GuideGen: A Text-Guided Framework for Paired Full-torso Anatomy and CT Volume Generation

Linrui Dai[1,2*†], **Rongzhao Zhang**[3*], **Yongrui Yu**[1], **Xiaofan Zhang**[1‡]

[1]Shanghai Jiao Tong University
[2]The University of Tokyo
[3]Shanghai Artificial Intelligence Laboratory
o-o111@g.ecc.u-tokyo.ac.jp, zhangrongzhao@pjlab.org.cn, {yuyongrui, xiaofan.zhang}@sjtu.edu.cn

Abstract

The recently emerging conditional diffusion models seem promising for mitigating the labor and expenses in building large 3D medical imaging datasets. However, previous studies on 3D CT generation primarily focus on specific organs characterized by a local structure and fixed contrast and have yet to fully capitalize on the benefits of both semantic and textual conditions. In this paper, we present GuideGen, a controllable framework based on easily-acquired text prompts to generate anatomical masks and corresponding CT volumes for the entire torso—from chest to pelvis. Our approach includes three core components: a text-conditional semantic synthesizer for creating realistic full-torso anatomies; an anatomy-aware high-dynamic-range (HDR) autoencoder for high-fidelity feature extraction across varying intensity levels; and a latent feature generator that ensures alignment between CT images, anatomical semantics and input prompts. Combined, these components enable data synthesis for segmentation tasks from only textual instructions. To train and evaluate GuideGen, we compile a multi-modality cancer imaging dataset with paired CT and clinical descriptions from 12 public TCIA datasets and one private real-world dataset. Comprehensive evaluations across generation quality, cross-modality alignment, and data usability on multi-organ and tumor segmentation tasks demonstrate GuideGen's superiority over existing CT generation methods.

Code — https://github.com/OvO1111/GuideGen

1 Introduction

The acquisition of a large quantity of medical images and their corresponding labels has always been critical for modern medical image analysis. However, due to data privacy issues and laborious annotation work, such datasets are often unavailable publicly, undermining the performance of subsequent tasks (Dai, Lei, and Zhang 2023). Recently trending conditional generative methods offer a promising solution to these problems: By eliminating privacy concerns and human interventions, they can generate vivid samples to an arbitrary scale, which provide significant boost to the performance of downstream applications like image segmentation and classification (Huang et al. 2024b; Hashmi et al. 2024).

Current conditional generative approaches can be categorized based on the nature of the conditions they incorporate. Semantic conditions, like organ and tumor maps, can be harnessed to generate images that adhere to specific locality constraints (Yao et al. 2021; Yang et al. 2023; Hu et al. 2023; Han et al. 2023; Guo et al. 2024). Textual conditions, on the other hand, have the advantage of enhanced generation diversity and reduced human effort at sample synthesis thanks to the versatility of natural language (Chambon et al. 2022; Balaji et al. 2022; Cho et al. 2024; Huang et al. 2024a; Guo et al. 2025). However, current advancements that adopts conditional pipelines seldom leverage the merits from both categories. Consequently, semantic-guided models typically exhibit limited sample diversity and depend on comprehensive anatomy masks that are costly to curate. Conversely, text-guided models, though much more flexible, struggle to fully capture exact spatial relationships among anatomical structures. Realizing this disadvantage, Kim *et al.* (Kim et al. 2024) proposes to use a text-guided controllable generation pipeline that generates CT images not only conforms to a general semantics, but also to a textual guidance on image specifics. However, it requires a paired input of semantic and textual guidance as input, which greatly limits its applicability. MedSyn (Xu et al. 2024b) alleviates the need of paired input at inference by setting a null semantic input and recovering CT from a learned joint distribution. However, it remains unclear whether its chest-only pipeline can broadcast to the entire torso with more complex anatomical structures. Also, current text-guided generative pipelines largely focus on straightforward usages in classification tasks, and their potential for downstream segmentation remains underexplored, which is both critical in clinical treatment planning and accurate diagnosis (Xu et al. 2024a).

To this end, we propose GuideGen, a novel framework that synthesizes full-torso anatomies and CT images based on medical strctured inputs. Naturally, GuideGen can synthesize paired samples that provides for dataset construc-

[*]These authors contributed equally.
[†]This work was done when the author was a master student at Shanghai Jiao Tong University.
[‡]Corresponding author.

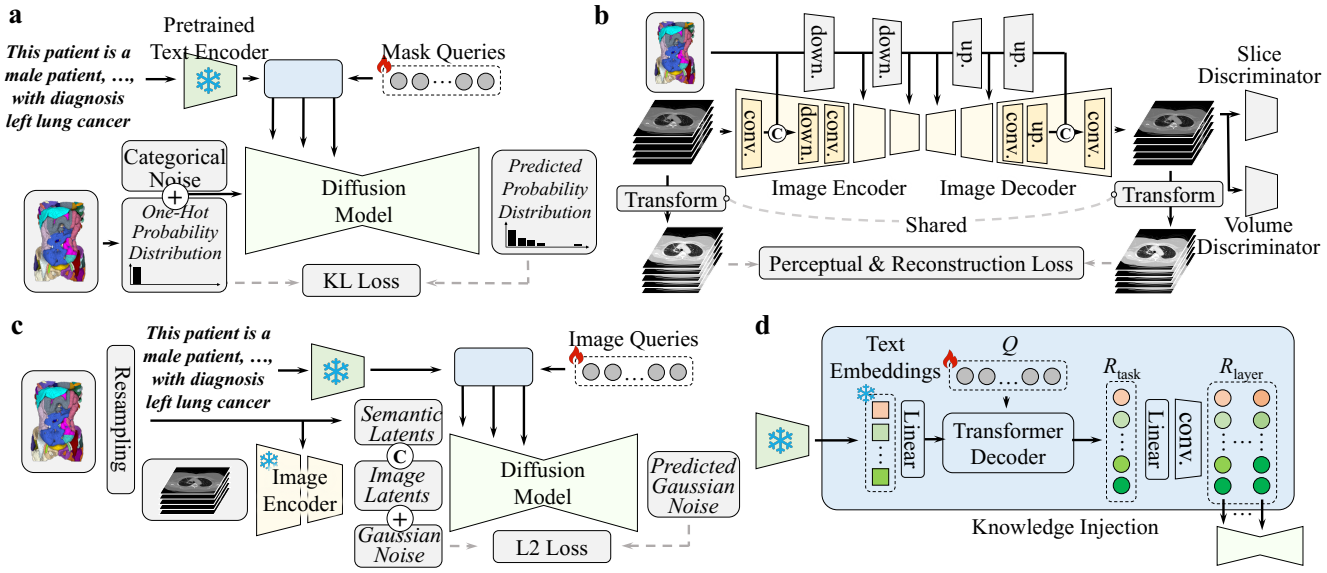

Figure 1: Overview of GuideGen's training pipeline: **(a)** Firstly, GuideGen learns to generate discrete semantic volumes that conforms to spatial features designated in the medical prompt (See Sec.3.1); **(b)** Secondly, GuideGen deploys a pyramidal autoencoding scheme to incorporate mask knowledge and reconstruct fine CT details with a high dynamic range (See Sec.3.2); **(c)** Finally, GuideGen combines the semantic latents derived from (a), image latents extracted in (b) and textual latents from the medical prompt to synthesize full-torso CT images (See Sec.3.3); **(d)** The internal structure of our knowledge injection module for extracting task-specific features from a structured input used in (a) and (c).

tion on downstream segmentation tasks. It is also more user-friendly than mask-based competitors as it necessitates only a textual input, empowering researchers to synthesize full-torso CT dataset with minor effort. GuideGen first offers a text-conditional semantic synthesizer based on a categorical diffusion model to generate discrete label indices, unlike current practices (Han et al. 2023) that suffer from an ambiguity problem (See Sec. 3.1). We also develop a knowledge injection module to help translate implicit information in medical prompts. Secondly, we train an anatomy-guided autoencoder that extracts comprehensive anatomical features from multiple contrast levels, avoiding the detail degradation by truncating intensities to a specific range that favors a region of interest (Chen et al. 2024; Yu et al. 2024) while preserving small organ and tumors described in the semantic condition. Finally, we utilize a diffusion latent generator that operates on the combined space of generated semantics and input textual features to recover valid image latents.

We train and validate GuideGen on an assembly of 12 public tumor datasets and one in-house colorectal cancer dataset, which contain the multi-modal information of CT images and medical descriptions. To testify GuideGen's downstream usability for segmentation tasks, we further include 2 multi-organ segmentation datasets and 3 tumor segmentation datasets. Our framework achieves state-of-the-art results in sample quality, conditional consistency as well as downstream usability on segmentation tasks across multiple datasets, all with the help of single textual inputs.

2 Related Works

Generative frameworks, such as Energy-Based Models (Du and Mordatch 2019; Guo et al. 2023), Generative Adversarial Networks (de Farias et al. 2021; Zhou et al. 2023), Normalizing Flows (Hajij et al. 2022; Jeevan, Nixon, and Sethi 2024), Variational Autoencoders (Kingma and Welling 2022), and Diffusion Models (Yoon et al. 2023; Hung et al. 2023; Iglesias et al. 2024) encouraged myriad researches on image generation due to their high authenticity and variability. Among these studies, the most advantageous feature that has emerged is arguably the capacity to steer the generation process via user-defined conditions (Po et al. 2024). This capability allows for a high degree of precision and customization in creating images, responding directly to the specific requirements set by the user, while providing a way for the generated images to be used for dataset augmentation purposes (Fang et al. 2024; Islam et al. 2024). In this work, we extend the ability of current diffusion models for a text-guided anatomy and CT synthesis, which mitigates data scarcity experienced in the field of medical image analysis.

Conditional Medical CT Synthesis, on which previous studies delivered convincing gains on selected downstream tasks (Hashmi et al. 2024; Kazerouni et al. 2023; Colleoni et al. 2024; Konz et al. 2024; Hu et al. 2023; Guo et al. 2024). Despite their pioneering efforts, their work mostly rely on external semantic guidance at inference, which is not costly to attain at clinical practices. On the other hand, text-based generation pipelines (Pinaya et al. 2022; Hamamci et al. 2025; Guo et al. 2025) usually supplies too little in-

formation to yield high-quality volumes and can only benefit downstream classification tasks. Although recent work has begun to combine the merits from both worlds (Kim et al. 2024; Xu et al. 2024b), they require paired inputs at inference or generate only regional patches and cannot provide for downstream segmentation tasks. While a step forward, this indicates room for further development to achieve a more automated and generalized solution. This paper intends to provide a text-guided full-torso generative framework that aids to a wider range of downstream tasks.

3 Method

As illustrated in Fig. 1, GuideGen separates the generative process into three stages to generate full-torso anatomy and CTs for downstream tasks with text-only inputs. We will elaborate on our designs in the following sections.

3.1 Text-conditional Semantic Synthesizer

Ambiguity-reducing Categorical Modeling Current generative frameworks that uses dedicated semantic synthesizers (Han et al. 2023; Chu et al. 2024) unintentionally introduce ambiguities as they struggle to capture the sharp transitions in labels near semantic boundaries due to an inaccurate data modeling. Contrary to these methods, our Text-Conditional Semantics Synthesizer (TCSS) provides a valid solution to reduce ambiguity by building upon a categorical diffusion model (Hoogeboom et al. 2021; Zbinden et al. 2023) which directly assumes a discrete formulation. As shown in Fig. 1(a), at its core, we train a categorical diffusion p_θ parameterized by θ that models the underlying distribution of mask volumes $\mathbf{m}$. At the first timestep, the diffusion variable $\mathbf{x}_0 = \mathbf{m} \in \{1, \cdots, N\}^{H\times W\times D}$, where H, W, D separately denotes mask height, width and depth. In the forward process, $\mathbf{x}_0$ is gradually transformed to a categorical noise $\mathbf{x}_T \sim \mathcal{C}_N(\mathbf{x}_T; \mathbf{1}/N)$ in T timesteps under a noise schedule $\beta_{1:T}$, where $\mathcal{C}_N$ denotes a categorical distribution of N categories (semantic classes). The forward probability transition can be described by

$$q(\mathbf{x}_t^i|\mathbf{x}_{t-1}^i) = \mathcal{C}_N(\mathbf{x}_t^i; (1-\beta_t)\mathbf{e}(\mathbf{x}_{t-1}^i) + \beta_t \cdot \frac{\mathbf{1}}{N}) \in [0,1]^N, \tag{1}$$

where the superscript i is the voxel index and $\mathbf{e}(\cdot)$ is a function that returns one-hot probability vectors from a categorical input. For compactness and ease of reading, from now on we will consider the diffusion process on an image with only one voxel to lose the superscript i, *i.e.* $H = W = D = 1$. Supposing voxels are independent, the formula derived below can be safely extended to images of any size. From the properties of Markov chains (Hoogeboom et al. 2021), the forward process of p_θ can be described as

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{C}_N(\mathbf{x}_t; \bar{\alpha}_t\mathbf{e}(\mathbf{x}_0) + (1-\bar{\alpha}_t)\cdot\frac{\mathbf{1}}{N}) \in [0,1]^N, \tag{2}$$

where $\bar{\alpha}_t = \prod_{\tau=1}^{t}(1-\beta_\tau)$.

Knowledge Injection To provide an interface for more fine-grained and versatile control via structured prompts, we extract relevant information in the medical prompt for semantic placement. Conventionally, it is widely adopted to use a dedicated text encoder $\mathcal{E}_T$ pretrained on medical knowledge (Sun et al. 2021) to map the original prompt $\mathbf{p}$ onto a latent space and perform cross-attention with the generation backbone (Chambon et al. 2022; Xu et al. 2024b; Hamamci et al. 2025). However, this alone is often insufficient to steer the model's focus to the most pertinent descriptions in input medical prompts. For instance, the model may fail to prioritize critical information like tumor location over less relevant descriptors, such as race and gender, during semantic synthesis. Recognizing this, we opt to use a separate module to extract task-specific features from encoded latents.

Specifically, as shown in Fig. 1(d), knowledge injection works by allowing learnable task-specific generation queries Q to interact with a series of transformer decoder blocks to retrieve relevant responses $R_{\text{task}} \in \mathbb{R}^{N\times C}$ for mask or image generative tasks from encoded medical prompts $\mathcal{E}_T(\mathbf{p})$. In addition, since different layers in the generative backbone perceive different levels of semantic information, we derive layer-wise responses $R_{\text{layer}} \in \mathbb{R}^{N\times(C\times L)}$ focusing on either global anatomies or local structures from R_{task}. N, C, L separately denote the number of query tokens, latent dimension and network depth of p_θ. This layer-wise guidance R_{layer} is then injected into the diffusion backbone p_θ: For each intermediate layer l in the diffusion backbone p_θ, denoting its textual guidance as R_{layer}^l, the generative latents $\mathbf{z}_l$ at layer l is computed by

$$\mathcal{Q} = W^q(\mathbf{z}_{l-1}), \mathcal{K} = W^k(R_{\text{layer}}^l), \mathcal{V} = W^v(R_{\text{layer}}^l) \tag{3}$$

$$\mathbf{z}_l = \text{MLP}(\text{LayerNorm}(\text{Softmax}(\mathcal{Q}\mathcal{K}^\top/\sqrt{d})\mathcal{V})), \tag{4}$$

where W^q, W^k, and W^v are learnable projections and d is the context dimension.

Training Objective We train $p_\theta : [0,1]^N \to [0,1]^N$ by minimizing the Kullback-Leiber divergence between posterior and generator distributions across the diffusion process as in (Ho, Jain, and Abbeel 2020). Denoting the posterior $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} \in [0,1]^N$ as $\boldsymbol{\pi}_t(\mathbf{x}_0)$, the overall training objective is

$$\mathcal{L} = \mathbb{E}_{t,\mathbf{x}_0}\left[\mathcal{D}_{\text{KL}}\left(\mathcal{C}_N(\mathbf{x}_{t-1}; \boldsymbol{\pi}_t(\mathbf{x}_0))||\mathcal{C}_N(\hat{\mathbf{x}}_{t-1}; p_\theta)\right)\right], \tag{5}$$

with the hat notation distinguishing generated variables.

To avoid training instability, we make a similar reparameterization as in DDPM (Ho, Jain, and Abbeel 2020) for enhanced performance by training a generator model on the same space as the input $\mathbf{x}_0$. Instead of training a model $p_\theta(\mathbf{x}_t, \mathbf{p})$ that models on the distribution $\mathcal{C}_N(\hat{\mathbf{x}}_{t-1}; p_\theta)$, we will train and draw samples from its reparameterized version, denoted as $f_\theta = \mathcal{C}_N(\hat{\mathbf{x}}_0; f_\theta(\mathbf{x}_t, \mathbf{p}))$. The training loss $\mathcal{L}$ needs to be modified accordingly: From the total probability formula, denoting $\mathbf{A} = (\boldsymbol{\pi}_t(1), \cdots, \boldsymbol{\pi}_t(N))$, we have $p_\theta = \mathbf{A}f_\theta$. The voxel-wise loss for TCSS can be written as

$$\mathcal{L}_1 = \mathbb{E}_{t,\mathbf{x}_0}\left[\mathcal{D}_{\text{KL}}\left(\mathcal{C}_N(\mathbf{x}_{t-1}; \boldsymbol{\pi}_t(\mathbf{x}_0))||\mathcal{C}_N(\hat{\mathbf{x}}_{t-1}; \mathbf{A}f_\theta)\right)\right]. \tag{6}$$

Under real settings where an image actually contains multiple voxels, $\mathcal{L}_1$ is averaged over each individual voxel.

At inference, TCSS samples from the distribution f_θ and computes the next denoising step according to a sampling

scheme S on the generator distribution $S(\mathbf{A}f_\theta(\hat{\mathbf{x}}_t, \mathbf{p}))$. Similar to (Zbinden et al. 2023), we sample by probability for timesteps $2 \sim T$ and take the index with max probability on the final denoising step to add stability to the generation process. Finally, generated masks can be retrieved via $\hat{\mathbf{m}} = \text{argmax}(\hat{\mathbf{x}}_0)$ along the channel dimension.

3.2 Anatomy-aware HDR Autoencoder

Anatomy Preservation Similar to previous works (Cho et al. 2024; Hamamci et al. 2025), we use a pair of image autoencoder to extract high-level texture details from CT volumes while lowering memory demands to incorporate larger CT volumes. However, current practices that directly apply off-the-shelf autoencoder architectures could undermine the preservation of spatial relations among anatomies of different sizes. For example, a small tumor visible in a higher resolution input may not be as readily recognized by the decoder half or the subsequent diffusion model in the latent space of the autoencoder. Therefore, we propose to use semantic masks to constrain the autoencoding process and mitigate this undesirable effect. As illustrated in Fig. 1(b), we follow a pyramidal approach to resample the semantic information to the resolution of layer-wise latents in the autoencoder and concatenate them in each layer. This helps the model focus on semantic details at encoding and reconstruct semantically accurate images in the decoding process.

HDR Accommodation Secondly, unlike concurrent works that truncate input CT volumes to a limited intensity range (Yu et al. 2024; Chen et al. 2024; Zhuang et al. 2023), we aim to accommodate the entire dynamic range of intensities in full-torso CTs to incorporate details for each organ. Similar to physicians examining different anatomies using different intensity windows, we devise an intensity transformation module h and constrain the reconstruction process under the full spectrum of intensity ranges. Formally,

$$h(\mathbf{x}) = k \max\left\{\min\left\{\frac{\mathbf{x} - w_c + w_r}{2w_r}, 1\right\}, 0\right\} + b, \quad (7)$$

where w_c and w_r are separately known as the window center and window radius. By definition, h truncates the intensity range of $\mathbf{x}$ to $[w_c - w_r, w_c + w_r]$ and rescales them to $[b, k + b]$. At each training step, we randomly sample w_c and w_r from the intensity range of $\mathbf{x}$ and use learnable coefficients k and b to map the truncated result back to the input space. By denoting the encoder and decoder halves of our autoencoder as $\mathcal{E}_I$ and $\mathcal{D}_I$, the l_1 reconstruction loss on an input volume $\mathbf{v}$ can be defined as

$$\mathcal{L}_{\text{rec}} = ||h(\mathcal{D}_I(\mathcal{E}_I(\mathbf{v}))) - h(\mathbf{v})||. \quad (8)$$

Aside from $\mathcal{L}_{\text{rec}}$, to reduce noise in the reconstructed images, we incorporate adversarial components into our autoencoders. Specifically, we include a frame discriminator $\mathbf{D}_f$ that randomly chooses several planar slices from the CT volume and tries to distinguish between authentic slices and synthetic ones, and a volume discriminator $\mathbf{D}_v$, that does the same job from a volumetric viewpoint. These discriminators, separately focusing on anatomical details within CT slices and the general structure of CT volumes, are not subject to intensity transformations h to avoid training instability and preserve the range of intensities for each organ. Furthermore, we use perceptual loss based on a VGG-16 network (Zhang et al. 2018; Simonyan and Zisserman 2014) $\mathcal{L}_{\text{perc}}$ for reducing artifacts as in (Chen et al. 2024). Overall, the loss function for an input CT volume $\mathbf{v}$ is

$$\begin{aligned}\mathcal{L}_2 =& \mathcal{L}_{\text{rec}}(h(\mathcal{D}_I(\mathcal{E}_I(\mathbf{v}))), h(\mathbf{v})) + \\ & \mathcal{L}_{\text{perc}}(h(\mathcal{D}_I(\mathcal{E}_I(\mathbf{v}))), h(\mathbf{v})) + \\ & \mathcal{L}_{\text{disc}}(\mathcal{D}_I(\mathcal{E}_I(\mathbf{v}))[i], \mathbf{v}[i], \mathbf{D}_f) + \\ & \mathcal{L}_{\text{disc}}(\mathcal{D}_I(\mathcal{E}_I(\mathbf{v})), \mathbf{v}, \mathbf{D}_v),\end{aligned} \quad (9)$$

where $\mathcal{L}_{\text{disc}}(\mathbf{x}', \mathbf{x}, \mathbf{D}) = \log \mathbf{D}(\mathbf{x}) + \log(1 - \mathbf{D}(\mathbf{x}'))$.

3.3 Latent-guided Feature Generator

In the final stage, we use a diffusion model to approximate the distribution of image latents $\mathbf{z}_0$ encoded by our autoencoder based on the textual and semantic guidance from input medical prompt and TCSS. In the forward process, $\mathbf{z}_0$ is gradually converted to Gaussian noise $\mathbf{z}_T \sim \mathcal{N}(0, \mathbf{I})$ through T timesteps following a noise schedule $\beta'_{1:T}$

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{1 - \beta'_t}\mathbf{z}_{t-1}, \beta'_t\mathbf{I}). \quad (10)$$

As illustrated in Fig. 1(c), We follow a similar way as in our TCSS to use knowledge injection and cross-attention layers to inject textual information into our diffusion backbone and concatenate resampled semantic to the diffusion variable $\mathbf{z}_t$ at each timestep (Fig. 1(c)). The generative objective is constructed as in (Ho, Jain, and Abbeel 2020):

$$\mathcal{L}_3 = \mathbb{E}_{t, \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})}\left[||\boldsymbol{\epsilon} - f_\varphi(\mathbf{z}_t; Resample(\hat{\mathbf{m}}), \mathbf{p})||_2^2\right], \quad (11)$$

where f_φ is the 3D U-Net (Çiçek et al. 2016) diffusion backbone parameterized by φ and controlled by $\hat{\mathbf{m}}$ after being resampled to a desired output resolution as well as textual conditions $\mathbf{p}$. At inference time, f_φ gradually recovers an image latent $\hat{\mathbf{z}}_0$ by reverse sampling from random noise, which is then decoded into a valid CT image by the decoder $\mathcal{D}_I$ described in Sec. 3.2. The generated sample pair $\{Resample(\hat{\mathbf{m}}), \mathcal{D}_I(\hat{\mathbf{z}}_0)\}$ can then be used for constructing or augmenting segmentation dataset.

4 Experiments

Dataset Construction (1) **Training**: We trained GuideGen on a compiled CT dataset from 12 public TCIA sources (Clark et al. 2013) and a private dataset (RJ), which supplements scarce public colorectal cancer data. Text prompts were generated by a private medical LLM that converted structured records (TCIA) and reports (RJ) into free-text descriptions using the template: "*The patient is* {*demographics group*}. *In this imaging, the patient's condition is described as* {*clinical information*}" For efficiency, ground-truth (GT) segmentation maps were pseudo-labeled using pre-trained networks (Wasserthal et al. 2023; Isensee et al. 2021). This dataset was split into 4534 training and 1179 validation cases. (2) **Inference**: For downstream task evaluation, we use BTCV (Landman et al. 2015) and AMOS22 (Ji et al. 2022) for multi-organ segmentation

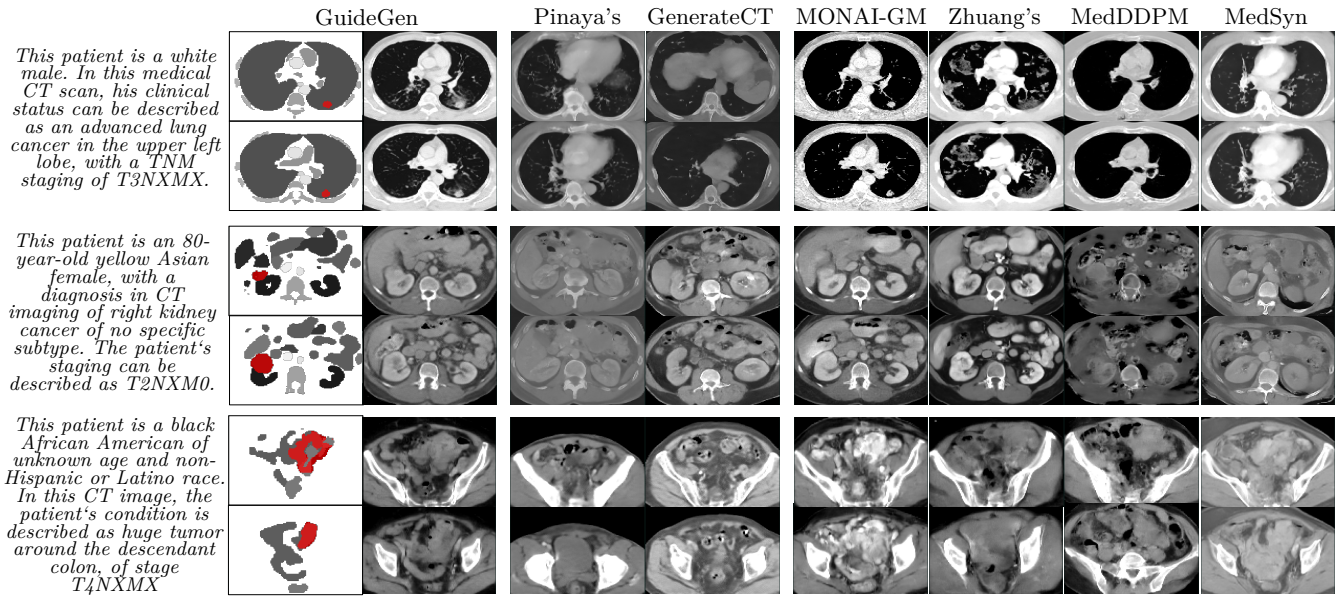

Figure 2: Qualitative results of different generation methods conditioned on the same textual prompts. Mask inputs to baseline models are generated with GuideGen with tumor semantics masked in red. See our project page for more qualitative results.

as well as lung tumors (LU), colon cancer (CO) in MSD dataset (Antonelli et al. 2022) and KiTS21 (Heller et al. 2023) for tumor segmentations. All other experiments use the validation split of our collected TCIA/RJ data.

Evaluation Metrics We use Perceptual Image Patch Similarity (LPIPS), Fréchet Inception Distance (FID, averaged on all slices from 3 axes) and Fréchet Video Inception Distance (FVD, axial slices as frames) to evaluate the generation quality (Park et al. 2023; Yu et al. 2024; Lei et al. 2025). Additionally, we report the Dice Similarity Coefficient (DSC) to evaluate image-mask consistency and the accuracy of several dedicated classifiers to measure alignment between generated CT and features described in the input prompt. Downstream segmentation performance is evaluated through DSC and 95% Hausdorff Distance (HD_{95}).

Implementation Details All generative trainings are performed under a constant learning rate of 2×10^{-5} with a batch size of 1 per GPU, a mask size of 128^3 and an image size of 256^3. Training processes use AdamW (Loshchilov and Hutter 2017) optimizers with a momentum of 0.99 and a weight decay of 1×10^{-5}. We implement each diffusion model with a cosine noise schedule and 1000 timesteps using PyTorch (Paszke et al. 2019). GuideGen is trained and tested on 4 NVIDIA 4090 GPUs with max VRAM usages of 23.2GB and 8.9GB separately for training and inference.

4.1 Main Results

Generation quality Mask quality: To validate that GuideGen's ambiguity-reducing modeling is indeed superior to current methods based on continuous formulation, we choose MedGen3D (Han et al. 2023) and LDM (Rombach et al. 2021) as baselines and infuse them with textual information via cross-attention modules. Their generation results are rounded to the nearest integer to represent semantic information. We also compare ours with Hu's deformation method for tumor generation (Hu et al. 2023). Results are shown in Tab. 1. It is evident that GuideGen performs better than baselines by a large margin on anatomy generation. This is not surprising as the ambiguity problem mentioned in Sec. 3.1 significantly hinders the performance of baselines when the number of semantic classes N is large. This problem becomes less prominent as N drops, as in the last two columns, where GuideGen maintains comparable performance to baseline methods. **Image quality**: We compare GuideGen's image generation quality with a series of generative methods, including Pinaya's and MedDDPM (Pinaya et al. 2022; Dorjsembe et al. 2024) for brain MRI generation, GenerateCT (Hamamci et al. 2025) and MedSyn (Xu et al. 2024b) for lung CT generation, Zhuang's (Zhuang et al. 2023) for abdominal CT generation and the generic medical generative framework MAISI (Guo et al. 2024). We make minor modifications to the baseline methods, retrain them on the same datasets as GuideGen and fix the output CT volume size of all methods at 128^3 to fit within our VRAM constraints. At inference, for methods that require a mask input,

Mask Synthesizers	Parameter Count(M)	Full Anatomy LPIPS↓	Full Anatomy FID↓	Tumor Only LPIPS↓	Tumor Only FID↓
MedGen3D(w. text)	48.8	0.70	201	**0.29**	<u>33.5</u>
LDM	115.2	<u>0.67</u>	<u>98.6</u>	0.30	69.1
Hu's	-	-	-	0.32	64.8
GuideGen	51.5	**0.33**	**7.1**	<u>0.29</u>	**27.9**

Table 1: Comparison of different semantic generation methods of full-torso anatomies and tumor-only binary masks. ↑ (↓) indicates higher(lower) values are better. We use **bold** and <u>underlined</u> metrics to indicate best and second bests.

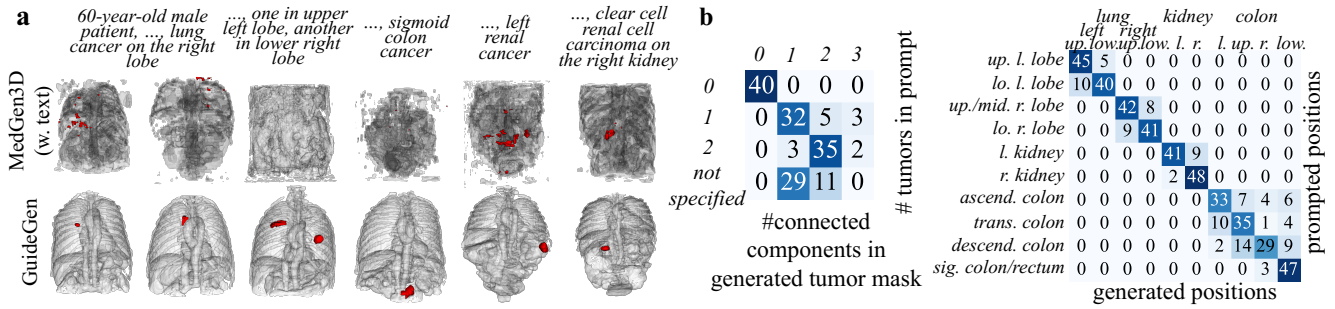

Figure 3: (a) Qualitative results of generated full-torso anatomical masks, with tumor masks masked in red. (b) Quantitative results evaluating GuideGen's mask-prompt alignment from two dimensions including the number of tumor and tumor location.

Methods	infer cond. mask	infer cond. text	Parameter Count(M)	CT generative metrics LPIPS↓	FID↓	FVD↓
Pinaya's	N	Y	196.1	0.465	92.0	2091
GenerateCT	N	Y	288.7	0.521	151.7	4634
MedSyn	N	Y	115.1	0.396	50.0	2012
MAISI	Y	N	166.5	0.406_{G}, 0.393_{R}	57.7_{G}, 54.6_{R}	1890_{G}, 1791_{R}
Zhuang's	Y	N	81.2	0.327_{G}, 0.335_{R}	$\underline{23.2}_{G}$, 39.1_{R}	$\underline{1094}_{G}$, 1366_{R}
MedDDPM	Y	N	80.5	0.354_{G}, 0.342_{R}	35.6_{G}, 45.8_{R}	1299_{G}, 1619_{R}
MedSyn	Y	Y	115.1	$\underline{0.304}_{G}$, $\underline{0.282}_{R}$	28.1_{G}, $\underline{26.7}_{R}$	1450_{G}, $\underline{1288}_{R}$
GuideGen	N	Y	164.1	$\mathbf{0.248}_{G}$, $\mathbf{0.256}_{R}$	$\mathbf{20.2}_{G}$, $\mathbf{19.4}_{R}$	$\mathbf{791}_{G}$, $\mathbf{745}_{R}$

Table 2: Comparison studies of different image generation frameworks. Metrics suffixed with 'G' or 'R' separately denote input semantic guidance from GuideGen-generated masks and real masks.

Methods	DSC↑ Spl.	Kid.	Liver	Sto.	Pan.	Lung	S.B.	Duo.	Colon	Heart	Avg.
MAISI	<u>0.73</u>	<u>0.72</u>	<u>0.80</u>	<u>0.60</u>	<u>0.43</u>	<u>0.84</u>	<u>0.49</u>	0.35	0.55	<u>0.40</u>	<u>0.59</u>
Zhuang's	0.43	0.43	0.62	0.36	0.21	0.69	0.37	0.23	0.26	0.24	0.38
MedDDPM	0.35	0.36	0.39	0.54	0.29	0.34	0.47	**0.43**	<u>0.67</u>	0.22	0.41
MedSyn	0.52	0.51	0.51	0.40	0.07	0.59	0.12	0.01	0.54	0.22	0.35
GuideGen	**0.75**	**0.72**	**0.90**	**0.63**	**0.46**	**0.84**	**0.51**	<u>0.41</u>	**0.70**	**0.53**	**0.65**

Table 3: Image-mask alignment between generated CT and their full-torso anatomical guidance. We report the DSC scores on ten major thoracic and abdominal organs.

we use both GuideGen-generated and real masks as conditions to ensure a fair comparison in Tab. 2. We can observe that GuideGen consistently outperforms existing methods in image quality across all metrics. Particularly, we notice that methods based solely on texts usually perform worse than those with semantic masks, as the highly diverse CT textures and intensity ranges among organs are difficult to recover without voxel-wise guidance, reflecting the necessity of our TCSS. Moreover, we see little discrepancy between metrics computed from CTs given by GuideGen-generated and real-world masks, reflecting that our TCSS can provide comparable quality to real masks. Quality results in Fig. 2 also show that GuideGen gives the most realistic and consistent CT generations.

Conditional alignment Image-mask alignment: Image-mask alignment can be quantitatively measured by comparing mask predictions from pretrained segmentation models (Wasserthal et al. 2023) on generated images with the actual masks given, with a higher alignment being a higher segmentation accuracy. We report the DSC of 10 major organs between predicted masks and input semantic condition in Tab. 3. It can be seen that GuideGen behaves desirably for semantic correspondence, achieving a 6% gain on DSC compared to its nearest competitor. **Image-prompt alignment**: To quantitatively evaluate the alignment between generated images and text prompt conditions, similar to (Gu et al. 2023; Hamamci et al. 2025), we use a series of pretrained classifiers on certain features to judge whether the generated images faithfully exhibit features specified in medical prompts. As shown in Tab. 4, images generated by our framework faithfully reflect the gender information and tumor location while outperforming other text-conditioned frameworks in preserving other textual features.

Methods	Accuracy↑ Age	Gender	Race	Tumor Loc.	Avg.
Pinaya's	0.06	0.35	0.10	0.17	0.17
GenerateCT	0.07	0.21	0.44	0.03	0.19
MedSyn	<u>0.17</u>	<u>0.74</u>	<u>0.51</u>	<u>0.47</u>	<u>0.47</u>
GuideGen	**0.39**	**0.90**	**0.60**	**0.89**	**0.69**

Table 4: Image-prompt alignment between generated CT and textual inputs. We report the accuracy of 4 dedicated classifiers on demographic and clinical features derived from input medical prompts.

Mask-prompt alignment: We measure mask-prompt alignment qualitatively using mask synthesizers adapted from MedGen3D (used in Tab. 1) and GuideGen in Fig. 3(a) by varying the number and location of tumor sites in our text prompt. We also quantify GuideGen's mask-prompt alignment in Fig. 3(b). It can be seen that our categorical modeling in TCSS not only generates masks of accurate shape and structure, but also preserves the spatial relationships between tumor and organs in the input prompts.

Downstream Usability To assess the usability in synthesizing datasets for segmentation, we use nnU-Net (Isensee et al. 2021) to train segmentation models with samples generated by different mask-based generative frameworks. Sample synthesis is based on the masks and/or textual prompts in the validation split of our TCIA and RJ datasets. **Multi-organ segmentation**: As shown in Tab. 5, GuideGen-generated samples yield segmentation performance comparable to real

Dataset	Method	No. Train Cases	DSC↑ Spleen	Kidneys	Liver	Sto.	Pan.	Adr.	Eso.	Aorta	IVC	Gall.	Duo.	Blad.	PV&SV	Avg.
BTCV	Real	24	0.92	0.79	0.94	0.86	0.7	0.6	0.71	0.89	0.81	0.52	-	-	0.52	0.74
	MAISI	200	0.91	0.89	0.94	0.80	0.61	0.44	0.60	0.84	0.78	0.29	-	-	0.48	0.69
	Zhuang's	200	0.90	0.88	0.95	0.83	0.65	0.54	0.69	0.88	0.82	0.49	-	-	0.56	0.74
	MedDDPM	200	0.92	0.90	0.95	0.87	0.66	0.54	0.68	0.88	0.84	0.39	-	-	0.49	0.74
	MedSyn	200	0.89	0.90	0.96	0.81	0.65	0.56	0.70	0.86	0.86	0.39	-	-	0.32	0.72
	GuideGen	200	**0.96**	**0.91**	**0.98**	**0.90**	**0.76**	**0.62**	**0.74**	**0.92**	**0.90**	**0.49**	-	-	**0.57**	**0.79**
AMOS	Real	240	0.95	0.94	0.96	0.89	0.81	0.67	0.78	0.92	0.87	0.72	0.76	0.82	-	0.84
	MAISI	200	0.83	0.84	0.91	0.74	0.60	0.50	0.60	0.81	0.71	0.34	0.53	0.43	-	0.65
	Zhuang's	200	0.85	0.87	0.90	0.66	0.63	0.46	0.61	0.82	0.73	0.26	0.55	0.62	-	0.66
	MedDDPM	200	0.86	0.89	0.90	0.74	0.61	0.46	0.61	0.86	0.76	0.43	0.55	0.60	-	0.69
	MedSyn	200	0.85	0.90	0.91	0.70	0.64	0.52	0.62	0.80	0.69	0.21	0.53	0.64	-	0.67
	GuideGen	200	**0.95**	**0.92**	**0.95**	**0.90**	**0.70**	**0.52**	**0.73**	**0.88**	**0.82**	**0.60**	**0.67**	**0.72**	-	**0.78**

Table 5: Segmentation performance on multi-organ segmentation tasks using real or synthetic data from each framework.

Method	No. Train Cases	DSC↑ LU	CO	KI	Avg.	HD_{95} ↓ LU	CO	KI	Avg.
Real	50/100/313	0.69	0.47	0.72	0.63	11.9	212.7	70.9	98.5
MAISI	200	0.48	0.10	0.24	0.27	31.7	284.9	293.5	203.4
Zhuang's	200	0.12	0.07	0.13	0.11	716.0	274.1	511.5	500.5
MedDDPM	200	0.10	0.09	0.29	0.16	776.0	273.6	119.4	389.7
MedSyn	200	0.44	0.11	0.39	0.31	33.1	267.4	309.0	203.2
GuideGen	200	**0.71**	**0.21**	**0.64**	**0.52**	**8.4**	**227.0**	**84.5**	**106.6**

Table 6: Segmentation performance on 3 tumor segmentation tasks using real or synthetic cases from each method.

Experiment Setup	LPIPS↓	FID↓	Avg. DSC↑	Avg. Acc.↑
w.o. mask input	0.42	54.3	-	0.32
LDM-generated mask input	0.34	38.4	0.34	0.40
MedGen3D-generated mask input	0.36	39.0	0.23	0.41
w.o. knowledge injection	0.26	21.7	0.25	0.57
w.o. anatomy preservation	0.27	32.4	0.40	0.61
w.o. HDR accommodation	0.33	40.9	0.36	0.64
GuideGen	**0.25**	**20.2**	**0.52**	**0.69**

Table 7: Ablations studies on GuideGen. DSC and Acc. are separately evaluated and averaged on 3 downstream tumor segmentation tasks and 4 image-prompt alignment tasks.

data and superior to samples from baseline frameworks, highlighting our method's generation quality. **Tumor segmentation**: For tumor segmentation (Tab. 6), GuideGen-generated lung tumors yield results competitive with real data, while on other tasks GuideGen-generated samples are more beneficial than those from other baselines.

4.2 Ablation Studies

Mask Synthesizers: By replacing the mask synthesizer (Stage-I) of our GuideGen with other off-the-shelf mask generators, we can ablate on the CT generative performance using different mask synthesizers, as illustrated by the first 3 rows and the last row of Tab. 7. It can be seen that CTs generated with semantic guidance from TCSS performs best, and a lack of semantic guidance degrades performance drastically. Comparing Tab. 7 with Tab. 1, we can see an interesting trend that generated CT quality is roughly in line with the quality of mask inputs. **Knowledge Injection**: The superiority of knowledge injection as opposed to conventional cross-attention guidance is illustrated in rows 4, 7 of Tab. 7, with a significant boost of alignment and downstream segmentation metrics indicating that knowledge injection better preserves textual features. **Anatomy-aware HDR Autoencoder**: The last 3 rows in Tab. 7 also shows a clear boost in generative metrics by integrating HDR accommodation and anatomy preservation to CT autoencoders for full-torso generation, which also promotes downstream usability as the image quality becomes more desirable. **Generation quantity**: From Fig. 4, we can clearly see a trend that with a larger quantity of generated samples, downstream segmentation performance can be better boosted across all tasks of choice. Also, for tasks with a simpler background (like LU), the performance of segmentation models trained using only GuideGen-generated samples can be comparable to or even better than those trained with real cases only.

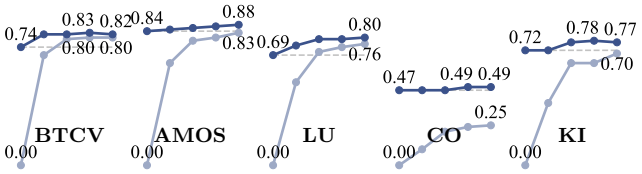

Figure 4: abcdefg1234567 Segmentation performance (DSC) using different number (0, 100, 200, 500, 1K) of GuideGen-generated samples as augmentation. Darker and lighter solid lines separately denote segmentation model trained with or without real data.

5 Conclusion

In this paper, we introduce GuideGen, a novel framework for generating paired 3D anatomies and CT images of the entire torso from structured medical prompts. Comprehensive evaluations demonstrate that GuideGen outperforms existing methods in generative quality, semantic alignment, and utility for training segmentation models. This approach advances the state-of-the-art in text-guided 3D CT generation, establishing a foundation for robust medical dataset synthesis. **Limitations** Currently, the model relies on structured prompts rather than free-text inputs; extending this capability remains a future direction for more versatile control.

Acknowledgements

This work was supported by National Natural Science Foundation of China (NSFC 62301311).

References

Antonelli, M.; Reinke, A.; Bakas, S.; Farahani, K.; Kopp-Schneider, A.; Landman, B. A.; Litjens, G.; Menze, B.; Ronneberger, O.; Summers, R. M.; et al. 2022. The medical segmentation decathlon. *Nature communications*, 13(1): 4128.

Balaji, Y.; Nah, S.; Huang, X.; Vahdat, A.; Song, J.; Kreis, K.; and Liu, M. 2022. Ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv 2022. *arXiv preprint arXiv:2211.01324*.

Chambon, P.; Bluethgen, C.; Delbrouck, J.-B.; Van der Sluijs, R.; Połacin, M.; Chaves, J. M. Z.; Abraham, T. M.; Purohit, S.; Langlotz, C. P.; and Chaudhari, A. 2022. Roentgen: vision-language foundation model for chest x-ray generation. *arXiv preprint arXiv:2211.12737*.

Chen, Q.; Chen, X.; Song, H.; Xiong, Z.; Yuille, A.; Wei, C.; and Zhou, Z. 2024. Towards generalizable tumor synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11147–11158.

Cho, J.; Zakka, C.; Kaur, D.; Shad, R.; Wightman, R.; Chaudhari, A.; and Hiesinger, W. 2024. Medisyn: Text-guided diffusion models for broad medical 2d and 3d image synthesis. *arXiv preprint arXiv:2405.09806*.

Chu, Y.; Yang, C.; Luo, G.; Qiu, Z.; and Gao, X. 2024. Anatomic-Constrained Medical Image Synthesis via Physiological Density Sampling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 69–79. Springer.

Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S. S.; Brox, T.; and Ronneberger, O. 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, 424–432. Springer.

Clark, K.; Vendt, B.; Smith, K.; Freymann, J.; Kirby, J.; Koppel, P.; Moore, S.; Phillips, S.; Maffitt, D.; Pringle, M.; et al. 2013. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Journal of digital imaging*, 26: 1045–1057.

Colleoni, E.; Matilla, R. S.; Luengo, I.; and Stoyanov, D. 2024. Guided image generation for improved surgical image segmentation. *Medical Image Analysis*, 103263.

Dai, L.; Lei, W.; and Zhang, X. 2023. Efficient Subclass Segmentation in Medical Images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 266–275. Springer.

de Farias, E. C.; Di Noia, C.; Han, C.; Sala, E.; Castelli, M.; and Rundo, L. 2021. Impact of GAN-based lesion-focused medical image super-resolution on the robustness of radiomic features. *Scientific reports*, 11(1): 21361.

Dorjsembe, Z.; Pao, H.-K.; Odonchimed, S.; and Xiao, F. 2024. Conditional Diffusion Models for Semantic 3D Brain MRI Synthesis. *IEEE Journal of Biomedical and Health Informatics*, 28(7): 4084–4093.

Du, Y.; and Mordatch, I. 2019. Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, 32.

Fang, H.; Han, B.; Zhang, S.; Zhou, S.; Hu, C.; and Ye, W.-M. 2024. Data augmentation for object detection via controllable diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1257–1266.

Gu, Y.; Yang, J.; Usuyama, N.; Li, C.; Zhang, S.; Lungren, M. P.; Gao, J.; and Poon, H. 2023. Biomedjourney: Counterfactual biomedical image generation by instruction-learning from multimodal patient journeys. *arXiv preprint arXiv:2310.10765*.

Guo, P.; Zhao, C.; Yang, D.; He, Y.; Nath, V.; Xu, Z.; Bassi, P. R.; Zhou, Z.; Simon, B. D.; Harmon, S. A.; et al. 2025. Text2CT: Towards 3D CT Volume Generation from Free-text Descriptions Using Diffusion Model. *arXiv preprint arXiv:2505.04522*.

Guo, P.; Zhao, C.; Yang, D.; Xu, Z.; Nath, V.; Tang, Y.; Simon, B.; Belue, M.; Harmon, S.; Turkbey, B.; et al. 2024. Maisi: Medical ai for synthetic imaging. *arXiv preprint arXiv:2409.11169*.

Guo, Q.; Ma, C.; Jiang, Y.; Yuan, Z.; Yu, Y.; and Luo, P. 2023. EGC: Image Generation and Classification via a Diffusion Energy-Based Model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9155–9164.

Hajij, M.; Zamzmi, G.; Paul, R.; and Thukar, L. 2022. Normalizing flow for synthetic medical images generation. In *2022 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT)*, 46–49. IEEE.

Hamamci, I. E.; Er, S.; Sekuboyina, A.; Simsar, E.; Tezcan, A.; Simsek, A. G.; Esirgun, S. N.; Almas, F.; Doğan, I.; Dasdelen, M. F.; et al. 2025. GenerateCT: text-conditional generation of 3D chest CT volumes. In *European Conference on Computer Vision*, 126–143. Springer.

Han, K.; Xiong, Y.; You, C.; Khosravi, P.; Sun, S.; Yan, X.; Duncan, J. S.; and Xie, X. 2023. Medgen3d: A deep generative framework for paired 3d image and mask generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 759–769. Springer.

Hashmi, A. U. R.; Almakky, I.; Qazi, M. A.; Sanjeev, S.; Papineni, V. R.; Mahapatra, D.; and Yaqub, M. 2024. XReal: Realistic Anatomy and Pathology-Aware X-ray Generation via Controllable Diffusion Model. *arXiv preprint arXiv:2403.09240*.

Heller, N.; Isensee, F.; Trofimova, D.; Tejpaul, R.; Zhao, Z.; Chen, H.; Wang, L.; Golts, A.; Khapun, D.; Shats, D.; Shoshan, Y.; Gilboa-Solomon, F.; George, Y.; Yang, X.; Zhang, J.; Zhang, J.; Xia, Y.; Wu, M.; Liu, Z.; Walczak, E.; McSweeney, S.; Vasdev, R.; Hornung, C.; Solaiman, R.; Schoephoerster, J.; Abernathy, B.; Wu, D.; Abdulkadir, S.; Byun, B.; Spriggs, J.; Struyk, G.; Austin, A.; Simpson, B.; Hagstrom, M.; Virnig, S.; French, J.; Venkatesh, N.; Chan, S.; Moore, K.; Jacobsen, A.; Austin, S.; Austin, M.; Regmi, S.; Papanikolopoulos, N.; and Weight, C. 2023. The KiTS21 Challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase CT. arXiv:2307.01984.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hoogeboom, E.; Nielsen, D.; Jaini, P.; Forré, P.; and Welling, M. 2021. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*, 34: 12454–12465.

Hu, Q.; Chen, Y.; Xiao, J.; Sun, S.; Chen, J.; Yuille, A. L.; and Zhou, Z. 2023. Label-free liver tumor segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7422–7432.

Huang, P.; Gao, X.; Huang, L.; Jiao, J.; Li, X.; Wang, Y.; and Guo, Y. 2024a. Chest-Diffusion: A Light-Weight Text-to-Image Model for Report-to-CXR Generation. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, 1–5. IEEE.

Huang, Y.; Zhu, J.; Hassan, H.; Su, L.; and Li, J. 2024b. Label-efficient multi-organ segmentation method with diffusion model. *arXiv preprint arXiv:2402.15216*.

Hung, A. L. Y.; Zhao, K.; Zheng, H.; Yan, R.; Raman, S. S.; Terzopoulos, D.; and Sung, K. 2023. Med-cDiff: Conditional medical image generation with diffusion models. *Bioengineering*, 10(11): 1258.

Iglesias, J. A.; Monterrubio, J. M.; Sesmero, M. P.; and Sanchis, A. 2024. Generation and Evaluation of Medical Images Based on Diffusion Models. In *2024 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, 1–8. IEEE.

Isensee, F.; Jaeger, P. F.; Kohl, S. A.; Petersen, J.; and Maier-Hein, K. H. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2): 203–211.

Islam, K.; Zaheer, M. Z.; Mahmood, A.; and Nandakumar, K. 2024. DiffuseMix: Label-Preserving Data Augmentation with Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27621–27630.

Jeevan, P.; Nixon, N.; and Sethi, A. 2024. Normalizing Flow Based Metric for Image Generation. *arXiv preprint arXiv:2410.02004*.

Ji, Y.; Bai, H.; Ge, C.; Yang, J.; Zhu, Y.; Zhang, R.; Li, Z.; Zhanng, L.; Ma, W.; Wan, X.; et al. 2022. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35: 36722–36732.

Kazerouni, A.; Aghdam, E. K.; Heidari, M.; Azad, R.; Fayyaz, M.; Hacihaliloglu, I.; and Merhof, D. 2023. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88: 102846.

Kim, K.; Na, Y.; Ye, S.-J.; Lee, J.; Ahn, S. S.; Park, J. E.; and Kim, H. 2024. Controllable text-to-image synthesis for multi-modality MR images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 7936–7945.

Kingma, D. P.; and Welling, M. 2022. Auto-Encoding Variational Bayes. arXiv:1312.6114.

Konz, N.; Chen, Y.; Dong, H.; and Mazurowski, M. A. 2024. Anatomically-controllable medical image generation with segmentation-guided diffusion models. *arXiv preprint arXiv:2402.05210*.

Landman, B.; Xu, Z.; Igelsias, J.; Styner, M.; Langerak, T.; and Klein, A. 2015. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, 12.

Lei, W.; Tian, H.; Dai, L.; Chen, H.; and Zhang, X. 2025. LesionDiffusion: Towards Text-Controlled General Lesion Synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 327–336. Springer.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Park, M.; Yun, J.; Choi, S.; and Choo, J. 2023. Learning to Generate Semantic Layouts for Higher Text-Image Correspondence in Text-to-Image Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7591–7600.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pinaya, W. H.; Tudosiu, P.-D.; Dafflon, J.; Da Costa, P. F.; Fernandez, V.; Nachev, P.; Ourselin, S.; and Cardoso, M. J. 2022. Brain imaging generation with latent diffusion models. In *MICCAI Workshop on Deep Generative Models*, 117–126. Springer.

Po, R.; Yifan, W.; Golyanik, V.; Aberman, K.; Barron, J. T.; Bermano, A.; Chan, E.; Dekel, T.; Holynski, A.; Kanazawa, A.; et al. 2024. State of the art on diffusion models for visual computing. In *Computer Graphics Forum*, volume 43, e15063. Wiley Online Library.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sun, Y.; Wang, S.; Feng, S.; Ding, S.; Pang, C.; Shang, J.; Liu, J.; Chen, X.; Zhao, Y.; Lu, Y.; et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Wasserthal, J.; Breit, H.-C.; Meyer, M. T.; Pradella, M.; Hinck, D.; Sauter, A. W.; Heye, T.; Boll, D. T.; Cyriac, J.; Yang, S.; et al. 2023. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5).

Xu, Y.; Quan, R.; Xu, W.; Huang, Y.; Chen, X.; and Liu, F. 2024a. Advances in medical image segmentation: A comprehensive review of traditional, deep learning and hybrid approaches. *Bioengineering*, 11(10): 1034.

Xu, Y.; Sun, L.; Peng, W.; Jia, S.; Morrison, K.; Perer, A.; Zandifar, A.; Visweswaran, S.; Eslami, M.; and Batmanghelich, K. 2024b. MedSyn: Text-guided Anatomy-aware Synthesis of High-Fidelity 3D CT Images. *IEEE Transactions on Medical Imaging*.

Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Zhang, W.; Cui, B.; and Yang, M.-H. 2023. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4): 1–39.

Yao, Q.; Xiao, L.; Liu, P.; and Zhou, S. K. 2021. Label-free segmentation of COVID-19 lesions in lung CT. *IEEE transactions on medical imaging*, 40(10): 2808–2819.

Yoon, J. S.; Zhang, C.; Suk, H.-I.; Guo, J.; and Li, X. 2023. Sadm: Sequence-aware diffusion model for longitudinal medical image generation. In *International Conference on Information Processing in Medical Imaging*, 388–400. Springer.

Yu, Y.; Chen, H.; Zhang, Z.; Xiao, Q.; Lei, W.; Dai, L.; Fu, Y.; Tan, H.; Wang, G.; Gao, P.; et al. 2024. CT Synthesis with Conditional Diffusion Models for Abdominal Lymph Node Segmentation. *arXiv preprint arXiv:2403.17770*.

Zbinden, L.; Doorenbos, L.; Pissas, T.; Huber, A. T.; Sznitman, R.; and Márquez-Neila, P. 2023. Stochastic segmentation with conditional categorical diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1119–1129.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

Zhou, T.; Li, Q.; Lu, H.; Cheng, Q.; and Zhang, X. 2023. GAN review: Models and medical image fusion applications. *Information Fusion*, 91: 134–148.

Zhuang, Y.; Hou, B.; Mathai, T. S.; Mukherjee, P.; Kim, B.; and Summers, R. M. 2023. Semantic image synthesis for abdominal ct. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 214–224. Springer.