasks. In the zero-shot classification setting, CT-CLIP achieves slightly better results in Accuracy and AUC, however our model maintains comparable performance, achieving a higher macro-averaged precision. We hypothesize that such a tradeoff may stem from the mismatch between the prompt styles used during training and those employed at evaluation time. Specifically, the reports in the CT-RATE dataset follow a structured format, typically divided into Findings and Impression sections, whereas in our evaluation setup, we rely on a significantly different prompting scheme. This discrepancy may lead to a misalignment between the textual cues learned by the model and those provided at inference, ultimately affecting generation quality. In the volume-to-volume retrieval task, our model substantially outperforms CT-CLIP across all MAP@K metrics. Notably, we reach a MAP@1 of 0.987, indicating that our embedding space effectively preserves semantic similarities between volumetric scans. This result underscores the strength of contrastive alignment in learning structured, clinically meaningful embeddings. Similarly, in the report-to-volume retrieval setting, our model consistently yields higher Recall@K scores. This improvement reflects stronger vision-language grounding and a better capacity to associate radiological reports with the correct anatomical content. These results confirm the effectiveness of our 3D-CLIP

architecture in learning clinically meaningful and modality-bridging representations, especially in retrieval-based scenarios.

6.4. Ablation on Contrastive Vision-Language Pretraining

Model	Axial ↓	Coronal ↓	Sagittal ↓	Avg. ↓	FID 3D ↓	AUC ↑	Precision ↑
without 3D-CLIP	4.778	5.842	3.274	4.632	0.004	0.665	0.367
with 3D-CLIP	4.597	3.803	2.545	3.648	0.003	0.745	0.477

Table 5: FID scores and classification performance of the proposed approach with and without 3D-CLIP contrastive training. Removing CLIP leads to a substantial drop in AUC and weighted precision. Best results for each column are highlighted in bold.

To evaluate the impact of contrastive vision-language pretraining in our architecture, we conducted an ablation study where we replaced the 3D-CLIP encoder with a text encoder pretrained on medical corpora but without joint alignment with CT volumes. This modification removes the modality-specific grounding between text and image, preventing the model from learning a shared semantic space. The results, reported in Table 5, revealed a notable decline in both image fidelity and clinical relevance. Specifically, the average 2.5D FID increased from 3.648 to 4.632, with the axial view rising from 4.597 to 4.778, the coronal view from 3.803 to 5.842, and the sagittal view from 2.545 to 3.274. In terms of clinical classification, the AUC decreased from 0.745 to 0.665, and precision dropped from 0.477 to 0.367. These findings underscore the importance of contrastive pretraining for effectively aligning vision and language modalities in the context of volumetric medical data. By fostering a semantically rich and anatomically grounded embedding space, contrastive training enables the generative model to better interpret and condition on textual prompts, leading to the synthesis of CT volumes that are not only visually faithful but also more diagnostically meaningful, bridging the gap between descriptive clinical language and 3D anatomical representation.

6.5. Effect of Guidance Scale on Generation and Clinical Utility

We further investigated the impact of the classifier-free guidance scale on the quality of generated CT scans and their clinical utility. Figure 3 reports the trend of FID 3D and Precision as the guidance scale varies from 0 to 15. Interestingly, while FID values remain consistently low across all settings, suggesting stable visual fidelity, the precision varies significantly and reaches

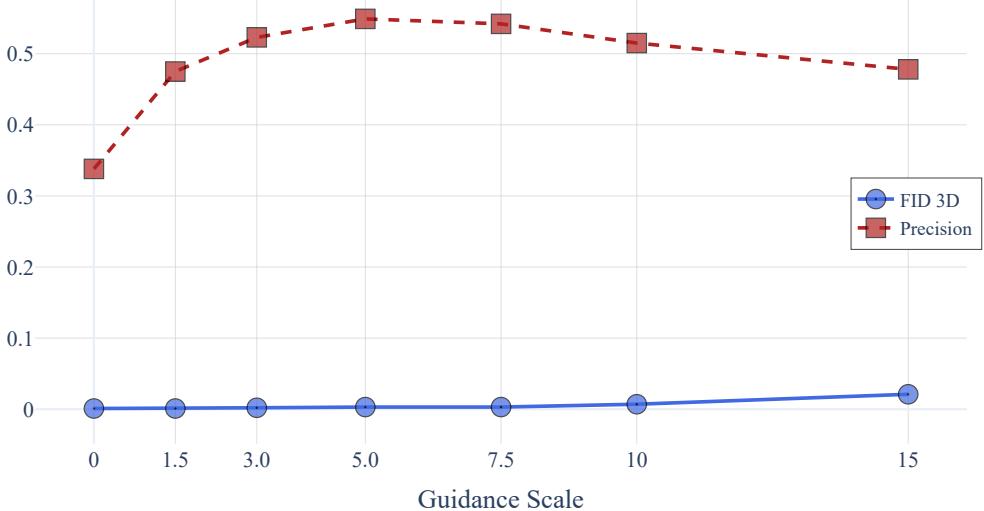


Figure 3: Effect of guidance scale on generation fidelity (FID 3D) and clinical relevance (Precision). While FID remains low and relatively stable across scales, precision peaks at a moderate guidance value, highlighting the importance of semantic controllability beyond visual similarity.

its maximum at a guidance scale of 5.0. This result confirms that evaluation based solely on image similarity metrics like FID may be misleading in the medical domain, where semantic alignment with clinical findings is critical.

6.6. Utility of Synthetic Data for Downstream Training

Training Set	AUC (Macro) \uparrow	AUC (Weighted) \uparrow	Precision (Macro) \uparrow	Precision (Weighted) \uparrow
Real only	0.751	0.741	0.484	0.564
Synthetic only	0.702	0.687	0.439	0.510
Real + Synthetic	0.778	0.765	0.512	0.584

Table 6: Performance of CT-Net trained on different datasets: only real CTs, only synthetic CTs generated by our model, and the union of both. All models are evaluated on the same test set of real CT scans. Best results for each column are highlighted in bold black, second-best in bold blue.

We further evaluated the utility of our synthetic CT volumes by using them to train a CT-Net classifier under three different configurations: using only real CT scans, using only synthetic scans generated by our model, and combining both datasets. In all cases, the model was evaluated on the same held-out test set composed exclusively of real CT volumes. Table 6,

shows that the classifier achieves reasonable performance when trained solely on synthetic data, demonstrating that the generated scans retain clinically meaningful features. More interestingly, combining real and synthetic data leads to a consistent improvement across all metrics, suggesting that synthetic data can be effectively used to augment real datasets, helping the classifier to generalize better. Although synthetic data alone may not yet be a substitute for real scans, it can serve as a valuable complement, especially in scenarios where annotated data is limited or class imbalance is a concern.

7. Conclusion

In this work, we have introduced a novel architecture for text-conditioned generation of high-resolution 3D CT scans, leveraging a combination of a latent diffusion model and a contrastively trained 3D-CLIP. Our approach enables the generation of anatomically coherent and semantically faithful volumes directly from clinical narratives, without relying on super-resolution stages. Through a comprehensive evaluation, we demonstrated the benefits of our method across multiple dimensions: image fidelity, clinical plausibility, and vision-language alignment. In particular, the inclusion of a 3D-CLIP module proved critical in bridging the semantic gap between text and volumetric anatomy, improving both generation quality and clinical utility. Additionally, we showed that synthetic CT volumes produced by our model can be effectively used to augment real datasets, yielding measurable improvements in classification performance.

Despite these encouraging results, several limitations remain. While our framework eliminates the need for explicit super-resolution, generating extremely large volumes still requires significant computational resources. Moreover, the current text encoder, though trained contrastively, could benefit from additional medical-specific pretraining to further enhance semantic precision. Future work may explore scalable diffusion backbones, more flexible conditioning schemes, and rigorous clinical validation through radiologist-driven studies. In addition, the opportunity to expand this framework to tackle the inverse task of CT-to-text generation, enabling automatic report synthesis from volumetric scans, represents an interesting direction for future research, with potential applications in clinical decision support and automated documentation. Additionally, exploring multimodal generation involving other imaging modalities such as X-Rays could further extend the applicability of this approach across diverse clinical contexts. Overall, our

findings support the feasibility and utility of direct Text-to-CT generation and open new avenues for data augmentation, training of diagnostic models, and automated scenario simulation in medical imaging.

Author Contributions

Daniele Molino: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing; **Camillo Maria Caruso:** Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Software, Validation, Writing – original draft, Writing – review & editing; **Filippo Ruffini:** Conceptualization, Formal analysis, Investigation, Methodology, Software; **Paolo Soda** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing; **Valerio Guerrasi:** Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing.

Acknowledgment

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. iii) Cancerforskningsfonden Norrland project MP23-1122; iv) Kempe Foundation project JCSMK24-0094; Resources are provided by the National Academic Infrastructure for Supercomputing in Sweden (NAIIS) 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.

References

- [1] Zaid Abdi Alkareem Alyasseri, Mohammed Azmi Al-Betar, Iyad Abu Doush, Mohammed A Awadallah, Ammar Kamal Abasi, Sharif Naser

- Makhadmeh, Osama Ahmad Alomari, Karrar Hameed Abdulkareem, Afzan Adam, Robertas Damasevicius, et al. Review on COVID-19 diagnosis models based on machine learning and deep learning approaches. *Expert systems*, 39(3):e12759, 2022.
- [2] Geert Litjens, Clara I Sánchez, Nadya Timofeeva, Meyke Hermsen, Iris Nagtegaal, Iringo Kovacs, Christina Hulsbergen-Van De Kaa, Peter Bult, Bram Van Ginneken, and Jeroen Van Der Laak. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific reports*, 6(1):26286, 2016.
 - [3] Ms Aayushi Bansal, Dr Rewa Sharma, and Dr Mamta Kathuria. A systematic review on data scarcity problem in deep learning: solution and applications. *ACM Computing Surveys (Csur)*, 54(10s):1–29, 2022.
 - [4] Pengfei Guo, Puyang Wang, Jinyuan Zhou, Shanshan Jiang, and Vishal M Patel. 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*, pages 2423–2432, 2021.
 - [5] Nima Tajbakhsh, Holger Roth, Demetri Terzopoulos, and Jianming Liang. Guest editorial annotation-efficient deep learning: the holy grail of medical imaging. *IEEE transactions on medical imaging*, 40(10):2526–2533, 2021.
 - [6] Nripendra Kumar Singh and Khalid Raza. Medical image generation using generative adversarial networks: A review. *Health informatics: A computational perspective in healthcare*, pages 77–96, 2021.
 - [7] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical image analysis*, 88:102846, 2023.
 - [8] Alejandro F Frangi, Sotirios A Tsaftaris, and Jerry L Prince. Simulation and synthesis in medical imaging. *IEEE transactions on medical imaging*, 37(3):673–679, 2018.

- [9] Phillip Chlap, Hang Min, Nym Vandenberg, Jason Dowling, Lois Holloway, and Annette Haworth. A review of medical image data augmentation techniques for deep learning applications. *Journal of medical imaging and radiation oncology*, 65(5):545–563, 2021.
- [10] Christian Bluethgen, Pierre Chambon, Jean-Benoit Delbrouck, Rogier van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P Langlotz, and Akshay S Chaudhari. A vision–language foundation model for the generation of realistic chest x-ray images. *Nature Biomedical Engineering*, pages 1–13, 2024.
- [11] Daniele Molino, Francesco Di Feola, Eliodoro Faiella, Deborah Fazzini, Domiziana Santucci, Linlin Shen, Valerio Guerrasi, and Paolo Soda. MedCoDi-M: A Multi-Prompt Foundation Model for Multimodal Medical Data Generation. *arXiv preprint arXiv:2501.04614*, 2025.
- [12] Daniele Molino, Francesco di Feola, Linlin Shen, Paolo Soda, and Valerio Guerrasi. Any-to-Any Vision-Language Model for Multimodal X-ray Imaging and Radiological Report Generation. *arXiv preprint arXiv:2505.01091*, 2025.
- [13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [14] Michalis Mazonakis and John Damilakis. Computed tomography: What and how does it measure? *European journal of radiology*, 85(8):1499–1504, 2016.
- [15] Ibrahim Ethem Hamamci, Sezgin Er, Anjany Sekuboyina, Enis Simsar, Alperen Tezcan, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Furkan Almas, Irem Doğan, Muhammed Furkan Dasdelen, et al. Generatect: Text-conditional generation of 3d chest ct volumes. In *European Conference on Computer Vision*, pages 126–143. Springer, 2024.
- [16] Yanwu Xu, Li Sun, Wei Peng, Shuyue Jia, Katelyn Morrison, Adam Perer, Afroz Zandifar, Shyam Visweswaran, Motahhare Eslami, and Kayhan Batmanghelich. Medsyn: Text-guided anatomy-aware synthesis

of high-fidelity 3d ct images. *IEEE Transactions on Medical Imaging*, 2024.

- [17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [18] Li Sun, Junxiang Chen, Yanwu Xu, Mingming Gong, Ke Yu, and Kayhan Batmanghelich. Hierarchical amortized GAN for 3D high resolution medical image synthesis. *IEEE journal of biomedical and health informatics*, 26(8):3966–3975, 2022.
- [19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [20] Divya Saxena and Jiannong Cao. Generative adversarial networks (gans) challenges, solutions, and future directions. *ACM Computing Surveys (CSUR)*, 54(3):1–42, 2021.
- [21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS ’20. Curran Associates Inc., 2020.
- [22] Firas Khader, Gustav Müller-Franzes, Soroosh Tayebi Arasteh, Tianyu Han, Christoph Haarburger, Maximilian Schulze-Hagen, Philipp Schad, Sandy Engelhardt, Bettina Baesler, Sebastian Foersch, et al. Denoising diffusion probabilistic models for 3D medical image generation. *Scientific Reports*, 13(1):7303, 2023.
- [23] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.
- [24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention

is all you need. *Advances in neural information processing systems*, 30, 2017.

- [25] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11315–11325, 2022.
- [26] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.
- [27] 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.
- [28] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- [29] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer, 2022.
- [30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.
- [31] Zihao Zhao, Yuxiao Liu, Han Wu, Mei Wang, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Zhiming Cui, Qian Wang, et al. Clip in medical imaging: A comprehensive survey. *arXiv preprint arXiv:2312.07353*, 2023.

- [32] Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K Hong, Woonhyuk Baek, and Byungseok Roh. Cxr-clip: Toward large scale chest x-ray language-image pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 101–111. Springer, 2023.
- [33] Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Seval Nil Esirgun, Irem Doga, Omer Faruk Durugol, Weicheng Dai, Murong Xu, et al. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834*, 2024.
- [34] Amarjeet Kumar, Hongxu Jiang, Muhammad Imran, Cyndi Valdes, Gabriela Leon, Dahyun Kang, Parvathi Nataraj, Yuyin Zhou, Michael D Weiss, and Wei Shao. A flexible 2.5 D medical image segmentation approach with in-slice and cross-slice attention. *Computers in Biology and Medicine*, 182:109173, 2024.
- [35] Minh H Vu, Guus Grimbergen, Tufve Nyholm, and Tommy Löfstedt. Evaluation of multislice inputs to convolutional neural networks for medical image segmentation. *Medical Physics*, 47(12):6216–6231, 2020.
- [36] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- [37] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [38] Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *MICCAI Workshop on Deep Generative Models*, pages 117–126. Springer, 2022.
- [39] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

- [40] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265. PMLR, 07–09 Jul 2015.
- [41] Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- [42] Tami D. DenOtter and Johanna Schubert. *Hounsfield Unit*. StatPearls Publishing, Treasure Island (FL), 2025.
- [43] Ramit Lamba, John P McGahan, Michael T Corwin, Chin-Shang Li, Tien Tran, J Anthony Seibert, and John M Boone. CT Hounsfield numbers of soft tissues on unenhanced abdominal CT scans: variability between two different manufacturers’ MDCT scanners. *American Journal of Roentgenology*, 203(5):1013–1020, 2014.
- [44] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [45] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [46] 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.
- [47] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nature biomedical engineering*, 6(12):1399–1406, 2022.
- [48] Chengkuan Chen, Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Andrew J Schaumberg, and Faisal Mahmood. Fast and scalable search of whole-slide images via self-supervised deep learning. *Nature Biomedical Engineering*, 6(12):1420–1434, 2022.

- [49] Shaoting Zhang, Ming Yang, Timothee Cour, Kai Yu, and Dimitris N Metaxas. Query specific fusion for image retrieval. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part II 12*, pages 660–673. Springer, 2012.
- [50] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Synthetic data augmentation using gan for improved liver lesion classification. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 289–293. IEEE, 2018.