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

Minye Shao[1], Xingyu Miao[1], Haoran Duan[2], Zeyu Wang[3], Jingkun Chen[4], Yawen Huang[5], Xian Wu[5], Jingjing Deng[6], Yang Long[1] ✉, and Yefeng Zheng[7]

[1] Department of Computer Science, Durham University, Durham, UK
[2] Department of Automation, Tsinghua University, Beijing, China
[3] College of Computer Science and Engineering, Dalian Minzu University, China
[4] Department of Engineering Science, University of Oxford, Oxford, UK
[5] Jarvis Research Center, Tencent YouTu Lab, Shenzheng, China
[6] School of Engineering Mathematics and Technology, University of Bristol, UK
[7] Medical Artificial Intelligence Laboratory, Westlake University, Hangzhou, China

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/VinyehShaw/TRACE.

Keywords: 3D Medical Imaging · Anatomical Fidelity · Multimodal Conditionally Guided Generation

1 Introduction

3D medical image generation is critical for diagnosis, personalized treatment, and surgical planning, yet privacy concerns and data scarcity hinder model development [1,16]. Diffusion models have shown promise by preserving data privacy but often assume isotropic generation [38,10], whereas CT volumes exhibit anisotropic shapes, with varying slice counts and scan lengths [24,28]. However,

✉ Corresponding author.

ensuring anatomical fidelity and spatiotemporal consistency remains computationally expensive [34]. Prior methods struggle to enforce structural constraints, generating slices that fail to capture complex anatomical variations, thereby limiting clinical applicability [18,5].

Recent works integrate transformers and diffusion models [9], drawing inspiration from video generation [13,26], but suffer from high computational overhead. Semantic synthesis methods, including GAN and diffusion-based models [3,6,35,26,37], provide anatomical control yet often lack strict pixel-wise constraints. SegGuidedDiff [18] conditions synthesis on segmentation masks, while lesion-focused approaches [19] improve localized synthesis. However, existing methods either fail to generalize to 3D medical images or require complex multi-stage pipelines [2,14,36]. To this end, we propose TRACE, a multimodal 2D diffusion framework that synthesizes 3D CT volumes as temporally aligned 2D slice sequences with anatomical and semantic guidance.

To our knowledge, this is the first work to generate unrestricted axial-length 3D medical imaging by synthesizing consistent 2D image sequences using conditionally guided 2D diffusion. Our contributions are summarized as follows: (1) Reconceptualizing 3D medical volume generation, we introduce a framework that models volumetric data as temporally coherent sequences, enabling flexible axial resolution while significantly improving computational efficiency through multi-conditioned 2D diffusion. (2) Ensuring anatomical fidelity and structural coherence, we incorporate frame skipping and positional encoding for temporal consistency, optical flow-based spatial alignment, and multimodal priors from segmentation masks and radiology reports, preserving anatomical structures across slices. (3) Advancing evaluation paradigms for generative medical imaging, we propose an anatomical fidelity assessment for generated volumes and demonstrate that our method achieves superior anatomical accuracy while reducing training and inference costs by 87.5% and 92.5%, respectively.

2 Methodology

We propose a 3D CT generation framework that leverages a 2D diffusion model within a video generation paradigm. The term “temporal” denotes the ordered progression of axial slices scanning sequence, enabling the model to capture inter-slice continuity via variable-interval frame pairs and ensure smooth transitions through Overlapping Frame Guidance during inference time. The generation operates in a 2.5D manner without full 3D computation, while the final outputs are assembled into 3D volumes. Anatomical fidelity synthesis is guided by optical flow for continuity, positional embeddings for slice ordering, text prompts for semantic context, and anatomical priors from NVIDIA VISTA3D [11].

2.1 Conditional 2D Diffusion for 3D CT Synthesis

Paired Frame Temporal Modeling. Our framework employs a 2D diffusion model to synthesize 3D CT images by processing frame pairs. Each pair

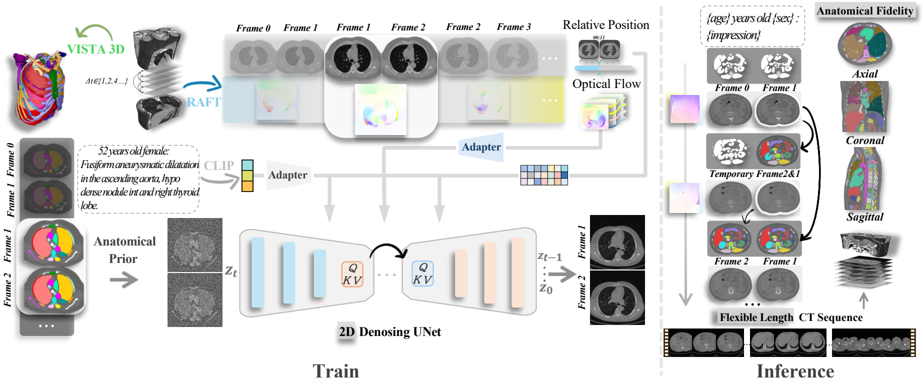

Fig. 1. TRACE models 3D CT volumes as sequences of frames, utilizing an efficient 2D diffusion model conditioned on multiple modalities to generate flexible length, coherent CT sequences. During training, it denoises frame pairs with varying skip intervals, guided by four modalities: anatomical masks (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.

is concatenated along the channel dimension, forming an input tensor of shape $(B, 2C, H, W)$, where B is the batch size, C the number of channels, and H, W the spatial dimensions. This enables convolutional filters to jointly process paired frames, implicitly capturing temporal dependencies within a 2D framework while maintaining computational efficiency. Unlike explicit 3D diffusion models, our approach encodes temporal relations through channel-wise feature interactions, enhancing temporal continuity without increasing dimensional complexity.

To accommodate CT-specific Z-spacing variations [23], we introduce a frame-skipping strategy that aligns with real-world axial resolutions. Given common spacings of 0.7, 1.5, and 3 mm in our dataset, we set frame intervals to 1, 2, and 4, respectively. Formally, the sampling strategy is defined as:

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

where (i, j) denotes a pair of frames sampled at interval k. Training with diverse frame pairs enhances the model's ability to capture short- and long-term dependencies, improving spatial consistency across slices.

For each pair $(i, j) \in P$, we compute the dense optical flow $f_{i \to j}$ using RAFT [30] and integrate it via a trainable convolutional adapter and injected through mid-block residuals to provide structural guidance for temporal alignment. The model learns coherence through latent space regularization: let $\mathbf{x}_t^{(i)}$ and $\mathbf{x}_t^{(j)}$ denote the latent representations of frames i and j at time step t. We enforce temporal consistency by minimizing the mean squared error (MSE) between latent representations of frame pairs:

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

This implicitly encourages adjacent frames to share aligned latent representations, promoting temporal smoothness across slices.

Anatomical Guidance. To explicitly preserve anatomical fidelity, we condition the generation of an image $\mathbf{x}_0 \in \mathbb{R}^{c\times h\times w}$ on a multi-class anatomical mask $\mathbf{m} \in \{0,\ldots,n-1\}^{1\times h\times w}$, where c is the channel count, h and w the spatial dimensions, and n the number of classes. Masks from VISTA3D (127 classes) cover diverse human structures, and our goal is to sample from $p(\mathbf{x}_0 \mid \mathbf{m})$ so that outputs conform to the provided anatomy. While the forward noising process $q(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ remains unchanged, both the reverse process $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{m})$ and noise prediction network ϵ_θ are conditioned on $\mathbf{m}$. Specifically, we optimize the shared MSE objective:

$$L_m = \mathbb{E}_{(\mathbf{x}_0,\mathbf{m}),t,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon} - \epsilon_\theta\left(\sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{1-\alpha_t}\boldsymbol{\epsilon},\, t,\mathbf{m}\right)\right\|^2\right], \tag{3}$$

ensuring that anatomical priors guide the denoising. For paired frame generation, anatomical masks $(\mathbf{m}_i, \mathbf{m}_j)$ for each frame pair are concatenated with the input at every denoising step so that

$$\epsilon_\theta(\mathbf{x}_t^{(i,j)}, t \mid \mathbf{m}_i, \mathbf{m}_j) : \mathbb{R}^{(2c+2)\times h\times w} \to \mathbb{R}^{2c\times h\times w}. \tag{4}$$

In addition, patient-specific text prompts are encoded by a frozen CLIP encoder into $\mathbf{v}_t$. This embedding is processed by a linear adapter,$\mathbf{v}'_t \;=\; \phi_2\big(W_2\, \phi_1(W_1\, \mathbf{v}_t)\big)$, where W_1, W_2 are learnable linear projections and ϕ_1, ϕ_2 are GELU activations. The adapter output $\mathbf{v}'_t$ is then injected via a lightweight cross-attention module into the U-Net encoder features and added to the existing hidden state: $\mathbf{h}_t = \mathbf{h}_t + \mathbf{v}'_t$, so that the model incorporates both precise anatomical priors and high-level textual semantics at each denoising step.

Temporal Position Encoding. We encode the temporal position of each frame using sinusoidal embeddings. For a sequence with start frame f_0 and end frame f_N, the normalized position of frame f_i is given by $r_i = \frac{f_i - f_0}{f_N - f_0}$. The embedding is defined as

$$E(r_i)_k = \sin\left(\frac{r_i}{10000\lambda^{2k/d}}\right), \quad E(r_i)_{k+1} = \cos\left(\frac{r_i}{10000\lambda^{2k/d}}\right), \tag{5}$$

where d is the embedding dimension, k indexes the dimensions, and λ is a scaling factor. For each frame pair (i,j), the concatenated embeddings $E(r_i, r_j)$ are used to condition the diffusion process, i.e., $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t^{(i,j)}, t, E(r_i, r_j))$. This facilitates temporal context integration, yielding smoother transitions and consistent temporal alignment in the generated CT sequences.

2.2 Inference via Overlapping Frame Guidance

Traditional diffusion models, formulated as Markovian processes, generate images step by step without memory of earlier states. In contrast, thoracic CT

Algorithm 1 Overlapping Frame Guidance Inference

```
Given: Total frames $N$, Anatomical Prior $\{M(n)\}_{n=0}^{N}$
Initialize:
$\tilde{\mathbf{x}}_t \leftarrow \text{Concat}(x_t,\ M(0),\ M(1))$
$(\mathbf{x}(0),\ \mathbf{x}(1)) \leftarrow \text{DDIM}(\tilde{\mathbf{x}}_t)$
for $n = 1$ to $N-1$ do
    $G(n) \leftarrow M(n) + [1 - M(n)] \cdot \mathcal{F}\left(\mathcal{H}\left(\mathbf{x}(n)^{\gamma}\right)\right)$
    $\tilde{\mathbf{x}}_t \leftarrow \text{Concat}(\mathbf{x}_t,\ G(n),\ M(n+1))$
    $(\tilde{\mathbf{x}}(n),\ \tilde{\mathbf{x}}(n+1)) \leftarrow \text{DDIM}(\tilde{\mathbf{x}}_t)$
    $G(n+1) \leftarrow M(n+1) + [1 - M(n+1)] \cdot \mathcal{F}\left(\mathcal{H}\left(\tilde{\mathbf{x}}(n)^{\gamma}\right)\right)$
    $\tilde{\mathbf{x}}_t \leftarrow \text{Concat}(\mathbf{x}_t,\ G(n),\ G(n+1))$
    $(\mathbf{x}(n),\ \mathbf{x}(n+1)) \leftarrow \text{DDIM}(\tilde{\mathbf{x}}_t)$
end for
Output: Frames $\{\mathbf{x}(n)\}_{n=0}^{N}$
```

imaging features small Z-axis spacing, resulting in subtle, consistent changes between slices that demand temporal and spatial coherence. To address this, we propose an overlapping frame guidance strategy, a non-Markovian approach where each generated frame directly informs the next to extend anatomical coherence across the sequence. As shown in fig. 1, overlapping frame pairs form a chain, with optical flow estimated between synthesised frames to ensure automatic alignment and continuity.

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 $\hat{\mathbf{x}}(n) = \mathcal{F}\left(\mathcal{H}\left(\mathbf{x}(n)^{\gamma}\right)\right)$, 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 emphasises 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.

3 Experiment and Results

Dataset. We use the CT-RATE [8] dataset, containing de-identified chest CT volumes and radiology reports from 21,314 patients [20]. The training set com-

Table 1. Quantitative and ablation results for 6 VBench dimensions, VRAM usage, throughput, and anatomical fidelity assessed by three segmentation metrics.

Method	SC	BC	TF	MS	IQ	OC	JI$\uparrow$	DC$\uparrow$	95HD$\downarrow$	FPS$\uparrow$	Train/Infer
GT	84.6	95.3	99.1	99.4	52.9	21.6	100.0	100.0	0.0	–	–
GenerateCT[9]	77.8	94.7	94.0	95.5	48.6	19.6	5.0	9.5	220.1	1.09	8×80GB/80GB
MedSyn[34]	74.3	91.7	95.6	95.7	50.3	20.8	35.4	47.8	59.9	3.91	4×48GB/96GB
PFM	67.3	90.7	94.8	73.2	32.7	15.3	20.9	25.3	87.4	**6.53**	80GB/6GB
DAG	54.5	86.4	85.7	90.2	50.3	18.5	52.7	59.4	33.2	5.52	80GB/6GB
PFM,DAG	67.7	92.8	95.5	96.2	51.2	19.1	54.6	63.5	30.9	4.98	80GB/6GB
50/[1]	77.3	92.9	93.4	94.8	51.5	21.0	58.8	73.4	19.2	4.27	80GB/6GB
100/[1,2]	78.3	94.6	95.3	96.0	50.3	19.8	59.8	69.4	21.8	3.56	80GB/6GB
50/[1,2,4]	80.1	95.1	96.0	**97.3**	**53.9**	19.3	62.8	73.5	14.3	3.89	80GB/6GB
PFM,DAG,OFG 100/[1,2,4](Ours)	**80.2**	**95.2**	**96.1**	96.7	52.6	**21.6**	**62.9**	**76.5**	**13.4**	3.45	80GB/6GB

prises 100 randomly selected CT volumes (512×512 pixels, 150–600 slices each; 22,000 scans total), with test volumes from unseen patients. All CT scans are preprocessed via HU recalibration, spatial orientation, and voxel spacing standardization. Radiology reports are parsed into text prompts formatted as "{age} years old {sex}: {impression}".

Implementation Details. We train our model with AdamW [22] at an initial rate of 1×10^{-5} (cosine annealed to 1×10^{-6} after 35,000 warmup steps) for 280 epochs with a batch size of 28, taking 10 days on one NVIDIA A100 GPU. CT slices are downsampled to 256×256, frame pairs are sampled at skip intervals [1, 2, 4] with optical flow computed at full resolution. Inference uses a frame embedding of 64 and a batch size of 1 on a NVIDIA 1660Ti (6GB).

3.1 Quantitative Evaluation

We extend GenerateCT with an optimized evaluation framework that quantifies temporal consistency, frame quality, semantic alignment, memory efficiency, and anatomical fidelity (see table 1).

Anatomical Fidelity. Generated 3D volumes are standardized to HU with uniform orientation and voxel spacing, then segmented via VISTA3D [11]. We assess fidelity by comparing generated and ground truth segmentations using Dice Coefficient (DC), Jaccard Index (JI) 95% Hausdorff Distance (95HD) [29].

VBench. VBench [15] evaluates key video properties: Subject Consistency (SC, DINO [4]), Background Consistency (BC, CLIP [25]), Temporal Flickering (TF), Motion Smoothness (MS, video interpolation priors [21]), Imaging Quality (IQ, MUSIQ [17] on SPAQ [7]), and Overall Consistency (OC, ViCLIP [33]). All results are reported as percentages, higher scores denote better performance.

FVD & FID. Follow GenerateCT, we use FID for slice-level fidelity and FVD (via I3D [31]) for video quality and temporal coherence; lower values indicate superior performance [12].

TRACE surpasses GenerateCT and MedSyn in Subject Consistency and Motion Smoothness. For GenerateCT, achieving approximately 705% and 1158% increase in Dice and Jaccard scores, respectively. FPS reflects the inference

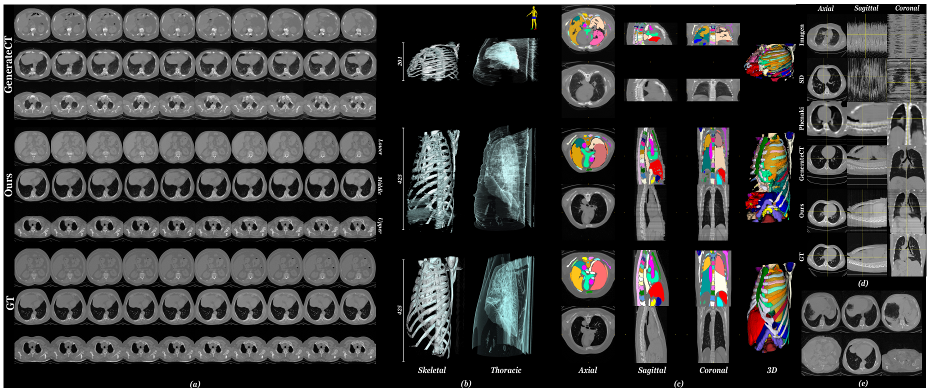

Fig. 2. Comparison of generated results for a "52-year-old male with thoracic aortic dilatation, hepatomegaly, hepatosteatosis, hiatal hernia, and a hypodense thyroid nodule". (a) Axial slices from GenerateCT, our method, and ground truth across the upper, middle, and lower thorax (frames 9–18, 174–183, 379–388), spanning diaphragm to clavicle. (b) 3D renderings of the skeleton, thoracic cavity, and lung structures. (c) Segmentation comparisons in axial, sagittal, and coronal views, with corresponding 3D visualizations. (d) Multiplanar slices from different methods for a 26-year-old male with COVID-19 pneumonia. (e) Ablation study on anatomical mask granularity.

throughput. Since our method generates 2D frame 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. The training and inference memory requirements, shown in column 12, are based on official implementations, where TRACE requires at least only a GTX 1660 Ti (6GB) for inference and a single A100 (80GB) for training, compared to 8 A100s for GenerateCT and 4 A6000s for MedSyn from their official configuration.

3.2 Qualitative Results

We conduct qualitative experiments to assess the spatiotemporal and anatomical consistency of TRACE in generating flexible-length 3D CT sequences.

Spatiotemporal Consistency. As shown in fig. 2(a), TRACE produces temporally coherent sequences with smooth transitions across frames in key regions (from diaphragm to clavicle, covering lower thorax, lung hilum, and upper thorax). In contrast, GenerateCT's result on the same case, as presented from their official release, exhibits abrupt discontinuities that may impede clinical interpretation. Notably, TRACE supports arbitrary sequence lengths without compromising coherence, demonstrating its scalability and robustness.

Anatomical Fidelity. Generated volumes are standardized to HU and segmented with VISTA3D [11]. In 2D views, as shown in fig. 2(c), TRACE maintains accurate organ structures, tissue contrast, and proper anatomical positioning across axial, sagittal, and coronal planes, while GenerateCT often exhibits

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

Method	Out	FID↓	FVD_{I3D}↓	CLIP↑
GT	-	-6.7	472.8	29.9
Imagen [27]	2D	160.8	3557.7	24.8
SD [26]	2D	151.7	3513.5	23.5
Phenaki [32]	3D	104.3	1886.8	25.2
GenerateCT [9]	3D	127.4	1382.4	27.4
Ours	3D	**98.6**	**1176.7**	**29.4**

misaligned or missing features. In 3D renderings, fig. 2(b) and fig. 2(c)-3D illustrate that TRACE preserves detailed thoracic anatomy and the original axial length, whereas GenerateCT suffers from significant information loss.

3.3 Additional Comparison Study

Existing Methods. We compare our method against four state-of-the-art approaches: Imagen [27] and Stable Diffusion (SD) [26], which generate high-resolution 2D slices based on text and slice indices; Phenaki [32], a text-to-video model adapted for 3D chest CT; and GenerateCT [9], the first framework for 3D chest CT synthesis from natural language prompts.

Results and Effectiveness. As shown in fig. 2(d), our method achieves superior spatiotemporal coherence in coronal and sagittal views when compared to the results reported by GenerateCT. By leveraging optical flow-guided frame skipping and overlapping frame guidance, our model attains a lower FVD_{I3D} of 1176.7, compared to 3557.7 for Imagen, 3513.5 for SD, and 1886.8 for Phenaki (see table 2). While Phenaki exhibits some spatial consistency, it lacks the anatomical detail required for clinical use. In contrast, our approach, guided by anatomical priors from reports and masks, produces images with enhanced structural integrity, particularly in the thoracic region and abdominal organs, closely approximating real images. GenerateCT, for instance, shows inconsistent sagittal scales and omits critical anatomical information, resulting in slightly inferior FID and CLIP scores (see fig. 2(d) and table 2). Noted that the negative FID value stems from highly similar slices. Overall, our method offers significant advantages in anatomical accuracy, image quality, and semantic consistency.

3.4 Ablation Study

To facilitate modular analysis, we abbreviated key components of TRACE as follows: (i) Paired Frame Modeling (PFM, Sec 2.1.1) vs. single-frame generation, (ii) Dual Anatomical Guidance (DAG, Sec 2.1.2) vs. unconditional generation, and (iii) Overlapping Frame Guidance (OFG, Sec 2.2) vs. Traditional Markovian inference. As depicted in table 1, 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) 2-class (anatomy vs. background, fig. 2(e) row 1), (ii) 128-class (fig. 2(e) row 2), and (iii) 3-class (lungs, other organs, background). The 3-class mask minimizes misalignment (e.g., generating the spleen as lung tissue), achieving optimal anatomical consistency.

Skipping Intervals and Sample Size. We test frame skipping with intervals of (i) $k = 1$, (ii) $k = 1, 2$, and (iii) $k = 1, 2, 4$. Notation such as "100/[1,2]" indicates a training set of 100 patients and the use of frame skip intervals of 1 and 2, respectively. As shown in table 1 (rows 8–10), the $k = 1, 2, 4$ strategy yields the best temporal consistency, particularly with a larger dataset (100 patients, row 11, ours), balancing efficiency and spatial coherence.

4 Conclusion

TRACE rethinks the problem of clinical 3D image generation by formulating chest CT volumes as temporally coherent sequences of 2D slices, enabling anatomical fidelity, flexible axial length, and computational efficiency through a multimodal 2D diffusion framework, providing a scalable solution for data generation in low-resource settings and informing future research in medical imaging.

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

References

1. Antonelli, M., Reinke, A., Bakas, S.e.a.: The medical segmentation decathlon. Nat. Commun. **13**(1), 4128 (2022)
2. Butte, S., Wang, H., Xian, M.e.a.: Sharp-gan: Sharpness loss regularized gan for histopathology image synthesis. In: Proc. IEEE Int. Symp. Biomed. Imaging. pp. 1–5 (2022)
3. Cao, S., Konz, N., Duncan, J.e.a.: Deep learning for breast mri style transfer with limited training data. J. Digit. Imaging **36**(2), 666–678 (2023)
4. Caron, M., Touvron, H., Misra, I.e.a.: Emerging properties in self-supervised vision transformers. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. pp. 9650–9660 (2021)
5. Chen, Y., Yang, X., Yue, X.e.a.: A general variation-driven network for medical image synthesis. Appl. Intell. **54**(4), 3295–3307 (2024)
6. Choi, Y., Choi, M., Kim, M.e.a.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 8789–8797 (2018)
7. Fang, Y., Zhu, H., Zeng, Y.e.a.: Perceptual quality assessment of smartphone photography. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 3677–3686 (2020)
8. Hamamci, I.E., Er, S., Almas, F.e.a.: A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. arXiv preprint arXiv:2403.17834 (2024)

9. Hamamci, I.E., Er, S., Sekuboyina, A.e.a.: Generatect: Text-conditional generation of 3d chest ct volumes. arXiv preprint arXiv:2305.16037 (2023)
10. Han, K., Xiong, Y., You, C.e.a.: Medgen3d: A deep generative framework for paired 3d image and mask generation. In: Proc. MICCAI. pp. 759–769 (2023)
11. He, Y., Guo, P., Tang, Y.e.a.: Vista3d: Versatile imaging segmentation and annotation model for 3d computed tomography. arXiv preprint arXiv:2406.05285 (2024)
12. Heusel, M., Ramsauer, H., Unterthiner, T.e.a.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Adv. Neural Inf. Process. Syst. **30** (2017)
13. Ho, J., Chan, W., Saharia, C.e.a.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
14. Hou, L., Agarwal, A., Samaras, D.e.a.: Robust histopathology image analysis: To label or to synthesize? In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 8533–8542 (2019)
15. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
16. Kazerouni, A., Aghdam, E.K., Heidari, M.e.a.: Diffusion models in medical imaging: A comprehensive survey. Med. Image Anal. **88**, 102846 (2023)
17. Ke, J., Wang, Q., Wang, Y.e.a.: Musiq: Multi-scale image quality transformer. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. pp. 5148–5157 (2021)
18. Konz, N., Chen, Y., Dong, H.e.a.: Anatomically-controllable medical image generation with segmentation-guided diffusion models. In: Proc. MICCAI. pp. 88–98 (2024)
19. Lalande, A., Chen, Z., Pommier, T.e.a.: Deep learning methods for automatic evaluation of delayed enhancement-mri. the results of the emidec challenge. Med. Image Anal. **79**, 102428 (2022)
20. Lamba, R., McGahan, J.P., Corwin, M.T.e.a.: Ct hounsfield numbers of soft tissues on unenhanced abdominal ct scans: variability between two different manufacturers' mdct scanners. Am. J. Roentgenol. **203**(5), 1013–1020 (2014)
21. Li, Z., Zhu, Z.L., Han, L.H.e.a.: Amt: All-pairs multi-field transforms for efficient frame interpolation. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 9801–9810 (2023)
22. Loshchilov, I.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
23. Moser, J., Sheard, S., Edyvean, S.e.a.: Radiation dose-reduction strategies in thoracic ct. Clin. Radiol. **72**(5), 407–420 (2017)
24. Pan, M., Gan, Y., Zhou, F.e.a.: Diffuseir: Diffusion models for isotropic reconstruction of 3d microscopic images. In: Proc. MICCAI. pp. 323–332 (2023)
25. Radford, A., Kim, J.W., Hallacy, C.e.a.: Learning transferable visual models from natural language supervision. In: Proc. Int. Conf. Mach. Learn. pp. 8748–8763 (2021)
26. Rombach, R., Blattmann, A., Lorenz, D.e.a.: High-resolution image synthesis with latent diffusion models. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 10684–10695 (2022)
27. Saharia, C., Chan, W., Saxena, S.e.a.: Photorealistic text-to-image diffusion models with deep language understanding. Adv. Neural Inf. Process. Syst. **35**, 36479–36494 (2022)

28. Shao, M., Wang, Z., Duan, H., Huang, Y., Zhai, B., Wang, S., Long, Y., Zheng, Y.: Rethinking brain tumor segmentation from the frequency domain perspective. IEEE Transactions on Medical Imaging (2025)
29. Taha, A.A., Hanbury, A.: Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. BMC Med. Imaging **15**, 1–28 (2015)
30. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: Proc. ECCV 2020. pp. 402–419 (2020)
31. Unterthiner, T., van Steenkiste, S., Kurach, K.e.a.: Fvd: A new metric for video generation (2019)
32. Villegas, R., Babaeizadeh, M., Kindermans, P.J.e.a.: Phenaki: Variable length video generation from open domain textual descriptions. In: Proc. Int. Conf. Learn. Represent. (2022)
33. Wang, Y., He, Y., Li, Y.e.a.: Internvid: A large-scale video-text dataset for multimodal understanding and generation. arXiv preprint arXiv:2307.06942 (2023)
34. Xu, Y., Sun, L., Peng, W.e.a.: Medsyn: text-guided anatomy-aware synthesis of high-fidelity 3d ct images. IEEE Trans. Med. Imaging (2024)
35. Yang, J., Dvornek, N.C., Zhang, F.e.a.: Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation. In: Proc. MICCAI 2019. pp. 255–263 (2019)
36. Yu, X., Li, G., Lou, W.e.a.: Diffusion-based data augmentation for nuclei image segmentation. In: Proc. MICCAI. pp. 592–602 (2023)
37. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
38. Zhu, L., Xue, Z., Jin, Z.e.a.: Make-a-volume: Leveraging latent diffusion models for cross-modality 3d brain mri synthesis. In: Proc. MICCAI. pp. 592–601 (2023)