Sketch2CT: Multimodal Diffusion for Structure-Aware 3D Medical Volume Generation

Delin An and Chaoli Wang
University of Notre Dame
{dan3, chaoli.wang}@nd.edu

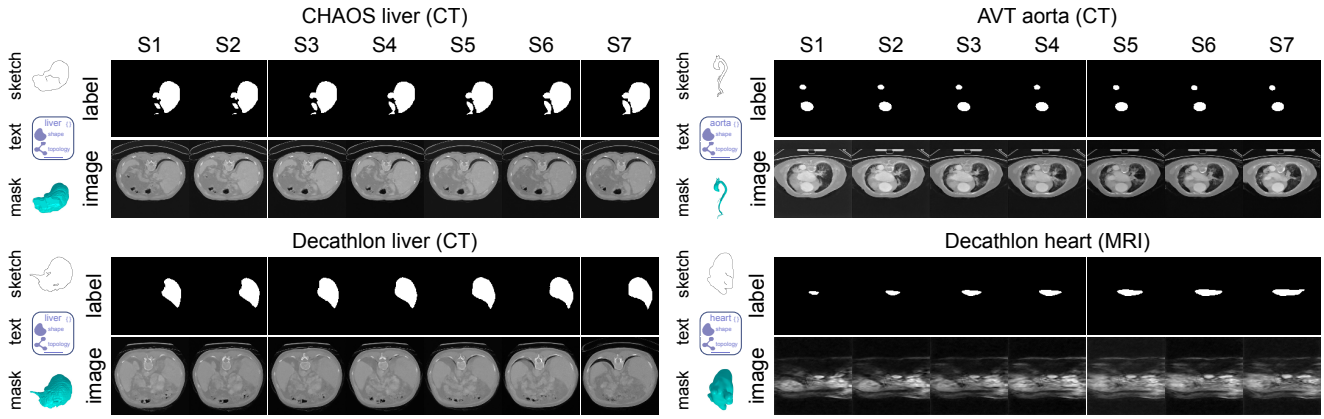


Figure 1. Multimodal sketch and text guided 3D medical image generation results from our Sketch2CT method. For each organ, a user-provided sketch and text description serve as structural and semantic conditions. Sketch2CT produces both a 3D segmentation mask and a synthesized medical volume that closely follows the geometry and anatomy implied by the input. S1-S7 denote seven consecutive axial slices from the synthesized 3D volume, illustrating spatial continuity.

Abstract

Diffusion probabilistic models have demonstrated significant potential in generating high-quality, realistic medical images, providing a promising solution to the persistent challenge of data scarcity in the medical field. Nevertheless, producing 3D medical volumes with anatomically consistent structures under multimodal conditions remains a complex and unresolved problem. We introduce Sketch2CT, a multimodal diffusion framework for structure-aware 3D medical volume generation, jointly guided by a user-provided 2D sketch and a textual description that captures 3D geometric semantics. The framework initially generates 3D segmentation masks of the target organ from random noise, conditioned on both modalities. To effectively align and fuse these inputs, we propose two key modules that refine sketch features with localized textual cues and integrate global sketch-text representations. Built upon a capsule-attention backbone, these modules leverage the complementary strengths of sketches and text to produce anatomically accurate organ shapes. The synthesized segmentation masks subsequently guide a latent diffusion model for 3D CT volume synthesis, enabling realistic reconstruction of organ appearances that are consistent with user-defined sketches and descriptions. Extensive experiments on public CT datasets demonstrate that Sketch2CT achieves superior performance in generating multimodal medical volumes. Its controllable, low-cost generation pipeline enables principled, efficient augmentation of medical datasets. Code is available at https://github.com/adlsn/Sketch2CT.

1. Introduction

Medical image synthesis has become a pivotal research field aimed at overcoming the challenge of data scarcity in medical imaging. In contrast to natural images, collecting large-scale, high-quality, and well-annotated medical datasets is hindered by privacy restrictions, significant acquisition costs, and dependence on expert annotations. Meanwhile, deep learning models, which underpin recent breakthroughs in medical image analysis tasks such as segmentation [3, 4, 14, 56], classification [8, 12], anomaly de-

tection [37, 55, 58], registration [9, 46], and cross-modality translation [30], require large volumes of labeled data for optimal performance. Thus, generating realistic and diverse synthetic medical images presents a promising avenue for enhancing model generalizability and robustness under limited data conditions.

Generative models, including generative adversarial networks (GANs) [19] and denoising diffusion probabilistic models (DDPMs) [25], have achieved significant progress in image synthesis [26], inpainting [27], and super-resolution [15]. Among these, diffusion models have garnered particular attention for their exceptional image quality, robust training, and strong capacity to represent complex data distributions. Conditional diffusion models [47] further enhance controllability by incorporating auxiliary modalities, such as text descriptions [11, 42], segmentation masks [29], or sketches [33, 43, 64], to guide generation across images, videos, and 3D point clouds. Building on their success in natural image domains, diffusion-based approaches have been adopted in medical imaging to help address the challenge of limited data.

Diffusion-based medical image synthesis can be categorized into 2D and 3D approaches, depending on the dimensionality of the generated data. 2D diffusion models [18] commonly produce individual image slices (e.g., CT or MRI), delivering visually realistic results but often failing to ensure anatomical continuity between adjacent slices. To improve realism, segmentation-guided 2D models [29] incorporate anatomical priors, enhancing local accuracy but still lacking global inter-slice consistency for full 3D volume reconstruction. Conversely, 3D diffusion models directly generate volumetric data to preserve spatial coherence; however, they require significant computational resources, limiting both achievable resolution and the scalability of the training process. Conditioning on 3D segmentation masks can enhance anatomical plausibility, although it introduces complexity and may destabilize the training process. Recently, latent diffusion models [36, 54, 63] have emerged as a promising alternative by conducting the diffusion process in a compressed latent space, reducing computational costs while retaining high fidelity.

While unconditional diffusion models are valuable for general data augmentation, generating paired image-annotation data through conditional diffusion is essential for downstream medical image analysis tasks [5, 22, 31, 45, 51]. Segmentation-guided diffusion frameworks [2, 13, 29] have demonstrated strong results for medical image synthesis; however, they are constrained by their reliance on predefined segmentation masks, which limit both the diversity and controllability of the output. To address segmentation availability constraints, methods such as MedGen3D [24] employ a two-stage pipeline in which random segmentation masks are generated before synthesizing corresponding medical volumes; however, the inherent randomness of these masks limits structural control.

To overcome these challenges, namely, (1) insufficient inter-slice consistency in 2D methods, (2) the high computational demands of full 3D diffusion, and (3) the limited controllability of current conditional models, we introduce Sketch2CT, a multimodal diffusion framework for structure-aware 3D medical volume generation. Sketch2CT comprises two stages: (1) a sketch-and-text conditioned latent diffusion model that produces anatomically consistent 3D segmentation masks from user-provided sketches and textual descriptions, and (2) a segmentation-conditioned latent diffusion model that synthesizes high-quality 3D CT volumes based on the generated masks. In our approach, sketches serve as an intuitive structural blueprint of the target anatomy, while textual descriptions provide complementary semantic and geometric details not captured by sketches alone. The fusion of these modalities enables the controllable, structure-preserving generation of realistic 3D medical volumes, as illustrated in Figure 1.

The principal contributions of this work are as follows:

- We effectively capture both local structural information from sketches and contextual semantics from text, thereby enhancing the model's attention to anatomically meaningful regions during segmentation generation.
- We jointly fuse multimodal representations from local and global perspectives, thereby facilitating deep interaction between the sketch and text modalities and improving structural fidelity.
- We enable low-cost, user-crafted sketches and text descriptions to generate high-quality, anatomically coherent 3D CT volumes, providing a practical, controllable solution for data augmentation.

2. Related Work

Generative models for natural image synthesis. Generative models can be broadly categorized into two major branches: GANs and diffusion models. GANs learn a mapping from a latent noise distribution to the target data distribution through an adversarial game between a generator and a discriminator, and have achieved remarkable success in early image synthesis tasks. However, they often suffer from training instability and mode collapse, which limits the diversity and fidelity of the generated samples [32]. Diffusion models have recently emerged as a more stable and expressive alternative, formulating data generation as the reversal of a gradual noising process. They have demonstrated superior image quality, robustness in training, and diversity compared to GANs. In the natural image domain, diffusion models have achieved state-of-the-art performance in 2D image generation and have been extended to conditional generation by introducing auxiliary modalities such as text, segmentation maps, or sketches to

control the synthesis process. Furthermore, several studies [16, 21, 34, 53] have explored generating 3D data from 2D or multimodal conditions, enabling geometry-aware synthesis and reconstruction. For instance, Wu et al. [57] proposed a sketch- and text-conditioned diffusion framework to generate colored 3D point clouds, demonstrating the effectiveness of combining coarse structural priors with semantic information.

Medical image synthesis with generative models. Compared to the extensive research on natural image generation, diffusion-based medical image synthesis remains underexplored. Most existing studies [1, 10, 23, 50, 61] focus on 2D slice-level synthesis, where diffusion models generate large numbers of CT or MRI slices with considerable diversity. However, the lack of inter-slice consistency and anatomical plausibility limits their usability for data augmentation and downstream analysis. A common strategy to improve anatomical realism is to incorporate explicit anatomical priors, such as segmentation masks, to guide the synthesis process. Following this idea, Wang et al. [54] employed pretrained latent diffusion models to generate medical images conditioned on segmentation maps, and Zhang et al. [63] proposed ControlNet to enable conditional 2D image generation. Although these methods demonstrate the benefits of conditioning, they are primarily designed for natural images and do not adapt well to medical data. Konz et al. [29] designed a segmentation-guided diffusion model (Seg-Diff) to synthesize medical images from segmentation masks, showing that the synthetic data can improve the performance of segmentation networks. Nevertheless, 2D-based approaches still struggle to maintain inter-slice continuity and cannot generate segmentation masks themselves.

To enhance controllability and data diversity, some studies generate segmentation masks as an intermediate step before image synthesis. For instance, Guibas et al. [20] used a dual-GAN architecture to produce 2D segmentation masks for retinal image generation, and Fernandes et al. [18] generated brain MRI images conditioned on masks from two coordinated generators. For volumetric synthesis, Subramaniam et al. [48] employed a Wasserstein GAN [7] to jointly generate 3D patches and their corresponding masks, which improves spatial coherence. However, the high computational cost of full 3D diffusion models makes large-scale volume synthesis challenging, particularly when incorporating 3D segmentation masks as structural priors. Han et al. [24] addressed this issue with MedGen3D, a multi-conditional diffusion framework that generates both 3D segmentation masks and medical volumes. Their approach first synthesizes sub-volumes and subsequently infers the complete volume from the generated regions, partially mitigating computational overhead.

Multimodal and structure-aware generation. Existing medical image synthesis approaches largely focus on either 2D image generation or 3D volume synthesis conditioned on segmentation masks. With the success of conditional diffusion models in computer vision, recent research has begun to incorporate multimodal guidance, such as text, sketches, or multimodal imaging, to enhance controllability and interpretability. For text-based generation, Xu et al. [60] proposed AttnGAN, and Zhang et al. [62] introduced StackGAN to synthesize images from textual descriptions through multi-stage refinement. In 3D reconstruction, image-based conditioning is widely used to recover 3D shapes from single or multi-view images [52], such as Pix2Vox++ [59] and Pixel2Mesh [53]. When image conditions are unavailable, sketches serve as an efficient and intuitive structural input. Guillard et al. [21] utilized sketches to refine object shapes, and Zhang et al. [64] reconstructed 3D models from single-view sketches. Recent works further demonstrate that combining sketches with text provides complementary information: sketches encode coarse geometry, while text conveys semantics and fine-grained details. For example, Wu et al. [57] employed capsule attention [40] to jointly process sketch and text features for controllable 3D point cloud generation.

Inspired by these advances, we aim to leverage multimodal information to guide the generation of 3D medical volumes. In particular, sketches offer intuitive structural guidance of the target anatomy, while text provides complementary semantic and morphological context. As diffusion models have proven superior to GANs in terms of image fidelity and stability, we adopt them as the backbone of our framework. We also employ a latent diffusion formulation that enables high-resolution 3D generation in a compressed latent space, thereby balancing efficiency and quality. Although latent compression may lose some fine-scale details, ensuring globally consistent and anatomically plausible 3D structure is more critical for medical image analysis tasks. To the best of our knowledge, Sketch2CT is the first framework to explore multimodal (sketch and text) diffusion for structure-aware 3D medical volume generation.

3. Sketch2CT

As shown in Figure 2, our *Sketch2CT* framework comprises two primary components: **segmentation mask generation** and **medical volume generation**. The first component is designed to synthesize 3D segmentation masks, conditioned on both sketches and textual descriptions. This stage includes four key modules: (1) **sketch extraction**, capturing the anatomical structure's outline; (2) **text acquisition**, providing complementary 3D semantic details; (3) **sketch-text feature fusion**, aligning structural and contextual information; and (4) **segmentation latent diffusion**, which generates a coherent 3D segmentation mask. The second component synthesizes realistic 3D medical volumes by conditioning a latent diffusion model on the segmentation repre-

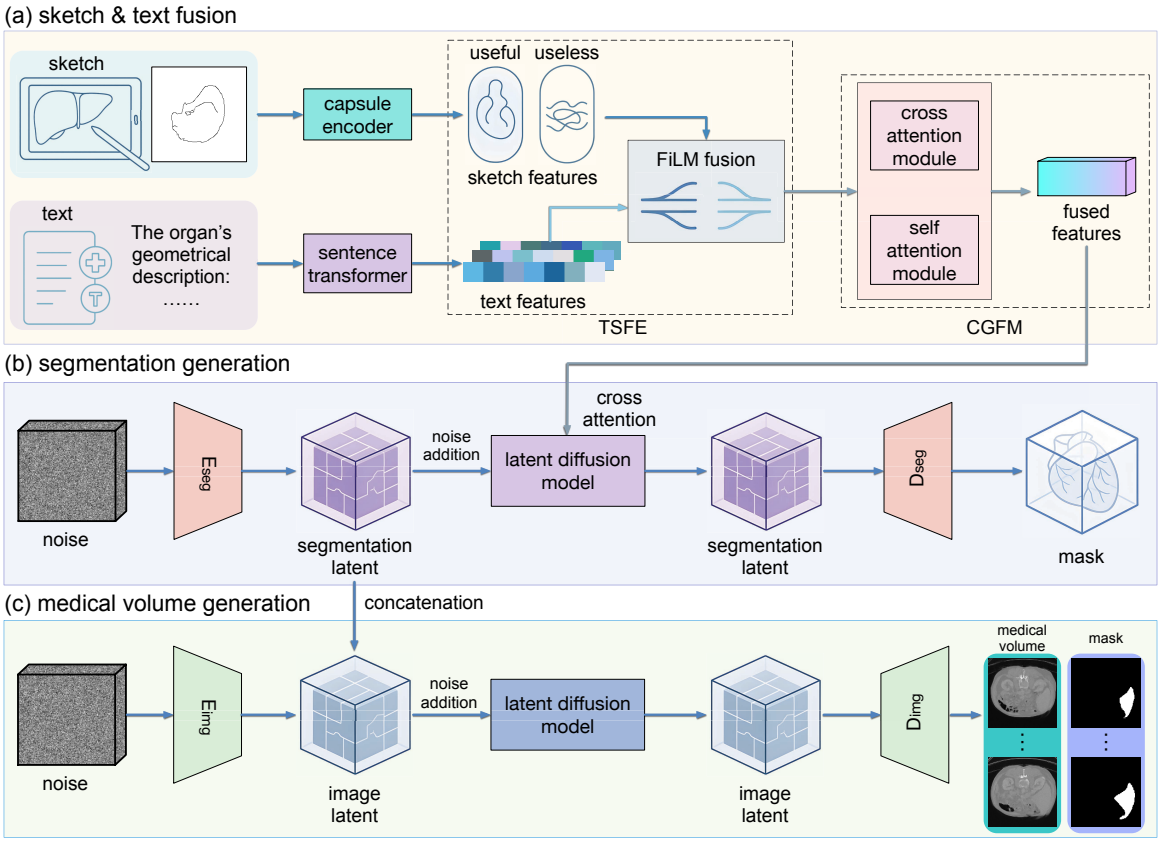


Figure 2. Overview of our *Sketch2CT* framework. (a) A capsule-based sketch encoder and a sentence transformer extract structural and semantic features, which are fused via a FiLM module. (b) The fused representation conditions a segmentation latent diffusion model to generate 3D organ masks. (c) The predicted segmentation latent guides an image latent diffusion model to synthesize 3D medical volumes.

sentation. The detailed design and implementation of each module are described in the following subsections.

3.1. Sketch Extraction

To provide intuitive and structure-aware guidance for generation, we first extract 2D sketches from 3D anatomical segmentations. The objective is to obtain representative contour depictions that capture the characteristic shape and boundary geometry of each organ, while keeping the sketches lightweight, abstract, and easily interpretable for user interaction. We begin by visualizing the 3D segmentation masks in a medical imaging environment (*3D Slicer* [17]) to obtain surface representations of the target anatomy. For volumetric organs such as liver and heart, 2D projections are captured along the **axial (transverse)** direction, which corresponds to the standard clinical viewing plane and effectively reveals the organ's overall morphology. For elongated, tubular structures such as the aorta, projections are instead taken along the **sagittal (longitudinal)** plane, which aligns with the vessel's principal axis and better preserves its continuous geometry.

The rendered projections are subsequently transformed into sketches through a sequence of image-processing operations that emphasize structural boundaries while suppressing texture and shading. Specifically, we employ the *PyVista* [49] library to render high-contrast surface projections with controlled lighting and orthogonal camera views, ensuring clear structural outlines. The rendered images are then processed using *OpenCV* operations, including grayscale conversion, histogram equalization for contrast enhancement, bilateral filtering for edge-preserving smoothing, and Canny-based edge detection to extract salient contours. The resulting binary edge maps are refined using morphological operations to yield clean, continuous sketches that delineate the organ boundaries. Through adjusting parameters such as edge detection thresholds and filtering kernel sizes, we can control the level of detail and abstraction in the sketches. These anatomy-aware sketches provide geometry-centric structural priors that condition the subsequent segmentation diffusion model.

3.2. Text Acquisition

Although sketches offer intuitive 2D structural information, they inherently lack depth and volumetric context, making

it difficult to reconstruct a complete 3D segmentation mask from a single projection. Obtaining sketches from multiple views could mitigate this limitation, but would substantially increase user effort and compromise practical usability. To balance expressiveness with simplicity, we introduce text descriptions as a complementary modality to enhance the spatial context that sketches alone cannot provide.

To construct these textual representations, we first generate a series of 2D snapshots for each segmentation mask along the three principal anatomical axes: axial, coronal, and sagittal. These projections highlight different facets of the organ's morphology and serve as comprehensive visual references for its 3D structure. Simultaneously, we calculate a suite of physical and geometric metrics from the segmentation mask, such as volume, surface area, principal-axis lengths, maximum and minimum diameters, sphericity, compactness, and, when applicable, centerline length. Together, these quantitative attributes succinctly characterize the organ's overall shape, size, and spatial proportions. During training, textual descriptions are generated automatically from this visual and geometric information to serve as textual conditions. At inference time, however, users can freely edit or provide custom text descriptions, enabling flexible, interactive control over the generation process without requiring reference segmentations.

The snapshots and geometric metrics are provided together as prompts to a large language model (LLM) specialized in visual-language understanding (GPT-4o-mini, OpenAI) [65]. The model is directed to serve as an expert geometric describer, generating a structured textual summary of the organ's shape characteristics, including global form, surface smoothness, symmetry, and topological continuity, while explicitly excluding any diagnostic or clinical context. The generated output is a structured JSON description that retains only geometry-related information. This approach ensures the resulting text focuses solely on spatial and morphological features, treating the segmentation as a generic 3D object; thus, domain-specific medical LLMs are unnecessary.

Finally, the textual descriptions are transformed into high-dimensional embeddings via a pretrained sentence transformer [39], capturing both global semantics and fine-grained geometric relationships conveyed in the text. These embeddings serve as textual conditioning features for the subsequent sketch-text fusion and segmentation-diffusion stages. By complementing sketch-based structural information with semantic representations of 3D geometry, the text modality enables more consistent, spatially aware mask generation.

3.3. Sketch-Text Feature Fusion

Although sketches provide an intuitive and compact representation of organ structure, their sparse, edge-only nature makes it challenging to extract stable, discriminative features for downstream 3D generation. Unlike natural images, medical sketches contain limited texture and shading details, and the relevant anatomical boundaries are often discontinuous or faint due to projection or noise. Directly encoding such sparse inputs with convolutional backbones can therefore lead to feature ambiguity and loss of structural cues. To address these challenges, we design two complementary modules for multimodal fusion: a **text-enhanced sketch feature extractor (TSFE)** for local text-guided refinement, and a **cross-modal global fusion module (CGFM)** for global semantic alignment between sketch and text.

TSFE aims to enrich the sparse sketch representation with semantic priors derived from textual descriptions. The sketch image is first encoded into a capsule-based embedding $\mathbf{f}_s \in \mathbb{R}^{d_s}$ using a convolutional backbone followed by primary capsules and attention-based routing, while the textual description is embedded as $\mathbf{f}_t \in \mathbb{R}^{d_t}$ using a pretrained sentence transformer. To adaptively modulate the sketch embedding, we employ a feature-wise linear modulation (FiLM) [35] mechanism that generates scale and shift parameters from the text embedding

$$\gamma, \beta = g(\mathbf{f}_t), \tag{1}$$

where $g(\cdot)$ denotes a lightweight projection network. The text-enhanced sketch feature is then obtained as

$$\tilde{\mathbf{f}}_s = \gamma \odot \mathbf{f}_s + \beta, \tag{2}$$

with $\odot$ indicating element-wise multiplication. This text-guided modulation adaptively amplifies semantically relevant channels and suppresses irrelevant ones, producing a more informative and stable sketch embedding.

To achieve global semantic alignment and joint reasoning across modalities, we introduce a two-level attention mechanism for the CGFM, comprising text-guided cross-attention and sketch-guided self-attention. Given the enhanced sketch features $\tilde{\mathbf{f}}_s$ and the text embedding $\mathbf{f}_t$, the cross-attention stage integrates fine-grained semantic features

$$\mathbf{F}_{\text{local}} = \text{Attention}(\tilde{\mathbf{f}}_s, \mathbf{f}_t, \mathbf{f}_t), \tag{3}$$

capturing localized correspondences between sketch contours and textual semantics. Subsequently, the self-attention stage aggregates these responses to form a global representation

$$\mathbf{F}_{\text{global}} = \text{SelfAttn}(\mathbf{F}_{\text{local}}), \tag{4}$$

which summarizes the organ's overall geometry and semantics. Finally, both representations are concatenated and projected to yield the joint multimodal feature

$$\mathbf{z}_{\text{fusion}} = \text{Proj}\big([\mathbf{F}_{\text{local}} \,\|\, \mathbf{F}_{\text{global}}]\big), \tag{5}$$

which serves as the conditioning input for the segmentation latent diffusion model.

Together, TSFE enhances the sparse sketch representation through text-driven channel modulation, while CGFM aligns global semantics via hierarchical attention. This two-stage fusion effectively bridges the modality gap, yielding a robust, geometry-aware multimodal embedding for generating anatomically consistent segmentations.

3.4. Segmentation Latent Diffusion

Under the multimodal conditions provided by the fused sketch-text representation, we aim to generate anatomically coherent 3D segmentation masks using a latent diffusion framework. Directly performing diffusion in voxel space is computationally demanding and memory-intensive for volumetric data. To address this, we use the 3D AutoencoderKL function from *MONAI* to project segmentation masks into a compact latent space, enabling efficient diffusion while preserving essential anatomical structures.

Formally, given a ground-truth segmentation volume $\mathbf{x}_0$, the autoencoder encodes it into a latent representation $\mathbf{z}_0 = E_{seg}(\mathbf{x}_0)$ and reconstructs it as $\hat{\mathbf{x}}_0 = D_{seg}(\mathbf{z}_0)$. During training, the latent diffusion model learns a parameterized denoising process that progressively reconstructs $\mathbf{z}_0$ from Gaussian noise $\mathbf{z}_T \sim \mathcal{N}(0, \mathbf{I})$ across T timesteps.

At each timestep t, the forward diffusion process adds Gaussian noise according to

$$q(\mathbf{z}_t \mid \mathbf{z}_{t-1}) = \mathcal{N}\big(\sqrt{\alpha_t}\,\mathbf{z}_{t-1},\, (1-\alpha_t)\mathbf{I}\big), \tag{6}$$

where $\{\alpha_t\}_{t=1}^{T}$ defines a variance schedule controlling the noise intensity. The reverse process is learned through a UNet-based denoising network ϵ_θ conditioned on the multimodal feature $\mathbf{z}_{\text{fusion}}$

$$p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, \mathbf{z}_{\text{fusion}}) = \mathcal{N}\big(\mu_\theta(\mathbf{z}_t, t, \mathbf{z}_{\text{fusion}}), \sigma_t^2\mathbf{I}\big), \tag{7}$$

where μ_θ and σ_t are learned mean and variance terms, and the conditioning is injected via cross-attention at multiple layers of the UNet.

Following the v-*prediction* parameterization [41], the network is trained to predict the velocity term

$$\mathbf{v}_t = \sqrt{\alpha_t}\,\boldsymbol{\epsilon} - \sqrt{1-\alpha_t}\,\mathbf{z}_0, \tag{8}$$

where $\boldsymbol{\epsilon}$ denotes the Gaussian noise added at step t. The training objective minimizes the mean squared error between the predicted and target velocities

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,\,\mathbf{z}_0,\,\boldsymbol{\epsilon}}\big[\|\epsilon_\theta(\mathbf{z}_t, t, \mathbf{z}_{\text{fusion}}) - \mathbf{v}_t\|_2^2\big]. \tag{9}$$

During inference, random latent noise is iteratively denoised through the learned reverse process, guided by the sketch–text condition $\mathbf{z}_{\text{fusion}}$, to produce a clean segmentation latent $\hat{\mathbf{z}}_0$. The pretrained autoencoder subsequently decodes this latent to yield the final 3D segmentation mask $\hat{\mathbf{x}}_0 = D(\hat{\mathbf{z}}_0)$.

By operating in the latent space and leveraging cross-modal conditioning, Sketch2CT efficiently synthesizes anatomically consistent segmentation masks that reflect both the structural constraints of the sketch and the semantic guidance of the text.

3.5. Conditional Medical Volume Generation

The final stage of Sketch2CT synthesizes 3D medical volumes conditioned on the generated segmentation structure. While the segmentation diffusion model reconstructs anatomical geometry, this stage translates the structural representation into corresponding image appearances, thereby bridging the gap between geometry and texture. To achieve this, we adopt a latent diffusion formulation similar to Section 3.4 but extend it to incorporate the segmentation latent as a structural prior.

Specifically, the segmentation mask $\hat{\mathbf{x}}_{\text{seg}}$ obtained from the previous stage is first encoded into a compact latent representation $\mathbf{z}_{\text{seg}} = E_{\text{seg}}(\hat{\mathbf{x}}_{\text{seg}})$ using a pretrained 3D AutoencoderKL. For the medical image volume $\mathbf{x}_{\text{img}}$, another AutoencoderKL is trained to map the medical volume into a perceptually meaningful latent space $\mathbf{z}_{\text{img}} = E_{\text{img}}(\mathbf{x}_{\text{img}})$.

During the diffusion process, the segmentation latent is concatenated with the noisy image latent at each timestep to guide the denoising trajectory

$$\mathbf{z}_{t-1} = \epsilon_\theta\big(\mathbf{z}_t \,\|\, \mathbf{z}_{\text{seg}},\, t\big), \tag{10}$$

where $\|$ denotes channel-wise concatenation. This conditioning scheme ensures that the generated image remains anatomically aligned with the segmentation, while enabling the realistic synthesis of tissue texture. The diffusion model is trained under the v-prediction parameterization following Equation 9. During inference, random noise is iteratively denoised under the segmentation guidance to produce a clean latent $\hat{\mathbf{z}}_{\text{img}}$, which is finally decoded into the voxel domain as $\hat{\mathbf{x}}_{\text{img}} = D_{\text{img}}(\hat{\mathbf{z}}_{\text{img}})$.

By conditioning the image diffusion model on the segmentation latent, Sketch2CT establishes a strong linkage between anatomical structure and visual realism. This design enables controllable, consistent generation of medical volumes, ensuring that synthesized images exhibit natural intensity variations while faithfully preserving the geometric fidelity of the generated segmentation.

4. Results and Discussion

We conduct our experiments on three public CT datasets and one MRI dataset: (1) CHAOS liver (CT) [28], which contains 20 CT scans annotated with liver; (2) AVT aorta (CT) [38], which contains 56 CT scans annotated with aorta; (3) Decathlon liver (CT) [6], which contains 131 CT scans annotated with liver; (4) Decathlon heart (MRI) [6], which contains 20 MRI scans annotated with heart. For

each dataset, we adopt an 8:2 train/test split. All volumes are resampled to a unified resolution of $128\times128\times128$.

Our Sketch2CT consists of an autoencoder and latent diffusion models for both segmentation and medical volume generation. The autoencoders use three resolution levels with channel widths $(32, 64, 64)$, a single residual block per level, and spatial attention only in the deepest layer. During diffusion training, the autoencoder is frozen and serves solely as the latent encoder-decoder. The denoising backbone is a 3D UNet provided by *MONAI* operating directly on the latent space. The UNet employs channel widths $(32, 64, 64)$ with one residual block per level, and applies spatial attention at the last two levels. Cross-attention is enabled with a conditioning dimension of 1024, and one transformer layer is used at each attention-enabled level. We train the model using a DDPM scheduler with 1000 steps and v-prediction parameterization. The optimization uses Adam with a learning rate of 1×10^{-4}, a batch size of 10, and mixed-precision training. The training runs for 300 epochs on an NVIDIA H200 GPU.

Additional experimental details, including the ablation studies of Sketch2CT, the effects of varying sketch granularity on mask generation, and other analyses, are provided in the supplementary materials.

4.1. Synthetic Image Quality Evaluation

To quantitatively assess the quality of our synthetic images, we compute two metrics: the Fréchet inception distance (FID) and the learned perceptual image patch similarity (LPIPS). Following established practices, such as those used in MedGen3D, both metrics are calculated based on multi-view projections of the generated volumes. Table 1 compares Sketch2CT with three representative baselines: Med-DDPM [13], MedGen3D [24], and Seg-Diff [29]. All baselines are evaluated on a synthesized dataset constructed from two sources: (1) sketches and text extracted from ground-truth segmentations, and (2) manually created samples—low-cost hand-drawn sketches paired with lightly refined text descriptions. For each dataset, we produce 20 such hand-crafted sketch–text pairs. Since the baselines cannot generate segmentation masks in a controlled manner, we use our synthesized masks as conditional inputs to ensure a fair comparison.

Sketch2CT achieves the best performance on most datasets. On the challenging Decathlon heart dataset, where noise leads to higher FID and lower LPIPS for all methods, Sketch2CT still surpasses all baselines by a clear margin. For the CHAOS liver and AVT aorta datasets, it exhibits consistent gains in both fidelity and perceptual similarity. Seg-Diff attains the best score on the Decathlon liver dataset, likely because its larger data volume favors a 2D diffusion model. However, its results often lack axial continuity. In contrast, Sketch2CT preserves full 3D spatial information while achieving image quality comparable to 2D methods.

	CHAOS liver (CT)		AVT aorta (CT)		Decathlon liver (CT)		Decathlon heart (MRI)	
method	FID↓	LPIPS↑	FID↓	LPIPS↑	FID↓	LPIPS↑	FID↓	LPIPS↑
Med-DDPM	114.4	0.220	119.6	0.213	115.3	0.207	128.7	0.192
MedGen3D	43.6	0.300	47.1	0.294	45.7	0.296	96.8	0.248
Seg-Diff	37.8	0.310	38.9	0.313	**34.8**	**0.335**	68.4	0.265
Sketch2CT	**33.7**	**0.332**	**36.9**	**0.321**	36.5	0.328	**65.1**	**0.269**

Table 1. Quantitative evaluation of synthetic images. We report FID (lower is better) and LPIPS (higher is better) across four datasets. Since baseline models cannot generate conditional segmentations, we provide synthesized masks as their inputs.

4.2. Benefits for Segmentation

As described in [29], we assess the quality of generated medical images by measuring how well they preserve structural information relevant for segmentation. This evaluation criterion aligns well with our Sketch2CT framework, which explicitly generates 3D segmentation masks conditioned on sketch-text pairs. Then it synthesizes corresponding medical images, ensuring that anatomical structures are preserved throughout the process.

Faithfulness to input masks. Table 2 reports the faithfulness of the generated images in relation to the input masks. For each method, we run an auxiliary segmentation network [29] (trained on real data) on the generated images and compute the Dice coefficient against (1) the input masks and (2) the segmentations predicted from the real images. Across all datasets, Sketch2CT achieves the highest Dice scores, demonstrating that the synthesized images most faithfully preserve spatial structure.

Downstream segmentation generalization. To evaluate the effectiveness of synthetic data, we train a segmentation network using either (1) real training images or (2) synthetic images generated by various methods, and test the models on real datasets, as shown in Table 3. Sketch2CT produces results closest to the real data baseline across all datasets, indicating that its generated images effectively replicate realistic anatomical details.

4.3. Qualitative Results

We also provide qualitative comparisons to complement the quantitative results and visually assess the anatomical plausibility of the synthesized segmentations and volumes. As shown in Figure 3, these examples reflect the trends observed in previous evaluations. For each case, we extract a sketch and a text description from the ground-truth segmentation to encode the organs' geometry. We then generate segmentation conditioned on these features and use it to guide the generation of medical images. We compare the synthesized results from various methods. Med-DDPM and MedGen3D often lose structural details, while Seg-Diff shows promising results for the single slice. In contrast,

model	CHAOS liver (CT)		AVT aorta (CT)		Decathlon liver (CT)		Decathlon heart (MRI)	
	Dice(m_{gen}^{pred}, m)	Dice($m_{gen}^{pred}, m_{real}^{pred}$)	Dice(m_{gen}^{pred}, m)	Dice($m_{gen}^{pred}, m_{real}^{pred}$)	Dice(m_{gen}^{pred}, m)	Dice($m_{gen}^{pred}, m_{real}^{pred}$)	Dice(m_{gen}^{pred}, m)	Dice($m_{gen}^{pred}, m_{real}^{pred}$)
Med-DDPM	0.501	0.492	0.487	0.474	0.512	0.495	0.421	0.408
MedGen3D	0.814	0.797	0.842	0.828	0.821	0.803	0.612	0.587
Seg-Diff	0.827	0.809	0.866	0.861	0.892	0.873	0.638	0.601
Sketch2CT	**0.868**	**0.852**	**0.894**	**0.887**	**0.912**	**0.904**	**0.642**	**0.614**

Table 2. Faithfulness of generated images to input masks. Dice is computed between segmentations predicted from generated images and the input masks (i.e., Dice(m_{gen}^{pred}, m)), and between segmentations predicted from generated images and segmentations predicted from real images (i.e., Dice($m_{gen}^{pred}, m_{real}^{pred}$)). Sketch2CT consistently achieves the best fidelity across all datasets.

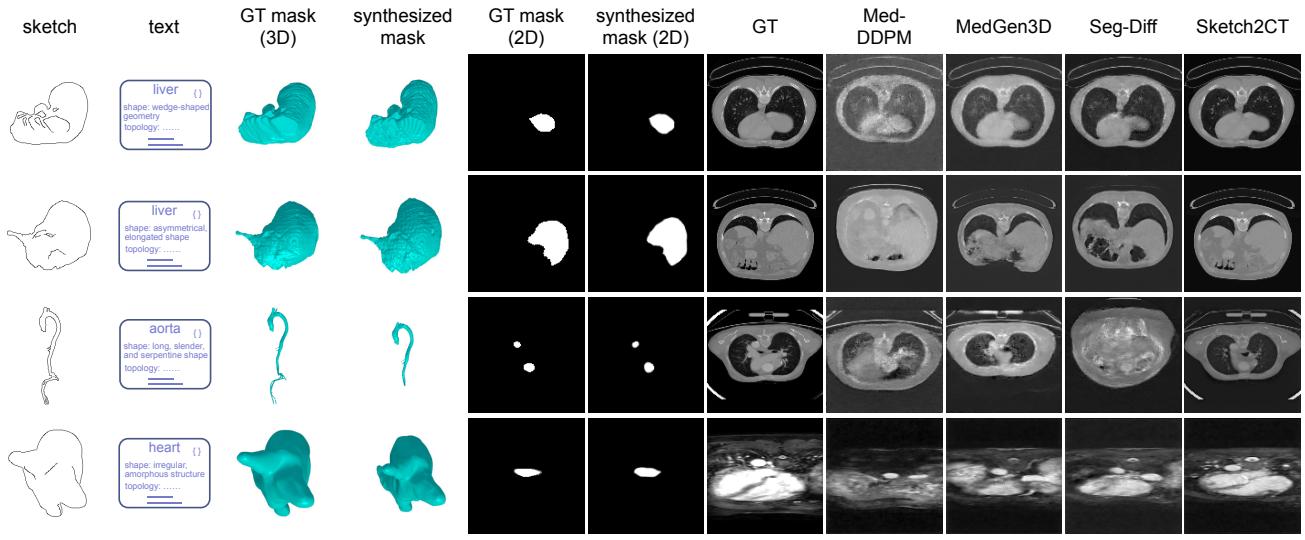


Figure 3. Qualitative comparison of baseline methods. For each case, we extract a sketch and text description from the ground-truth mask to encode organ geometry. These features guide the generation of 3D masks and the synthesis of 3D volumes.

	CHAOS liver (CT)	AVT aorta (CT)	Decathlon liver (CT)	Decathlon heart (MRI)
real training set	0.897	0.904	0.912	0.823
Med-DDPM	0.574	0.561	0.598	0.412
MedGen3D	0.814	0.801	0.823	0.681
Seg-Diff	0.826	0.812	0.887	0.702
Sketch2CT	**0.893**	**0.889**	**0.904**	**0.711**

Table 3. Downstream segmentation performance (Dice) on real datasets. A segmentation model trained on synthetic images generated by Sketch2CT achieves the closest performance to the model trained on real data, demonstrating superior anatomical realism.

Sketch2CT produces spatially coherent shapes with more precise boundaries.

5. Conclusions and Future Work

We have presented Sketch2CT, a multimodal diffusion framework that generates structure-aware 3D medical volumes from sketches and textual descriptions. The framework comprises a segmentation generator that reconstructs anatomical structures based on multimodal inputs and a volume generator that synthesizes realistic CT appearances guided by the segmentation latent. The proposed TSFE and CGFM modules effectively align sketch and text features with 3D anatomical semantics, thereby facilitating efficient, high-fidelity reconstruction. Our experiments demonstrate that Sketch2CT enhances both image realism and data availability. A key advantage of this framework is its ability to utilize low-cost, user-created sketches and simple text descriptions to produce anatomically coherent segmentation masks and corresponding CT volumes. This level of control yields reliable synthetic data that enhances data augmentation and supports downstream segmentation tasks. However, the current implementation is limited to a few organs and focuses exclusively on the synthesis of single organs. In addition, two medical imaging experts qualitatively evaluated the generated results in terms of anatomical realism, structural continuity, and clinical plausibility, confirming their overall reliability. Future work will extend Sketch2CT to multi-organ generation, incorporate disease-specific sketch editing to simulate pathological variations, and conduct broader expert evaluations.

Acknowledgments. This research was supported in part by the U.S. National Science Foundation through grants IIS-2101696, OAC-2104158, and IIS-2401144, and the U.S. National Institutes of Health through grant 7R01HL177814-02.

References

[1] Kumar Abhishek and Ghassan Hamarneh. Mask2Lesion: Mask-constrained adversarial skin lesion image synthesis. In *Proceedings of International Workshop on Simulation and Synthesis in Medical Imaging*, pages 71–80, 2019. 3

[2] Suhyun Ahn, Wonjung Park, Jihoon Cho, and Jinah Park. Volumetric conditioning module to control pretrained diffusion models for 3D medical images. In *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 85–95, 2025. 2

[3] Delin An, Pan Du, Pengfei Gu, Jian-Xun Wang, and Chaoli Wang. Hierarchical LoG Bayesian neural network for enhanced aorta segmentation. In *Proceedings of IEEE International Symposium on Biomedical Imaging*, pages 1–5, 2025. 1

[4] Delin An, Pengfei Gu, Milan Sonka, Chaoli Wang, and Danny Z Chen. Sli2Vol+: Segmenting 3D medical images based on an object estimation guided correspondence flow network. In *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3624–3634, 2025. 1

[5] Delin An, Pan Du, Jian-Xun Wang, and Chaoli Wang. AortaDiff: Volume-guided conditional diffusion models for multi-branch aortic surface generation. *IEEE Transactions on Visualization and Computer Graphics*, 32(1):922–932, 2026. 2

[6] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern H. Menze, Olaf Ronneberger, Ronald M. Summers, et al. The medical segmentation decathlon. *arXiv preprint arXiv:2106.05735*, 2021. 6

[7] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017. 3

[8] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, Vivek Natarajan, and Mohammad Norouzi. Big self-supervised models advance medical image classification. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pages 3458–3468, 2021. 1

[9] Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John V. Guttag, and Adrian V. Dalca. VoxelMorph: A learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, 2019. 2

[10] Christoph Baur, Shadi Albarqouni, and Nassir Navab. MelanoGANs: High-resolution skin lesion synthesis with GANs. *arXiv preprint arXiv:1804.04338*, 2018. 3

[11] Kevin Chen, Christopher B. Choy, Manolis Savva, Angel X. Chang, Thomas A. Funkhouser, and Silvio Savarese. Text2Shape: Generating shapes from natural language by learning joint embeddings. In *Proceedings of Asian Conference on Computer Vision*, pages 100–116, 2018. 2

[12] Ana Davila, Jacinto Colan, and Yasuhisa Hasegawa. Comparison of fine-tuning strategies for transfer learning in medical image classification. *Image and Vision Computing*, 146: 105012, 2024. 1

[13] Zolnamar Dorjsembe, Hsing-Kuo Pao, Sodtavilan Odonchimed, and Furen Xiao. Conditional diffusion models for semantic 3D brain MRI synthesis. *IEEE Journal of Biomedical and Health Informatics*, 28(7):4084–4093, 2024. 2, 7

[14] Pan Du, Delin An, Chaoli Wang, and Jian-Xun Wang. AI-powered automated model construction for patient-specific CFD simulations of aortic flows. *Science Advances*, 11(36): eadw2825, 2025. 1

[15] Nahla M. H. Elsaid and Yu-Chien Wu. Super-resolution diffusion tensor imaging using SRCNN: A feasibility study. In *Proceedings of IEEE Engineering in Medicine and Biology Society Conference*, pages 2830–2834, 2019. 2

[16] Haoqiang Fan, Hao Su, and Leonidas J. Guibas. A point set generation network for 3D object reconstruction from a single image. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2463–2471, 2017. 3

[17] Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jennings, Fiona Fennessy, Milan Sonka, et al. 3D slicer as an image computing platform for the quantitative imaging network. *Magnetic Resonance Imaging*, 30(9):1323–1341, 2012. 4

[18] Virginia Fernandez, Walter Hugo Lopez Pinaya, Pedro Borges, Petru-Daniel Tudosiu, Mark S. Graham, Tom Vercauteren, and M. Jorge Cardoso. Can segmentation models be trained with fully synthetically generated data? In *Proceedings of International Workshop on Simulation and Synthesis in Medical Imaging*, pages 79–90, 2022. 2, 3

[19] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[20] John T. Guibas, Tejpal S. Virdi, and Peter S. Li. Synthetic medical images from dual generative adversarial networks. *arXiv preprint arXiv:1709.01872*, 2017. 3

[21] Benoît Guillard, Edoardo Remelli, Pierre Yvernay, and Pascal Fua. Sketch2Mesh: Reconstructing and editing 3D shapes from sketches. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pages 13003–13012, 2021. 3

[22] Marija Habijan, Irena Galić, Hrvoje Leventić, Krešimir Romić, and Danilo Babin. Abdominal aortic aneurysm segmentation from CT images using modified 3D U-Net with deep supervision. In *Proceedings of International Symposium on Electronics in Marine*, pages 123–128, 2020. 2

[23] Changhee Han, Hideaki Hayashi, Leonardo Rundo, Ryosuke Araki, Wataru Shimoda, Shinichi Muramatsu, Yujiro Furukawa, Giancarlo Mauri, and Hideki Nakayama. GAN-based synthetic brain MR image generation. In *Proceedings of IEEE International Symposium on Biomedical Imaging*, pages 734–738, 2018. 3

[24] 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 *Proceedings of Interna-*

tional Conference on Medical Image Computing and Computer Assisted Interventions, pages 759–769, 2023. 2, 3, 7

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of Advances in Neural Information Processing Systems*, 2020. 2

[26] Hongxu Jiang, Muhammad Imran, Linhai Ma, Teng Zhang, Yuyin Zhou, Muxuan Liang, Kuang Gong, and Wei Shao. Fast-DDPM: Fast denoising diffusion probabilistic models for medical image-to-image generation. *arXiv preprint arXiv:2405.14802*, 2024. 2

[27] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. BrushNet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *Proceedings of European Conference on Computer Vision*, pages 150–168, 2024. 2

[28] A. Emre Kavur, Naciye Sinem Gezer, Mustafa Baris, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savas Özkan, et al. CHAOS challenge: Combined (CT-MR) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021. 6

[29] Nicholas Konz, Yuwen Chen, Haoyu Dong, and Maciej A. Mazurowski. Anatomically-controllable medical image generation with segmentation-guided diffusion models. In *Proceedings of International Conference on Medical Image Computing and Computer Assisted Interventions*, pages 88–98, 2024. 2, 3, 7

[30] Qing Lyu and Ge Wang. Conversion between CT and MRI images using diffusion and score-matching models. *arXiv preprint arXiv:2209.12104*, 2022. 2

[31] Andriy Myronenko, Dong Yang, Yufan He, and Daguang Xu. Aorta segmentation from 3D CT in MICCAI SEG.A. 2023 challenge. In *Proceedings of International Conference on Medical Image Computing and Computer Assisted Interventions*, pages 13–18, 2023. 2

[32] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of International Conference on Machine Learning*, pages 8162–8171, 2021. 2

[33] Luke Olsen, Faramarz F. Samavati, Mario Costa Sousa, and Joaquim A. Jorge. Sketch-based modeling: A survey. *Computers & Graphics*, 33(1):85–103, 2009. 2, 3

[34] Junyi Pan, Xiaoguang Han, Weikai Chen, Jiapeng Tang, and Kui Jia. Deep mesh reconstruction from single RGB images via topology modification networks. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pages 9963–9972, 2019. 3

[35] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 3942–3951, 2018. 5

[36] Walter H. L. Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F. Da Costa, Virginia Fernandez, Parashkev Nachev, Sébastien Ourselin, and M. Jorge Cardoso. Brain imaging generation with latent diffusion models. In *Proceedings of MICCAI Workshop on Deep Generative Models*, pages 117–126, 2022. 2

[37] Walter H. L. Pinaya, Petru-Daniel Tudosiu, Robert J. Gray, Geraint Rees, Parashkev Nachev, Sébastien Ourselin, and M. Jorge Cardoso. Unsupervised brain imaging 3D anomaly detection and segmentation with transformers. *Medical Image Analysis*, 79:102475, 2022. 2

[38] Lukas Radl, Yuan Jin, Antonio Pepe, Jianning Li, Christina Gsaxner, Fen-Hua Zhao, and Jan Egger. AVT: Multicenter aortic vessel tree CTA dataset collection with ground truth segmentation masks. *Data in Brief*, 40:107801, 2022. 6

[39] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, pages 3980–3990, 2019. 5

[40] Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic routing between capsules. In *Proceedings of Advances in Neural Information Processing Systems*, pages 3856–3866, 2017. 3

[41] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2023. 6

[42] Aditya Sanghi, Hang Chu, Joseph G. Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. CLIP-Forge: Towards zero-shot text-to-shape generation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18582–18592, 2022. 2

[43] Yuefan Shen, Changgeng Zhang, Hongbo Fu, Kun Zhou, and Youyi Zheng. DeepSketchHair: Deep sketch-based 3D hair modeling. *IEEE Transactions on Visualization and Computer Graphics*, 27(7):3250–3263, 2021. 2

[44] Hoo-Chang Shin, Neil A. Tenenholtz, Jameson K. Rogers, Christopher G. Schwarz, Matthew L. Senjem, Jeffrey L. Gunter, Katherine P. Andriole, and Mark Michalski. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *Proceedings of MICCAI Workshop on Simulation and Synthesis in Medical Imaging*, pages 1–11, 2018. 3

[45] Isaac Shiri, Giovanni Baj, Pooya Mohammadi Kazaj, Marius R Bigler, Anselm W Stark, Waldo Valenzuela, Ryota Kakizaki, Matthias Siepe, Stephan Windecker, Lorenz Räber, et al. AI-based detection and classification of anomalous aortic origin of coronary arteries using coronary CT angiography images. *Nature Communications*, 16(1):3095, 2025. 2

[46] Hanna Siebert, Christoph Großbröhmer, Lasse Hansen, and Mattias P. Heinrich. ConvexAdam: Self-configuring dual-optimization-based 3D multitask medical image registration. *IEEE Transactions on Medical Imaging*, 44(2):738–748, 2025. 2

[47] Vedant Singh, Surgan Jandial, Ayush Chopra, Siddharth Ramesh, Balaji Krishnamurthy, and Vineeth N. Balasubramanian. On conditioning the input noise for controlled image generation with diffusion models. *arXiv preprint arXiv:2205.03859*, 2022. 2

[48] Pooja Subramaniam, Tabea Kossen, Kerstin Ritter, Anja Hennemuth, Kristian Hildebrand, Adam Hilbert, Jan Sobesky, Michelle Livne, Ivana Galinovic, Ahmed A. Khalil, Jochen B. Fiebach, Dietmar Frey, and Vince I. Madai. Generating 3D TOF-MRA volumes and segmentation labels using generative adversarial networks. *Medical Image Analysis*, 78:102396, 2022. 3

[49] Bane Sullivan and Alex Kaszynski. PyVista: 3D plotting and mesh analysis through a streamlined interface for the visualization toolkit (VTK). *Journal of Open Source Software*, 4 (37):1450, 2019. 4

[50] 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. 3

[51] Theodoros P. Vagenas, Konstantinos Georgas, and George K. Matsopoulos. Deep learning-based segmentation and mesh reconstruction of the aortic vessel tree from CTA images. In *Proceedings of International Conference on Medical Image Computing and Computer Assisted Interventions*, pages 80–94, 2023. 2

[52] Dan Wang, Xinrui Cui, Xun Chen, Zhengxia Zou, Tianyang Shi, Septimiu E. Salcudean, Z. Jane Wang, and Rabab Ward. Multi-view 3D reconstruction with transformers. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pages 5702–5711, 2021. 3

[53] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2Mesh: Generating 3D mesh models from single RGB images. In *Proceedings of European Conference on Computer Vision*, pages 55–71, 2018. 3

[54] 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. 2, 3

[55] Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C. Cattin. Diffusion models for medical anomaly detection. In *Proceedings of International Conference on Medical Image Computing and Computer Assisted Interventions*, pages 35–45, 2022. 2

[56] Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, and Philippe C. Cattin. Diffusion models for implicit image segmentation ensembles. In *Proceedings of International Conference on Medical Imaging with Deep Learning*, pages 1336–1348, 2022. 1

[57] Zijie Wu, Yaonan Wang, Mingtao Feng, He Xie, and Ajmal Mian. Sketch and text guided diffusion model for colored point cloud generation. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pages 8895–8905, 2023. 3

[58] Julian Wyatt, Adam Leach, Sebastian M. Schmon, and Chris G. Willcocks. AnoDDPM: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 649–655, 2022. 2

[59] Haozhe Xie, Hongxun Yao, Shengping Zhang, Shangchen Zhou, and Wenxiu Sun. Pix2Vox++: Multi-scale context-aware 3D object reconstruction from single and multiple images. *International Journal of Computer Vision*, 128(12): 2919–2935, 2020. 3

[60] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text-to-image generation with attentional generative adversarial networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1316–1324, 2018. 3

[61] Chenyu You, Weicheng Dai, Yifei Min, Fenglin Liu, David A. Clifton, S. Kevin Zhou, Lawrence H. Staib, and James S. Duncan. Rethinking semi-supervised medical image segmentation: A variance-reduction perspective. In *Proceedings of Advances in Neural Information Processing Systems*, 2023. 3

[62] Han Zhang, Tao Xu, and Hongsheng Li. StackGAN: Text-to-photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pages 5908–5916, 2017. 3

[63] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pages 3813–3824, 2023. 2, 3

[64] Song-Hai Zhang, Yuan-Chen Guo, and Qing-Wen Gu. Sketch2Model: View-aware 3D modeling from single free-hand sketches. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6012–6021, 2021. 2, 3

[65] Justin Zhao, Timothy Wang, Wael Abid, Geoffrey Angus, Arnav Garg, Jeffery Kinnison, Alex Sherstinsky, Piero Molino, Travis Addair, and Devvret Rishi. LoRA Land: 310 fine-tuned LLMs that rival GPT-4, A technical report. *arXiv preprint arXiv:2405.00732*, 2024. 5

Sketch2CT: Multimodal Diffusion for Structure-Aware 3D Medical Volume Generation

Supplementary Material

A. Structured Semantic Text Description

To obtain geometry-focused textual descriptions for multimodal conditioning, we provide GPT-4o-mini with (1) three canonical 3D renderings of each organ (axial, sagittal, and coronal views), and (2) a carefully designed prompt instructing the model to extract purely geometric information. Examples of the text descriptions are presented in Figure A1.

Prompt design. GPT-4o-mini is directed to serve as an expert in the geometric interpretation of anatomical 3D structures. The prompt emphasizes strict geometric reasoning and forbids any clinical or physiological interpretation. In summary, the model is tasked with:

- analyzing the organ's 3D geometry using the three provided views;
- describing the volumetric shape, dominant axes, convexity, and proportions;
- characterizing surface morphology, including curvature patterns, smoothness, ridges, protrusions, and indentations;
- assessing bilateral or rotational symmetry and quantifying asymmetry;
- identifying high-level topological traits such as cavities, branching, or major geometric landmarks; and
- providing additional structural features that could assist downstream geometry-aware diffusion models.

The prompt explicitly limits the model to geometric and morphological observations, instructing it to avoid using medical terminology, diagnoses, clinical relevance, or biological functions. The aim is to produce a text description that focuses solely on the structural characteristics of the input 3D shape.

Output structure. GPT-4o responds with a structured description organized into the following conceptual components:

- an identifier for the input organ instance;
- a summary of global volumetric shape and coarse structural proportions;
- a characterization of surface-level geometric features;
- an analysis of symmetry or asymmetry patterns;
- a summary of key topological features; and
- high-level geometric guidance relevant for shape-conditioned generative modeling.

Although internally expressed in a structured format, we use only the distilled textual content for multimodal conditioning in the Sketch2CT framework.

B. Sketch Granularity Analysis

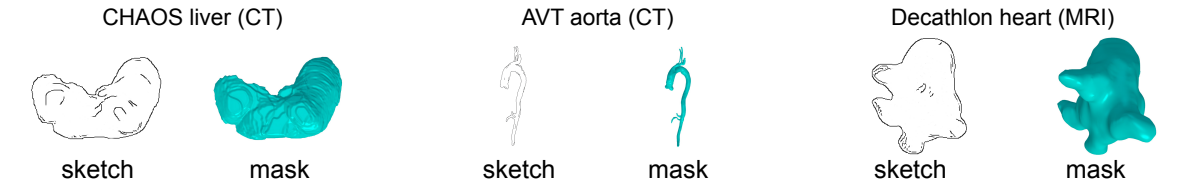
To investigate how sketch detail influences the quality of generated segmentation masks, we conduct a controlled analysis using sketches derived from 2D organ snapshots. As described in the main paper, sketch contours are generated using an edge-based extractor with a sensitivity parameter ranging from 0 to 10. Lower values produce sparse and coarse structural outlines, while higher values generate denser sketches with fine-grained details.

We select three representative parameter values across this range to create sketches with increasing levels of granularity. These sketches are then input into our segmentation generation module, while maintaining consistency with all other model components and conditions. The results are shown in Figure A2. Our results reveal a clear trend:

- **Coarse sketches** (low parameter settings) yield segmentation masks with smooth, global surfaces and minimal small-scale fluctuations. The generated structures faithfully capture the overall anatomical form, with stable topology and consistent volumetric shape.
- **Detailed sketches** (high parameter settings) introduce additional local variations, resulting in segmentation masks with slightly rougher or more irregular surface patterns. These high-detail sketches accurately capture fine contour variations and transfer them into the predicted mask.

Despite the surface-level differences, the global anatomical fidelity remains highly consistent across various levels of detail. The final reconstructed 3D medical volumes derived from these segmentation masks are visually similar, and quantitative metrics indicate minimal variation across different levels of sketch detail. In particular, the Dice score between coarse and medium sketches fluctuates within 0.02 ± 0.01. In contrast, the Dice difference between coarse and fine sketches remains similarly small at approximately 0.06 ± 0.02, confirming that increased sketch detail leads to only marginal changes in segmentation quality.

Considering both performance and computational efficiency, we recommend using sketches with low to moderate granularity. Coarse sketches offer sufficient structural guidance for the multimodal diffusion model while minimizing unnecessary local noise and reducing preprocessing overhead. This supports a practical use case where users can provide simple, clean sketches without compromising the quality of downstream generation.



global shape: The structure forms a broad, horizontally oriented mass composed of two major volumetric lobes connected by a shallower central region. Its footprint is wide and flattened, with the dominant axes extending laterally. The left portion appears thicker and more convex, while the right portion displays a more elongated and tapered contour.
surface characteristics: The surface exhibits a mix of smooth convex regions and numerous layered, terrace-like undulations that wrap across the volume.
symmetry: The shape is strongly asymmetric. The two large lobes differ in size, thickness, and curvature distribution. The right region has more pronounced layered striations, while the left region is smoother and more globular, reinforcing the asymmetric mass distribution.
topology: The topology is that of a single connected solid without holes or branching protrusions. It contains multiple shallow depressions, internal-looking concave surfaces, and a central notch-like indentation separating the two primary volumetric lobes.

global shape: An elongated, curved tubular structure with a dominant vertical axis and a pronounced C-shaped bend near the upper region. The main volume tapers gradually from a thicker proximal segment to a narrower distal end. Several smaller offshoots extend laterally from the main tube.
surface characteristics: Surface appears smooth and continuously curved with no sharp ridges. The upper curved segment shows broad, gentle convexity, while the descending portion exhibits a consistent cylindrical profile.
symmetry: The structure is strongly asymmetric. The primary curvature and the lateral branch orientations both deviate from bilateral symmetry. No axial or rotational symmetry is apparent due to the varied placement and orientation of side branches.
topology: Topologically a single connected tubular form with multiple smaller branches extending from the main path. No cavities or closed loops are visible. The topology is tree-like, with one dominant trunk and several secondary bifurcations.

global shape: The structure has an overall compact, bulbous mass with several elongated protrusions extending outward in different directions. The main body is rounded and slightly tapered toward its lower end, while three major extensions emerge asymmetrically from its mid-to-upper regions.
surface characteristics: Protrusions blend gradually into the main body with soft curvature transitions. The extended arms vary slightly in thickness and curvature, some bending gently while others remain nearly straight. The main mass exhibits gentle convexity with mild surface undulations but no prominent indentations.
symmetry: The structure is highly asymmetric. The protrusions differ in length, direction, and placement, resulting in no bilateral or rotational symmetry.
topology: It is a single connected solid with multiple outward extensions, but no holes, cavities, or branching bifurcations that split into sub-branches. The global topology is compact with several single-branch appendage-like forms.

Figure A1. Examples of textual geometry descriptions used as the text-based conditioning input in Sketch2CT. For each organ, we show the input sketch, the generated mask, and the corresponding structured text description that captures the global shape, surface characteristics, symmetry, and topology.

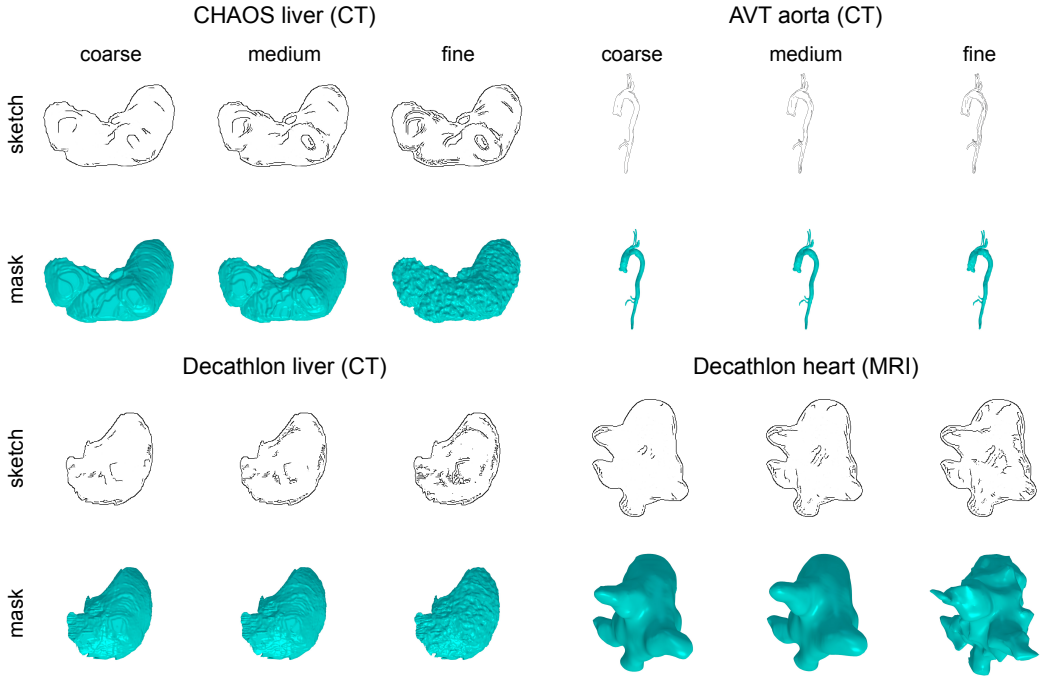


Figure A2. Effect of sketch granularity on segmentation mask generation. For each dataset, we vary the sketch detail in three levels, coarse, medium, and fine, and visualize the corresponding 3D masks produced by Sketch2CT. Increased sketch detail introduces more local structural variations, while the overall anatomical geometry remains consistent across granularity levels.

C. Ablation Study

To evaluate the contribution of the core components in Sketch2CT, we conduct an ablation study aligned with the modules defined in the main framework. Specifically, we examine the impact of removing: (1) TSFE, which refines sparse sketch embeddings via text-guided FiLM modulation; (2) CGFM, which performs global semantic alignment through hierarchical cross- and self-attention; and (3) the segmentation latent diffusion model, which reconstructs coherent 3D masks in latent space under multimodal conditioning. For each ablated variant, synthetic images are generated and used to train the same downstream segmentation network [18, 44] as in the main experiments. Table A1 reports Dice scores on real test sets. Removing any single module leads to a consistent decrease in performance across all datasets, demonstrating that all three components play complementary roles in ensuring accurate multimodal alignment and anatomically faithful 3D mask generation. The full model achieves the best performance across all benchmarks.

	CHAOS liver (CT)	AVT aorta (CT)	Decathlon liver (CT)	Decathlon heart (MRI)
full model	**0.893**	**0.889**	**0.904**	**0.711**
w/o TSFE	0.864	0.859	0.872	0.683
w/o CGFM	0.825	0.818	0.831	0.671
w/o Seg-LDM	0.642	0.629	0.633	0.545

Table A1. Ablation study of Sketch2CT. Removing TSFE, CGFM, or the segmentation latent diffusion model (Seg-LDM) reduces downstream segmentation performance, confirming that all components are essential for generating anatomically coherent and text-aligned 3D masks.

We further analyze the individual contributions of sketches and text by evaluating two additional variants: sketch-only and text-only conditioning. As shown in Figure A3, relying solely on sketches leads to incomplete or distorted 3D structures due to the inherent limitations of single-view contours, which lack depth information and volumetric context. Conversely, using text alone removes spatial constraints entirely, resulting in incorrect global shape, misplaced structures, and anatomically implausible geometries. These observations highlight that sketches and text provide complementary forms of guidance. Sketches anchor the spatial structure [21, 33, 64], while text provides semantic and morphological context, both of which are essential for accurate and stable 3D mask generation.

D. Qualitative Evaluation of Generative Diversity

To qualitatively demonstrate the diversity of the synthesized volumes generated by Sketch2CT, we repeat the generation process three times under the same sketch and text conditions. As shown in Figure A4, each run produces anatomically consistent 3D segmentation masks that follow the shared multimodal conditioning, while the synthesized CT volumes exhibit natural variability in texture, intensity distribution, and fine-scale appearance. This stochasticity reflects the inherent randomness of the diffusion process, enabling Sketch2CT to produce diverse yet structurally faithful volumetric data. The qualitative results are consistent with the quantitative findings reported in the main paper, indicating that Sketch2CT maintains stable geometry while supporting realistic appearance-level variation, which is valuable for creating diverse synthetic datasets.

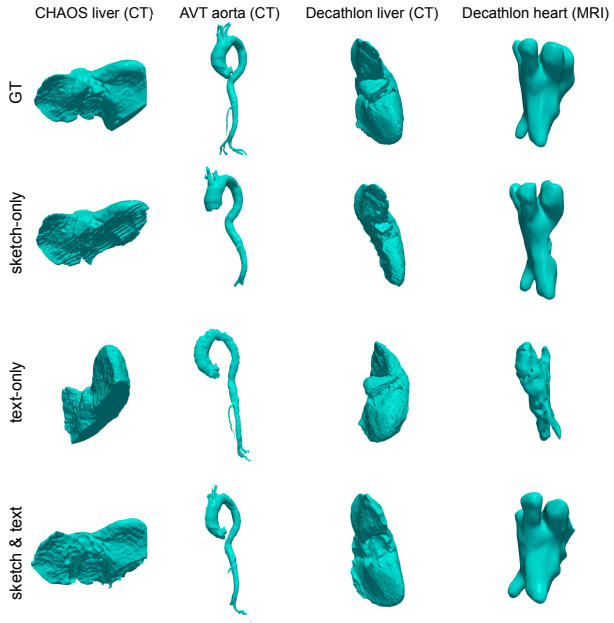


Figure A3. Comparison of segmentation masks generated using sketch-only and text-only conditions. Sketch-only guidance fails to recover full 3D geometry due to single-view ambiguity. In contrast, text-only guidance yields incorrect global shape and spatial placement, underscoring the need to combine sketches and text.

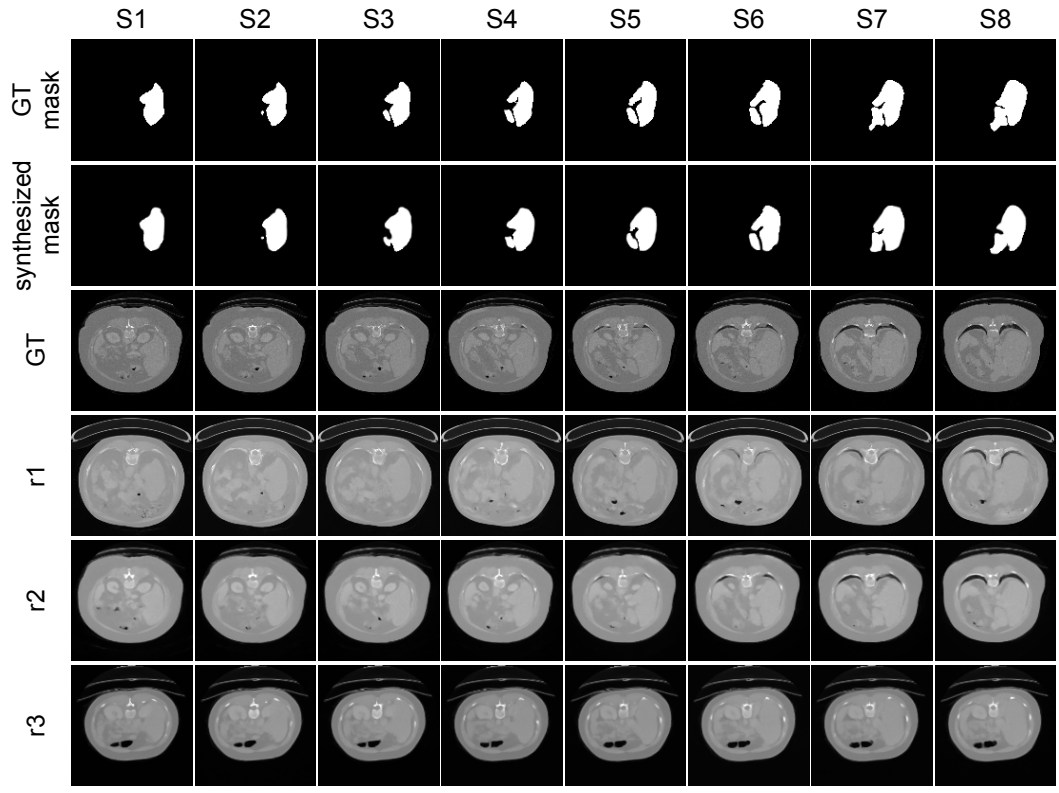


Figure A4. Qualitative diversity demonstration using a single Decathlon liver case. Under identical sketch and text conditions, three independent runs (r1-r3) generate anatomically consistent yet appearance-varying CT volumes, illustrating the stochastic diversity of Sketch2CT.