

TRACE: Temporally Reliable Anatomically-Conditioned 3D CT Generation with Enhanced Efficiency

Minye Shao¹ , Xingyu Miao¹ , Haoran Duan² , Zeyu Wang³ , Jingkun Chen⁴,

Yawen Huang⁵, Xian Wu⁵, Jingjing Deng⁶ , Yang Long^{1*} , Yefeng Zheng⁷

¹Department of Computer Science, Durham University, Durham, UK;

²Department of Automation, Tsinghua University, Beijing, China;

³College of Computer Science and Engineering, Dalian Minzu University, Dalian, China;

⁴Department of Engineering Science, University of Oxford, Oxford, UK;

⁵Jarvis Research Center, Tencent YouTu Lab, Shenzhen, China;

⁶School of Engineering Mathematics and Technology, University of Bristol, Bristol, UK;

⁷Medical Artificial Intelligence Laboratory, School of Engineering, Westlake University, Hangzhou, China

yang.long@ieee.org *

Abstract

3D medical image generation is essential for data augmentation and patient privacy, calling for reliable and efficient models suited for clinical practice. However, current methods suffer from limited anatomical fidelity, restricted axial length, and substantial computational cost, placing them beyond reach for regions with limited resources and infrastructure. We introduce **TRACE**, a framework that generates 3D medical images with spatiotemporal alignment using a 2D multimodal-conditioned diffusion approach. TRACE models sequential 2D slices as video frame pairs, combining segmentation priors and radiology reports for anatomical alignment, incorporating optical flow to sustain temporal coherence. During inference, an overlapping-frame strategy links frame pairs into a **flexible length** sequence, reconstructed into a spatiotemporally and anatomically aligned 3D volume. Experimental results demonstrate that TRACE effectively balances **computational efficiency** with preserving **anatomical fidelity** and spatiotemporal consistency. Code is available at: <https://github.com/VinyeShaw/TRACE>.

1. Introduction

3D medical image generation is essential for precise diagnosis, personalized treatment, and surgical planning. However, privacy constraints and limited data sharing hinder the development of efficient generative models for medical imaging [1]. The clinical demand for high-quality,

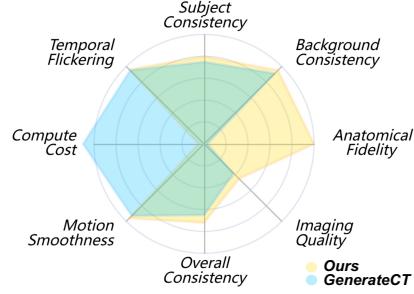


Figure 1. TRACE delivers **Flexible-Length** CT volumes, greatly cuts **Compute Cost**, boosts **Anatomical Fidelity**.

privacy-preserving diagnostic support has led to the exploration of generative methods that can effectively learn from small datasets [26]. This approach not only addresses data scarcity but also optimizes computational resources to enhance diagnostic precision and personalized care [37].

3D diffusion models have demonstrated potential in medical image generation by preserving data privacy. To simplify model design and enhance computational efficiency, most implementations assume isotropic generation dimensions[17, 70]. However, 3D CT volumes often exhibit anisotropic shapes along the axial direction, with scan length and slice count adjusted according to the region of interest and pathology extent [65]. Consequently, the isotropic assumption in traditional 3D diffusion models may limit the precise characterization of spatial distribution features inherent to Chest CT volumes [45]. Additionally, clinical applications demand high anatomical fidelity, requiring generated images to accurately capture organ shapes and details while allowing controlled variability to reflect individ-

*Corresponding author

ual differences, however, previous methods often struggle to enforce these anatomical constraints, leading to generated slices that may not faithfully represent complex anatomical structures, thus limiting clinical applicability [9, 28, 62]. Finally, spatiotemporal consistency across slices is essential to maintain anatomical alignment throughout the volume, whereas achieving this continuity imposes substantial computational costs, significantly increasing training time and model complexity, which makes generating 3D CT images with both anatomical fidelity and spatiotemporal coherence a persistent challenge [62].

In this paper, we propose TRACE, a novel framework for 3D CT image generation leveraging 2D diffusion models within a video generation paradigm to address the challenges of anatomical fidelity, computational cost, and spatiotemporal consistency in medical imaging. Our approach transforms 3D CT volume generation into a sequential process of 2D slice synthesis, treating each slice as a frame in a video. Additionally, we implement an overlapping frame guidance strategy during inference to ensure temporal coherence and spatial alignment across slices while avoiding the computational burden of traditional 3D models. Moreover, we introduce multimodal conditional guidance, incorporating detailed anatomical priors from segmentation masks and semantic guidance from radiology reports, to ensure anatomical accuracy in each axial slice and maintain consistent anatomical structures throughout the 3D volume, generating 3D CT images that capture realistic anatomical features with controlled variability, thereby ensuring anatomical fidelity, spatiotemporal consistency, and clinical relevance. To our knowledge, this is the first work to generate variable-length medical imaging videos using a conditionally guided 2D generation model. Our contributions are summarized as follows:

- Enabling the generation of 3D CT volumes with flexible axial resolution, we propose a novel framework that models 3D medical volumes as sequences of continuous frames, leveraging multi-conditioned 2D diffusion to achieve substantial computational efficiency.
- Achieving anatomical fidelity and spatiotemporal consistency, our framework integrates frame skipping and positional encoding for temporal coherence, spatial alignment through optical flow, and anatomical priors derived from segmentation masks and radiology reports. This ensures precise alignment of each frame and its transitions with anatomical structures in the generated volumes.
- We enhance previous evaluation methods on this task by introducing an anatomical fidelity assessment for generated volumes. Extensive experiments demonstrate our method’s robust performance on limited training data, achieving substantial improvements in anatomical fidelity while reducing training and inference costs by 87.5% and 92.5%, respectively.

TRACE supports essential applications in medical imaging, including data augmentation, where anatomically fidelity synthetic images enhance training sets and improve model generalization in data-limited contexts. In privacy-preserving data sharing, TRACE enables research without compromising patient identity. For personalized medicine, clinicians can define anatomy via text and masks, aiding diagnosis, surgical planning, and training. These applications underscore TRACE’s utility in expanding data resources, safeguarding privacy, and supporting clinical workflows.

2. Related Work

2.1. Medical Generative Modeling

With the growing maturity of deep learning [6–8, 11, 12, 32–34, 40, 41, 46, 56, 67, 68], generative modeling has witnessed rapid advances. Recent advancements in medical generative modeling have leveraged diffusion-based methods [14, 25, 39, 44, 47], demonstrating impressive capabilities in producing refined medical images. Conditional diffusion models, specifically text-guided diffusion models [5, 31], generate medical images based on textual descriptions, aligning the outputs with specific clinical narratives and supporting applications such as data augmentation and diagnostic training. Text-conditioned video generation, including diffusion-based and transformer-based autoregressive methods, has seen notable progress [20–22, 49, 55, 60, 61]. Diffusion-based models typically employ 3D U-Nets to generate short, low-resolution videos with fixed frame counts, subsequently enhancing resolution and duration through cascaded techniques [20]. In contrast, transformer-based models offer flexibility with variable frame counts and longer sequences, though often at lower resolutions [63]. These advances have inspired adapting video generation techniques to 3D medical imaging, treating 2D slices as sequential frames. GenerateCT [15] integrates transformer and diffusion models to synthesize coherent 3D CT volumes by employing spatial and temporal attention. However, its high computational cost arises from using a causal vision transformer for 3D encoding, vision-language alignment, and cascaded diffusion for super-resolution, with each stage contributing to the model’s complexity.

2.2. Semantic Synthesis

Semantic synthesis, or generating images from segmentation masks, is a critical task in medical image generation, enabling precise control over anatomical structures in synthesized images. Traditional image-to-image translation methods, including GAN-based models [3, 10, 64] and diffusion-based models [49, 59, 69], have been employed for this purpose but often lack the ability to enforce strict pixel-wise anatomical constraints essential for medical applications. Recent works have attempted to

fine-tune large pre-trained latent diffusion models (LDMs) for segmentation-conditioned synthesis on natural images [57, 69]; however, these models adapt poorly to medical images due to the complex anatomical variations and potential loss of spatial details when converting conditioning masks into abstract latent spaces. To address these challenges, Konz et al. [28] proposed SegGuidedDiff, a diffusion model that conditions multi-class anatomical segmentation masks at each sampling step, ensuring images adhere closely to anatomical constraints. Similarly, LeFusion [29] introduces a lesion-focused diffusion model that synthesizes images by concentrating on lesion areas, improving performance on cardiac MRI datasets. In histopathology image synthesis, methods such as InsMix [36] and DiffMix [43], which apply random label perturbations to generate images from modified labels, while others generate randomly distributed labels for image synthesis [2, 23, 66]. Although these approaches enhance control over the synthesis process, they often focus on single-class data or involve complex two-stage diffusion processes, which may not capture the spatial and structural correlations essential for anatomically consistent 3D medical images.

3. Methods

We propose a novel 3D CT generation approach leveraging a 2D diffusion model within a video generation framework. Our method captures diverse temporal relationships through variable interval frame pairs and ensures smooth transitions through an Overlapping Frame Guidance strategy. High-fidelity generation is achieved by integrating optical flow for temporal coherence, text prompts for semantic context, positional embeddings for temporal positioning, and anatomical priors from NVIDIA VISTA3D [18] for structural accuracy, resulting in anatomical fidelity and customized CT synthesis.

3.1. Preliminaries

3.1.1 Diffusion Probabilistic Models (DPMs)

DPMs are generative models that approximate the data distribution by learning a reverse denoising process [51]. The forward process progressively adds Gaussian noise to the data, transitioning from the original data to pure noise over T steps, defined as:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}), \quad (1)$$

where β_t is a variance schedule. The reverse process learns to remove this noise step-by-step.

Denoising Diffusion Implicit Models (DDIM) enable faster sampling by modifying the forward process, allowing deterministic generation:

$$\mathbf{x}_{t-1} = \sqrt{\alpha_{t-1}} f_\theta(\mathbf{x}_t, t) + \sqrt{1 - \alpha_{t-1}} \epsilon_\theta(\mathbf{x}_t, t), \quad (2)$$

where setting $\sigma_t = 0$ ensures consistent generation, useful for video-like applications.

3.1.2 3D Medical Imaging as Video

To enable flexible axial resolution, allowing the generation of 3D CT volumes with variable numbers of slices, we treat 3D medical images as a sequence of 2D slices, analogous to frames in a video. Let $V \in \mathbb{R}^{H \times W \times D}$ represent a 3D volume, where H , W , and D denote height, width, and depth (number of slices), respectively. By considering the depth dimension as a temporal axis, diffusion models can effectively generate 3D data. This approach not only provides customizable axial lengths to meet diverse clinical requirements but also significantly reduces computational costs by producing only the necessary slice sequences.

3.2. Conditional 2D Diffusion for 3D CT Synthesis

3.2.1 Paired Frame Temporal Modeling

In our framework, we utilize a 2D diffusion model to generate 3D CT image sequences by operating on frame pairs. Specifically, each consecutive frame pair is concatenated along the channel dimension, forming an input tensor of shape $(B, 2C, H, W)$, where B is the batch size, C is the original number of channels (e.g., grayscale or RGB), and H, W denote the spatial dimensions. By concatenating two frames as separate channels, the model's convolutional filters can jointly process both frames, allowing it to detect spatial-temporal patterns, such as motion or changes between frames, through standard 2D convolutions. This approach enables the model to capture temporal dynamics effectively while maintaining the computational efficiency of a 2D architecture. Unlike conventional 3D diffusion models that explicitly handle the temporal dimension, our method leverages the channel dimension to encode temporal information by treating each frame pair as correlated features. This formulation encourages the model to establish relationships between the two frames, effectively modeling temporal continuity without requiring an explicit temporal dimension.

To further enhance the model's ability to capture both short and long-term spatial dependencies, we introduce a frame-skipping strategy designed to adapt to the Z-axis spacing (Z-spacing) characteristics typical of CT imaging. In real clinical settings, CT images often vary in Z-spacing depending on anatomical requirements [42], and in our dataset, most samples are set to spacings of 0.7, 1.5, or 3 mm. To align with this characteristic, our frame-skipping intervals are set to 1, 2, and 4, respectively, simulating the variation in Z-spacing. Formally, this strategy is defined as:

$$P = \{(i, j) \mid j = i + k, i \bmod k = 0, k \in \{1, 2, 4\}\}. \quad (3)$$

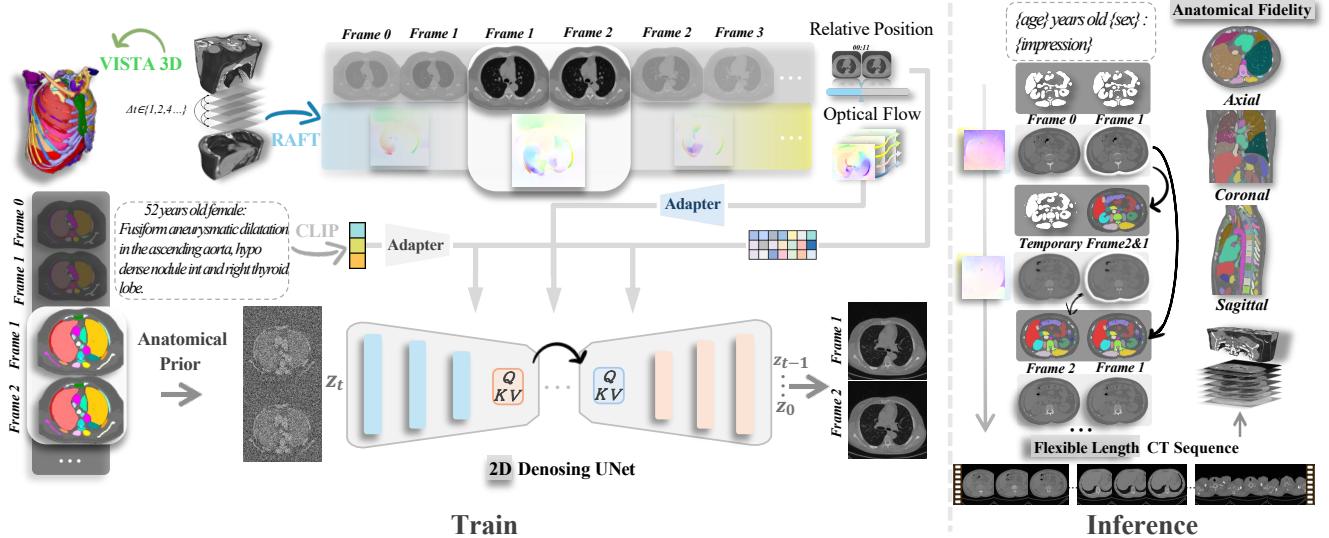


Figure 2. TRACE models 3D CT volumes as sequences of frames, utilizing an efficient 2D diffusion model conditioned on multiple modalities to generate flexible length, coherent CT sequences. During training, it denoises frame pairs with varying skip intervals, guided by four modalities: anatomical masks (VISTA3D), optical flow between frames (RAFT), report embeddings (CLIP), and relative position embeddings. The optical flow and text embeddings pass through trainable adapters before entering the diffusion model. Inference employs an overlapping-frame guidance strategy to synthesize semantically aligned frame pairs, generating anatomically consistent CT sequences, which are then reconstructed back to 3D volumes.

By training on these diverse frame pairs, the model gains a richer understanding of temporal relationships, reinforcing spatial consistency across different temporal spans and resulting in more coherent and robust outputs during inference.

For each frame pair $(i, j) \in P$, we compute a dense optical flow field $f_{i \rightarrow j}$ (mapping a pixel in frame i to its corresponding location in frame j) using RAFT [53]. This optical flow is then incorporated into the model through a trainable adapter, providing additional guidance for temporal alignment between frames.

The learned temporal coherence can be further explained through the shared representation in the latent space. During training, the concatenated frames are treated as a joint entity, and the denoising process iteratively updates them to minimize inconsistencies. Let $\mathbf{z}_t^{(i)}$ and $\mathbf{z}_t^{(j)}$ represent the latent representations of frames i and j at time step t . The temporal consistency learned during training can be interpreted as the minimization of an objective function that enforces similarity between the two frames in the latent space:

$$T_{\text{coherence}} = \mathbb{E}_{(i,j) \in P} \left[\|\mathbf{x}_t^{(i)} - \mathbf{x}_t^{(j)}\|^2 \right]. \quad (4)$$

This objective encourages the latent representations of the two frames to be similar, thus enforcing temporal coherence during both training and inference.

3.2.2 Anatomical Guidance

We condition the generation of an image $\mathbf{x}_0 \in \mathbb{R}^{c \times h \times w}$ on a multi-class anatomical mask $\mathbf{m} \in \{0, \dots, n-1\}^{h \times w}$, where c is the number of channels, h and w are the height and width of the image, n represents the number of anatomical classes, including background. These anatomical masks, obtained from VISTA3D with 127 classes, cover comprehensive human structures, including organs, bones, and lesions. We aim to sample from the distribution $p(\mathbf{x}_0 | \mathbf{m})$, ensuring the generated images adhere to the provided anatomical context.

The forward noising process $q(\mathbf{x}_t | \mathbf{x}_{t-1})$ remains unchanged. However, the reverse process $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{m})$ and the noise prediction network ϵ_θ are conditioned on the anatomical mask. This conditioning modifies the denoising process to incorporate anatomical priors effectively. The training objective is formulated as:

$$L_m = \mathbb{E}_{(\mathbf{x}_0, \mathbf{m}), t, \epsilon} \left[\|\epsilon - \epsilon_\theta(\sqrt{\alpha_t} \mathbf{x}_0 + \sqrt{1-\alpha_t} \epsilon, t, \mathbf{m})\|^2 \right], \quad (5)$$

where x_0 is paired with its corresponding anatomical mask m for each training instance. This objective ensures that the anatomical structures provided by the mask are respected during the generation.

For paired frame generation, anatomical masks for both frames are computed using VISTA3D and concatenated channel-wise with the network input at every denoising step, allowing the model to access anatomical information at the pixel level. Let $(\mathbf{m}_i, \mathbf{m}_j)$ denote the anatomical

masks for frame pair (i, j) . The conditioning is integrated as follows:

$$\epsilon_\theta(\mathbf{x}_t^{(i,j)}, t \mid \mathbf{m}_i, \mathbf{m}_j) : \mathbb{R}^{(2c+2) \times h \times w} \rightarrow \mathbb{R}^{2c \times h \times w}. \quad (6)$$

Additionally, we incorporate semantic information from textual guidance to enhance anatomical fidelity further. The text prompt includes patient-specific information such as age, gender, and diagnostic impressions, providing valuable semantic cues for the model. The text is processed using a pre-trained CLIP model to extract a text embedding \mathbf{v}_t , which encodes rich semantic information.

The text embedding \mathbf{v}_t is passed through a trainable adapter designed to transform the semantic features for effective conditioning. Let $\mathbf{W}_1, \mathbf{W}_2$ represent the learnable parameters of the adapter, consisting of a series of linear transformations with nonlinear activations to capture complex semantic interactions: $v'_t : \phi_2(\mathbf{W}_2 \cdot \phi_1(\mathbf{W}_1 v_t))$, where ϕ_1 and ϕ_2 are nonlinear activation functions applied to introduce nonlinearity and improve expressiveness. The transformed vector v'_t is subsequently used to condition the model by integrating it into the encoder hidden states h_t : $h_t = h_t + v'_t$.

This dual approach, combining anatomical mask guidance and textual semantic information, ensures the generated images respect both pixel-level anatomical structures and high-level semantic context, enhancing spatial accuracy and temporal coherence, crucial for medical imaging applications.

3.2.3 Temporal Position Encoding

To encode the temporal position of each frame pair within the entire video sequence, we employ sinusoidal positional embeddings. Given the start frame f_0 and the end frame f_N , the relative position r_i for frame f_i is computed as:

$$E(r_i)_k = \sin\left(\frac{r_i}{10000\lambda^{2k/d}}\right), \quad E(r_i)_{k+1} = \cos\left(\frac{r_i}{10000\lambda^{2k/d}}\right), \quad (7)$$

where $r_i = \frac{f_i - f_0}{f_N - f_0}$ represents the normalized relative position, d is the embedding dimension, k indexes the embedding dimension, and λ is a scaling factor. This formulation provides a compact and continuous representation of the frame's temporal position.

These embeddings are concatenated for both frames in a pair and integrated as conditional guidance within the diffusion model: $\epsilon_\theta(\mathbf{x}_t^{(i,j)}, t, E(r_i, r_j))$.

This design allows the model to adjust the denoising dynamics based on the temporal position of the frame pair, enhancing temporal coherence throughout the generated sequence. By embedding temporal position implicitly into the conditioning, the model better understands the relative context, leading to smoother transitions and consistent temporal alignment in the generated CT images.

3.3. Inference via Overlapping Frame Guidance

Traditional diffusion models, often formulated as Markovian processes, result in a “memoryless” progression, where each generation step depends solely on the previous state. However, as mentioned, thoracic CT imaging typically has a relatively small Z-axis spacing to allow detailed observation of fine anatomical structures, leading to a characteristic of approximately consistent and subtle changes between frames. Similarly, generated CT images require continuity and coherence across time and frames. To address this need and align with these characteristics, we propose the overlapping frame guidance strategy, a non-Markovian approach where each generated frame directly contributes to forming subsequent frames. Through this layered guidance strategy, our method extends anatomical coherence across the entire sequence, allowing each frame to shape the next in a controlled and continuous manner, thereby fulfilling the specific requirements of CT imaging.

As shown in Algorithm 1, the inference process begins by generating the initial frame pair $(\mathbf{x}(0), \mathbf{x}(1))$, relying on the anatomical priors $M(0)$ and $M(1)$. These segmentation masks serve as essential guides, structuring the initial input as $\tilde{\mathbf{x}}_t = \text{Concat}(\mathbf{x}_t, M(0), M(1))$. This input, combining random noise with anatomical context, is passed through the DDIM process to produce the initial frames $(\mathbf{x}(0), \mathbf{x}(1))$ that anchor the generation sequence.

For each subsequent frame pair, the process leverages overlapping frame guidance to ensure continuity. Specifically, for frame $\mathbf{x}(n+1)$, the previously generated frame $\mathbf{x}(n)$ and its mask $M(n)$ are incorporated to create a guidance map $G(n)$. This map is constructed by processing $\mathbf{x}(n)$ through a transformation $\hat{\mathbf{x}}(n) = \mathcal{F}(\mathcal{H}(\mathbf{x}(n)^\gamma))$, where γ amplifies high-intensity features, \mathcal{H} selectively smooths the background, and \mathcal{F} normalizes the values. This results in $G(n) = M(n) + [1 - M(n)] \cdot \hat{\mathbf{x}}(n)$, a map that emphasizes anatomical regions while smoothing transitions elsewhere.

The generated guidance map $G(n)$, together with the segmentation mask $M(n+1)$, forms the input $\tilde{\mathbf{x}}_t = \text{Concat}(\mathbf{x}_t, G(n), M(n+1))$ for the next DDIM step, producing frames $(\tilde{\mathbf{x}}(n), \tilde{\mathbf{x}}(n+1))$. This recursive setup, detailed in the pseudocode, ensures each new frame is conditioned on the prior frame, promoting temporal coherence across the sequence $\{\mathbf{x}(n)\}_{n=0}^N$ while preserving anatomical fidelity.

Through overlapping frame guidance, each frame sequentially shapes the next, creating a cascade of context-driven synthesis that aligns well with anatomical structures and segmentation masks, achieving coherence and continuity throughout the sequence.

Table 1. Comparison of quantitative and ablation experiments evaluates scores between our method and GenerateCT across six VBench dimensions. Segmentation results of generated volumes are assessed using three common medical segmentation metrics, serving as reference indicators for anatomical fidelity evaluation.

Method	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Imaging Quality	Overall Consistency	$JI \uparrow$	$DC \uparrow$	$95HD \downarrow$	$FPS \uparrow$	Train/Infer VRAM
GT	84.58%	95.26%	99.08%	99.41%	52.89%	21.58%	100%	100%	0.00	-	-
GenerateCT [15]	77.78%	94.73%	94.01%	95.50%	48.56%	19.58%	4.98%	9.49%	220.07	1.09	8x80GB/80GB
MedSyn [62]	74.30%	91.70%	95.60%	95.70%	50.30%	20.80%	35.40%	47.80%	59.90	3.91	4x48GB/96GB
PFM	67.30%	90.70%	94.80%	73.20%	32.70%	15.30%	20.90%	25.30%	87.40	6.53	80GB/6GB
DAG	54.50%	86.40%	85.70%	90.20%	50.30%	18.50%	52.70%	59.40%	33.20	5.52	80GB/6GB
PFM, DAG	67.70%	92.80%	95.50%	96.20%	51.20%	19.10%	54.60%	63.50%	30.90	4.98	80GB/6GB
50/[1]	77.31%	92.93%	93.37%	94.82%	51.46%	21.02%	58.78%	73.43%	19.18	4.27	80GB/6GB
100/[1,2]	78.25%	94.61%	95.32%	96.04%	50.32%	19.78%	59.76%	69.38%	21.80	3.56	80GB/6GB
50/[1,2,4]	80.07%	95.08%	95.98%	97.29%	53.91%	19.33%	62.77%	73.48%	15.25	3.89	80GB/6GB
100/[1,2,4] (Ours)	80.13%	95.17%	96.07%	96.69%	52.58%	21.63%	62.94%	76.50%	13.37	3.45	80GB/6GB

Algorithm 1 Overlapping Frame Guidance Inference

```

1: Given: Total frames  $N$ , Anatomical Prior  $\{M(n)\}_{n=0}^N$ 
2: Initialize:
3:  $\tilde{\mathbf{x}}_t \leftarrow \text{Concat}(x_t, M(0), M(1))$ 
4:  $(\mathbf{x}(0), \mathbf{x}(1)) \leftarrow \text{DDIM}(\tilde{\mathbf{x}}_t)$ 
5: for  $n = 1$  to  $N - 1$  do
6:    $G(n) \leftarrow M(n) + [1 - M(n)] \cdot \mathcal{F}(\mathcal{H}(\mathbf{x}(n)^\gamma))$ 
7:    $\tilde{\mathbf{x}}_t \leftarrow \text{Concat}(\mathbf{x}_t, G(n), M(n + 1))$ 
8:    $(\tilde{\mathbf{x}}(n), \tilde{\mathbf{x}}(n + 1)) \leftarrow \text{DDIM}(\tilde{\mathbf{x}}_t)$ 
9:    $G(n + 1) \leftarrow M(n + 1) + [1 - M(n + 1)] \cdot \mathcal{F}(\mathcal{H}(\tilde{\mathbf{x}}(n)^\gamma))$ 
10:   $\tilde{\mathbf{x}}_t \leftarrow \text{Concat}(\mathbf{x}_t, G(n), G(n + 1))$ 
11:   $(\mathbf{x}(n), \mathbf{x}(n + 1)) \leftarrow \text{DDIM}(\tilde{\mathbf{x}}_t)$ 
12: end for
13: Output: Frames  $\{\mathbf{x}(n)\}_{n=0}^N$ 

```

4. Experiment and Results

4.1. Dataset and Implementation Details

Dataset. We utilized the CT-RATE [16] dataset, comprising de-identified chest CT volumes and radiology reports from 21,314 patients under diverse imaging conditions [30]. The training set consisted of 100 randomly selected CT volumes (512×512 pixels, 150–600 slices each), totaling approximately 22,000 scans. The test sets consist solely of volumes from unseen patients. We performed Hounsfield Unit (HU) value recalibration and standardized the spatial orientation and voxel spacing (X, Y, Z intervals) of the CT volumes to ensure consistency in image density and anatomical spatial accuracy. Radiology reports were parsed into text prompts formatted as “{age} years old {sex}: {impression}”.

Implementation Details. We employed the AdamW optimizer [38] with an initial learning rate of 1×10^{-5} and a cosine annealing schedule (35,000 warmup steps, minimum learning rate 1×10^{-6}). Training spanned 300 epochs with a batch size of 28, taking approximately 14 days on a single NVIDIA A100 GPU. CT slices were downsampled to 256×256 pixels for memory efficiency, which also matches the resolution of generated images and segmentation maps.

Frame pairs used skip intervals of [1, 2, 4] to enhance temporal diversity, while optical flow was computed at the original resolution. A frame embedding dimension of 64 was utilized. During inference, a batch size of 1 enabled efficient processing on GPUs with 6 GB of memory.

4.2. Quantitative Evaluation

Building on GenerateCT, we introduce an optimized evaluation framework that improves assessments of temporal consistency, frame quality, semantic alignment, and memory efficiency, adding segmentation metrics to measure anatomical fidelity, see Tab. 1.

Anatomical Fidelity. Generated 3D volumes are standardized to HU with uniform orientation and voxel spacing, then segmented with VISTA3D [18]. We quantify fidelity by comparing generated and ground truth segmentations using Dice Coefficient (DC), Jaccard Index (JI), and 95th percentile Hausdorff Distance (95HD) [52].

VBench. VBench [24] evaluates key video aspects: Subject Consistency (DINO [4]), Background Consistency (CLIP [48]), Temporal Flickering, Motion Smoothness (video interpolation priors [35]), Imaging Quality (MUSIQ [27] on SPAQ [13]), and Overall Consistency (ViCLIP [58]), all metrics scoring higher indicating better result. This framework precisely assesses CT sequence continuity and consistency.

FVD&FID. Fréchet Inception Distance (FID) measures slice-level fidelity, while Fréchet Video Distance (FVD) extends FID to video, assessing quality and temporal coherence using I3D [54]. Lower distances reflect better fidelity in medical image generation [19].

As shown in Tab. 1, our method outperforms GenerateCT and MedSyn across benchmark video evaluation metrics, notably in **Subject Consistency** and **Motion Smoothness**, demonstrating superior frame continuity and smooth motion dynamics. For anatomical fidelity, our approach achieves Dice and Jaccard scores that surpass GenerateCT by 1163% and 706%, respectively, indicating substantial improvements in anatomical accuracy. FPS reflects the inference throughput. Since our method generates 2D frame

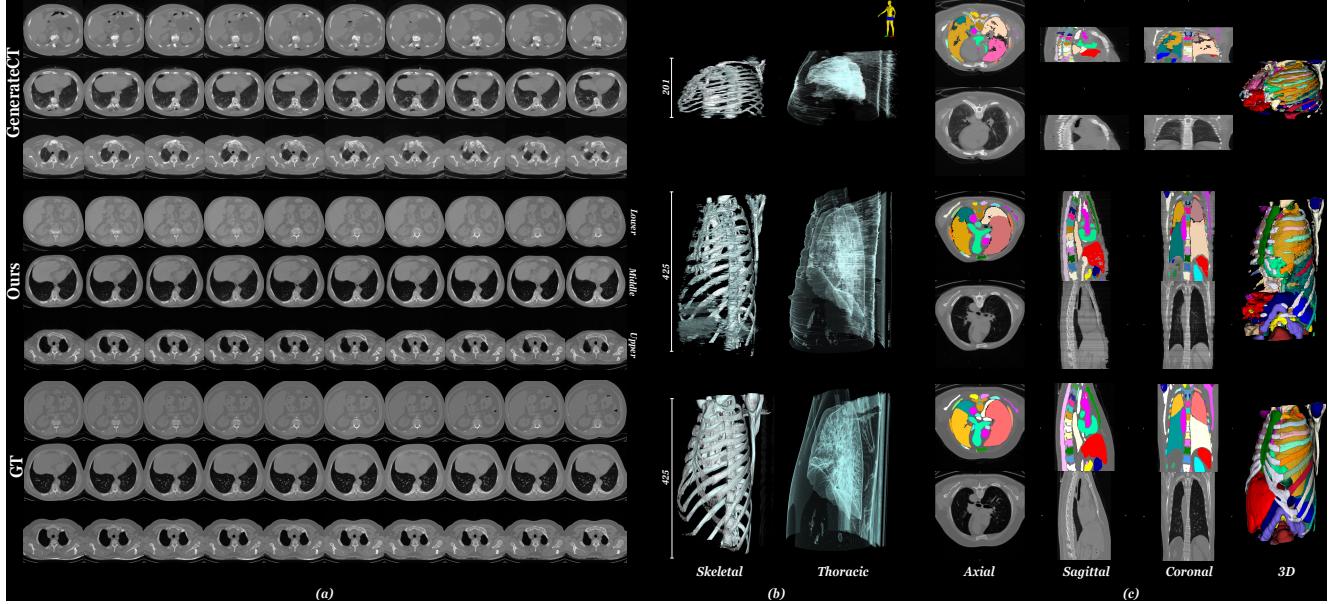


Figure 3. Comparison of generated results from multiple perspectives for “52 years old male: Fusiform dilatation in the thoracic aorta. Hepatomegaly, hepatosteatosis. Hiatal hernia. Hypodense nodule in the right thyroid lobe.” (a) Comparison of axial slices from GenerateCT, our method, and ground truth (GT), arranged left to right from diaphragm to clavicle, with each method displaying the upper, middle, and lower thorax (frames 9-18, 174-183, and 379-388). (b) 3D rendering comparison highlighting the skeleton, thoracic cavity, and key lung structures. (c) Segmentation results on generated volumes in axial, sagittal, and coronal views, with corresponding 3D renderings.

pairs, we report the average slices per second at the CT volume level for fair comparison across methods, following the official settings and using a single NVIDIA H20 GPU (96GB) for all methods. Regarding computational efficiency, TRACE only requires a GTX 1660 Ti (6GB) for inference and one A100 (80GB) for training, while GenerateCT needs 8 A100s (80GB each) for training and one A100 for inference, highlighting our efficiency advantage.

4.3. Qualitative Results

To demonstrate the effectiveness of TRACE in aligning with the requirements of clinical 3D medical image generation, we designed qualitative experiments focusing on spatiotemporal and anatomical consistency when generating sequences of flexible lengths.

Spatiotemporal Consistency. Fig. 3 (a) illustrates the temporal coherence of the generated sequences and the adaptability of TRACE in producing volumes of variable lengths. We extracted multiple axial slices from three key anatomical regions, extending from the diaphragm to the clavicle, including the lower thorax (lung base), middle thorax (lung hilum), and upper thorax (lung apex). As seen in the middle row of Fig. 3 (a), TRACE demonstrates smooth transitions across frames, capturing subtle anatomical variations while maintaining spatial consistency throughout the volume. This consistency is evident in the continuity of anatomical structures such as ribs, vertebrae, and lung fields across adjacent slices. In contrast, GenerateCT exhibits no-

ticeable discontinuities and inconsistencies between frames, with abrupt anatomical changes that may impede clinical interpretation. Furthermore, TRACE generates volumes of arbitrary lengths without sacrificing temporal coherence, an essential feature in medical imaging, where scan lengths vary according to diagnostic requirements. The consistent quality across sequences of different lengths highlights the scalability and robustness of our approach.

Anatomical Fidelity. We evaluate anatomical fidelity from both orthogonal planes and a 3D perspective. Segmentation results are derived from the VISTA3D segmentation model applied to the generated volumes, allowing for anatomical consistency assessment through key organ and skeletal structure recognition. The generated volumes were standardized to HU with aligned spatial orientation and voxel spacing, using DICOM metadata parameters from the ground truth, ensuring a consistent basis for comparison. For 2D evaluation (Fig. 3 (c)), we extract and compare the segmentation and original images from Axial, Sagittal, and Coronal views generated by TRACE and GenerateCT. TRACE accurately represents anatomical structures and pathologies, maintaining appropriate tissue contrast and correct anatomical positioning, ensuring slice-by-slice consistency. In contrast, GenerateCT may present distortions or misaligned anatomical features, with organs or skeletal structures incorrectly positioned or missing, which limits its diagnostic relevance. For 3D evaluation, as shown in Fig. 3 (b), TRACE produces detailed and spatially co-

herent renderings of thoracic anatomy, including the lung parenchyma, airways, rib cage, and vertebral column, while retaining the original axial length. In contrast, GenerateCT is limited to 201 axial slices compared to 425 in the original volume, resulting in significant information loss. Figure Fig. 3 (c-3D) further contrasts the 3D segmentations of the thoracic region, showing that TRACE preserves known anatomical structures effectively, whereas GenerateCT’s reduced axial resolution and positional inconsistencies lead to less reliable segmentation, most notably with complete omission of structures like the sternum, small intestine, and colon.

4.4. Additional Comparison Study

Existing Methods. We selected four state-of-the-art methods: Imagen [50], Stable Diffusion (SD) [49], Phenaki [55], and GenerateCT [15]. Imagen is a 2D image generation method conditioned on text prompts, generating high-resolution medical images slice by slice based on slice indices and textual cues. Stable Diffusion is a pre-trained 2D text-to-image model fine-tuned for medical image generation to assess its performance in 3D generation tasks. Phenaki, an advanced text-to-video generation model, was adapted for 3D chest CT image generation to evaluate its effectiveness in this domain. GenerateCT provided reference results for fine-tuning and experimentation with the above methods as the first framework to generate high-resolution 3D chest CT images based on natural language prompts and our previous work.

Results and Effectiveness. Compared to Imagen and Stable Diffusion, as shown in Fig. 4, the coronal and sagittal planes reflect superior spatiotemporal coherence in our method. Leveraging optical flow-guided frame skipping and overlapping frame guidance, our model effectively captures spatiotemporal correlations, as evidenced by the notably lower FVD_{I3D} in Tab. 2: 1176.7 for our method versus 3557.7 for Imagen, 3513.5 for Stable Diffusion, and 1886.8 for Phenaki, demonstrating robust coherence across frames. While Phenaki offers some spatial consistency, it lacks the anatomical detail required for clinical use, whereas our approach achieves an optimal balance between resolution, computational efficiency, and anatomical fidelity.

Compared to GenerateCT, our method demonstrates notable improvements in anatomical fidelity, primarily due to anatomical priors provided by reports and masks, which guide the generation to produce images with structural integrity in the thoracic region and clearer abdominal organs, closely approximating real images. As shown in Fig. 4, GenerateCT exhibits substantial differences in scale across sagittal slices compared to our method and the ground truth, focusing only on the lung area, which leads to a loss of critical information and less clarity in abdominal organs. Additionally, in Tab. 2, GenerateCT’s performance in FID at

Table 2. Quantitative comparison of our method with baseline methods evaluated using the framework from GenerateCT.

Method	Out	FID \downarrow	FVD _{I3D} \downarrow	CLIP \uparrow
GT/GT	-	-6.7	472.778	29.9
Base w/ Imagen [50]	2D	160.8	3557.7	24.8
Base w/ SD [49]	2D	151.7	3513.5	23.5
Base w/ Phenaki [55]	3D	104.3	1886.8	25.2
Base w/ GenerateCT [15]	3D	127.4	1382.4	27.4
Ours	3D	98.6	1176.7	29.4

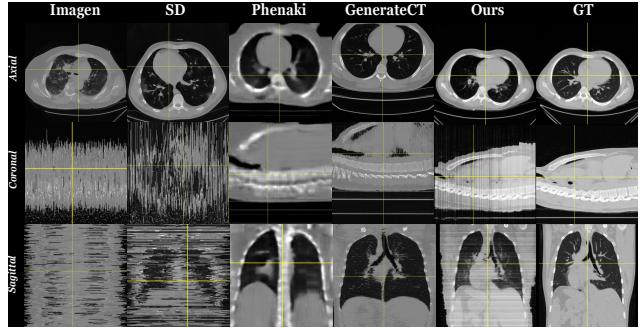


Figure 4. Axial, sagittal, and coronal slices of 3D CT volumes generated by various methods, case: “26-year-old male: Findings compatible with COVID-19 pneumonia”.

127.4 and CLIP score at 27.4 slightly lag behind ours, further validating our comprehensive advantages in anatomical accuracy, image quality, and semantic consistency. Noted that the negative FID value stems from highly similar slices.

4.5. Ablation Study

To facilitate modular analysis, we abbreviated key components of TRACE as follows: (i) Paired Frame Modeling (PFM, Sec 3.2.1) vs. single-frame generation, (ii) Dual Anatomical Guidance (DAG, Sec 3.2.2) vs. unconditional generation, and (iii) Overlapping Frame Guidance (OFG, Sec 3.3) vs. Traditional Markovian inference. As depicted in Tab. 1 (row 5-7), PFM improves background consistency and reduces flickering by leveraging inter-frame dependencies, while DAG enhances imaging quality with explicit anatomical priors. Their combination yields smoother motion but lacks long-range interactions, which OFG addresses by recursively fusing local and global cues for coherent anatomy.

Anatomical Mask Granularity.

We further evaluated mask granularity by comparing (i) a two-class setup (all anatomical structures vs. background) as shown in Fig. 5, row 1, (ii) a 128-class detailed mask setup in Fig. 5, row 2, and (iii) a three-class setup (lungs, other organs/skeletal structures, background) shown in Fig. 3(a). The results indicate that the three-class setup provides the best consistency and quality. In contrast, the

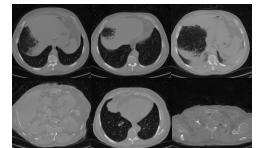


Figure 5. Ablation results for anatomical mask granularity.

Table 3. Expert Evaluation Scores for TRACE-Generated 3D Chest CT Volumes by Two Radiologists.

Evaluation Criteria	Radiologist 1	Radiologist 2
Anatomical Fidelity		
Structural Accuracy	8.8 ± 0.9	8.5 ± 1.0
Tissue Contrast	9.0 ± 0.8	8.9 ± 0.7
Pathological Representation	7.8 ± 1.1	7.5 ± 1.2
Spatiotemporal Consistency		
Intra-slice Consistency	8.9 ± 0.7	9.1 ± 0.6
Inter-slice Continuity	9.2 ± 0.6	9.0 ± 0.7
Temporal Artifacts	8.7 ± 0.8	8.8 ± 0.8
Overall Diagnostic Utility		
Overall Image Quality	8.5 ± 1.0	8.7 ± 0.9
Confidence Level	8.6 ± 0.9	8.8 ± 0.8

two-class setup reduces the separation between the lungs and other organs, leading to misclassifications (e.g., generating the spleen as lung tissue) and disrupting lung continuity. The 128-class setup performs well in the middle thorax region but fails to converge in the upper and lower thorax, resulting in incomplete or abnormal structures in these areas.

Skipping Intervals and Sample Size. To evaluate the impact of training sample size on model performance and to assess the role of frame skipping in capturing varying CT volume Z-axis spacing and enhancing spatial understanding, we designed this ablation study. We tested three skip intervals: (i) consecutive pairs ($k = 1$), (ii) skips of 1 and 2 frames ($k = 1, 2$), and (iii) skips of 1, 2, and 4 frames ($k = 1, 2, 4$), using two different training sizes (50 and 100 patients). As shown in Tab. 1 (row 8-11), the combined skip interval (1, 2, 4) best captures temporal dependencies, particularly achieving the highest performance with the larger dataset (100 patients, row 11, ours).

4.6. Expert evaluation

To rigorously assess the clinical applicability and reliability of TRACE, an expert evaluation was conducted involving two board-certified thoracic radiologists with 4 and 6 years of specialized experience, respectively. The radiologists independently evaluated fifty 3D chest CT volumes generated by TRACE against the ground truth, scoring them across three primary criteria: anatomical fidelity, spatiotemporal consistency, and overall diagnostic utility, using a 10-point scale. As summarized in Tab. 3, TRACE demonstrated strong performance, achieving average scores of 8.7 for anatomical fidelity, 9.0 for spatiotemporal consistency, and 8.7 for diagnostic utility. These results highlight TRACE’s ability to meet stringent quality standards, underscoring its potential for integration into clinical workflows.

5. Conclusion

In this work, we introduce TRACE, a framework for generating anatomically faithful 3D chest CT volumes with 2D multimodal-conditioned diffusion. TRACE effectively addresses key challenges in 3D CT generation, addressing anatomical fidelity, spatiotemporal coherence, adaptable

axial resolution, and reduced computational costs. Our approach further enhances evaluation by incorporating a fidelity assessment of generated volumes. Experimental results, supported by expert evaluations, indicate that TRACE surpasses existing methods in fidelity and efficiency, highlighting its potential for clinical applications in data augmentation and personalized modeling.

References

- [1] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022. 1
- [2] Sujata Butte, Haotian Wang, Min Xian, and Aleksandar Vakanski. Sharp-gan: Sharpness loss regularized gan for histopathology image synthesis. In *2022 IEEE 19th International symposium on biomedical imaging (ISBI)*, pages 1–5. IEEE, 2022. 3
- [3] Shixing Cao, Nicholas Konz, James Duncan, and Maciej A Mazurowski. Deep learning for breast mri style transfer with limited training data. *Journal of Digital imaging*, 36(2):666–678, 2023. 2
- [4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 6
- [5] Pierre Chambon, Christian Bluethgen, Jean-Benoit Delbrouck, Rogier Van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P Langlotz, and Akshay Chaudhari. Roentgen: vision-language foundation model for chest x-ray generation. *arXiv preprint arXiv:2211.12737*, 2022. 2
- [6] Ziyi Chang, George Alex Koulis, and Hubert PH Shum. On the design fundamentals of diffusion models: A survey. *arXiv preprint arXiv:2306.04542*, 2023. 2
- [7] Ziyi Chang, He Wang, George Alex Koulis, and Hubert P. H. Shum. Large-scale multi-character interaction synthesis. In *Proceedings of the 2025 ACM SIGGRAPH*. ACM, 2025.
- [8] Shuang Chen, Amir Atapour-Abarghouei, and Hubert PH Shum. Hint: High-quality inpainting transformer with mask-aware encoding and enhanced attention. *IEEE Transactions on Multimedia*, 2024. 2
- [9] Yufei Chen, Xiaoyu Yang, Xiaodong Yue, Xiang Lin, Qi Zhang, and Hamido Fujita. A general variation-driven network for medical image synthesis. *Applied Intelligence*, 54(4):3295–3307, 2024. 2
- [10] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 2

- [11] Haoran Duan, Yang Long, Shidong Wang, Haofeng Zhang, Chris G Willcocks, and Ling Shao. Dynamic unary convolution in transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12747–12759, 2023. 2
- [12] Haoran Duan, Shuai Shao, Bing Zhai, Tejal Shah, Jungong Han, and Rajiv Ranjan. Parameter efficient fine-tuning for multi-modal generative vision models with möbius-inspired transformation. *International Journal of Computer Vision*, pages 1–14, 2025. 2
- [13] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3677–3686, 2020. 6
- [14] Alper Güngör, Salman UH Dar, Şaban Öztürk, Yilmaz Korkmaz, Hasan A Bedel, Gokberk Elmas, Muzaffer Ozbey, and Tolga Çukur. Adaptive diffusion priors for accelerated mri reconstruction. *Medical image analysis*, 88:102872, 2023. 2
- [15] Ibrahim Ethem Hamamci, Sezgin Er, Anjany Sekuboyina, Enis Simsar, Alperen Tezcan, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Furkan Almas, Irem Dogan, Muhammed Furkan Dasdelen, et al. Generatct: Text-conditional generation of 3d chest ct volumes. *arXiv preprint arXiv:2305.16037*, 2023. 2, 6, 8
- [16] Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Bastian Wittmann, Enis Simsar, Mehmet Simsar, et al. A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. *arXiv preprint arXiv:2403.17834*, 2024. 6
- [17] Kun Han, Yifeng Xiong, Chenyu You, Pooya Khosravi, Shanlin Sun, Xiangyi Yan, James S Duncan, and Xiaohui Xie. Medgen3d: A deep generative framework for paired 3d image and mask generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 759–769. Springer, 2023. 1
- [18] Yufan He, Pengfei Guo, Yucheng Tang, Andriy Myronenko, Vishwesh Nath, Ziyue Xu, Dong Yang, Can Zhao, Benjamin Simon, Mason Belue, et al. Vista3d: Versatile imaging segmentation and annotation model for 3d computed tomography. *arXiv preprint arXiv:2406.05285*, 2024. 3, 6
- [19] 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. 6
- [20] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2
- [21] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.
- [22] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 2
- [23] Le Hou, Ayush Agarwal, Dimitris Samaras, Tahsin M Kurc, Rajarsi R Gupta, and Joel H Saltz. Robust histopathology image analysis: To label or to synthesize? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8533–8542, 2019. 3
- [24] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 6
- [25] Lan Jiang, Ye Mao, Xiangfeng Wang, Xi Chen, and Chao Li. Cola-diff: Conditional latent diffusion model for multi-modal mri synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 398–408. Springer, 2023. 2
- [26] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Haci-haliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88:102846, 2023. 1
- [27] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 6
- [28] Nicholas Konz, Yuwen Chen, Haoyu Dong, and Maciej A Mazurowski. Anatomically-controllable medical image generation with segmentation-guided diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 88–98. Springer, 2024. 2, 3
- [29] Alain Lalande, Zhihao Chen, Thibaut Pommier, Thomas Decourselle, Abdul Qayyum, Michel Salomon, Dominique Ginhac, Youssef Skandarani, Arnaud Boucher, Khawla Brahim, et al. Deep learning methods for automatic evaluation of delayed enhancement-mri. the results of the emidec challenge. *Medical Image Analysis*, 79:102428, 2022. 3
- [30] 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. 6
- [31] Hyungyung Lee, Wonjae Kim, Jin-Hwa Kim, Tackeun Kim, Jihang Kim, Leonard Sunwoo, and Edward Choi. Unified chest x-ray and radiology report generation model with multi-view chest x-rays. *arXiv preprint arXiv:2302.12172*, 3(7):8, 2023. 2
- [32] Ruochen Li, Stamos Katsigianis, Tae-Kyun Kim, and Hubert PH Shum. Bp-sgcn: Behavioral pseudo-label informed sparse graph convolution network for pedestrian and heterogeneous trajectory prediction. *IEEE Transactions on Neural Networks and Learning Systems*, 2025. 2
- [33] Ruochen Li, Tanqiu Qiao, Stamos Katsigianis, Zhanxing Zhu, and Hubert PH Shum. Unified spatial-temporal edge-

- enhanced graph networks for pedestrian trajectory prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.
- [34] Yuchen Li, Fan Wan, and Yang Long. Sid-nerf: Few-shot nerf based on scene information distribution. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024. 2
- [35] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9801–9810, 2023. 6
- [36] Yi Lin, Zeyu Wang, Kwang-Ting Cheng, and Hao Chen. Insmix: towards realistic generative data augmentation for nuclei instance segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 140–149. Springer, 2022. 3
- [37] Yanbin Liu, Girish Dwivedi, Farid Boussaid, and Mohammed Bennamoun. 3d brain and heart volume generative models: A survey. *ACM Computing Surveys*, 56(6):1–37, 2024. 1
- [38] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [39] Qing Lyu and Ge Wang. Conversion between ct and mri images using diffusion and score-matching models. *arXiv preprint arXiv:2209.12104*, 2022. 2
- [40] Xingyu Miao, Haoran Duan, Yang Bai, Tejal Shah, Jun Song, Yang Long, Rajiv Ranjan, and Ling Shao. Laser: Efficient language-guided segmentation in neural radiance fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2
- [41] Xingyu Miao, Haoran Duan, Yang Long, and Jungong Han. Rethinking score distilling sampling for 3d editing and generation. *arXiv preprint arXiv:2505.01888*, 2025. 2
- [42] JB Moser, SL Sheard, Sue Edyvean, and I Vlahos. Radiation dose-reduction strategies in thoracic ct. *Clinical radiology*, 72(5):407–420, 2017. 3
- [43] Hyun-Jic Oh and Won-Ki Jeong. Diffmix: Diffusion model-based data synthesis for nuclei segmentation and classification in imbalanced pathology image datasets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 337–345. Springer, 2023. 3
- [44] Muzaffer Özbeý, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Öztürk, Alper Güngör, and Tolga Çukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 2023. 2
- [45] Mingjie Pan, Yulu Gan, Fangxu Zhou, Jiaming Liu, Ying Zhang, Aimin Wang, Shanghang Zhang, and Dawei Li. Diffuseir: Diffusion models for isotropic reconstruction of 3d microscopic images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 323–332. Springer, 2023. 1
- [46] Jiayao Pu, Haoran Duan, Junzhe Zhao, and Yang Long. Rules for expectation: Learning to generate rules via social environment modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):6874–6887, 2024. 2
- [47] Kunpeng Qiu, Zhiqiang Gao, Zhiying Zhou, Mingjie Sun, and Yongxin Guo. Noise-consistent siamese-diffusion for medical image synthesis and segmentation. In *CVPR*, 2025. 2
- [48] 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. 6
- [49] 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. 2, 8
- [50] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 8
- [51] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3
- [52] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15:1–28, 2015. 6
- [53] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 4
- [54] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 6
- [55] Ruben Villegas, Mohammad Babaizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*, 2022. 2, 8
- [56] Fan Wan, Xingyu Miao, Haoran Duan, Jingjing Deng, Rui Gao, and Yang Long. Sentinel-guided zero-shot learning: A collaborative paradigm without real data exposure. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2
- [57] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022. 3
- [58] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 6
- [59] Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C Cattin. Diffusion models for medical anomaly

- detection. In *International Conference on Medical image computing and computer-assisted intervention*, pages 35–45. Springer, 2022. 2
- [60] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021. 2
- [61] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Dixin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *European conference on computer vision*, pages 720–736. Springer, 2022. 2
- [62] 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. 2, 6
- [63] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 2
- [64] Junlin Yang, Nicha C Dvornek, Fan Zhang, Julius Chapiro, MingDe Lin, and James S Duncan. Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 255–263. Springer, 2019. 2
- [65] Jiancheng Yang, Yi He, Xiaoyang Huang, Jingwei Xu, Xiaodan Ye, Guangyu Tao, and Bingbing Ni. Alignshift: bridging the gap of imaging thickness in 3d anisotropic volumes. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*, pages 562–572. Springer, 2020. 1
- [66] Xinyi Yu, Guanbin Li, Wei Lou, Siqi Liu, Xiang Wan, Yan Chen, and Haofeng Li. Diffusion-based data augmentation for nuclei image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 592–602. Springer, 2023. 3
- [67] Bing Zhai, Haoran Duan, Yu Guan, Huy Phan, and Wai Lok Woo. Dsleepnet: Disentanglement learning for personal attribute-agnostic three-stage sleep classification using wearable sensing data. *IEEE Journal of Biomedical and Health Informatics*, 2025. 2
- [68] Francis Xiatian Zhang, Shuang Chen, Xianghua Xie, and Hubert PH Shum. Depth-aware endoscopic video inpainting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 143–153. Springer, 2024. 2
- [69] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 3
- [70] Lingting Zhu, Zeyue Xue, Zhenchao Jin, Xian Liu, Jingzhen He, Ziwei Liu, and Lequan Yu. Make-a-volume: Leveraging latent diffusion models for cross-modality 3d brain mri synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 592–601. Springer, 2023. 1