Dual-Guided 3D Liver CT Image Generation for Medical Analysis

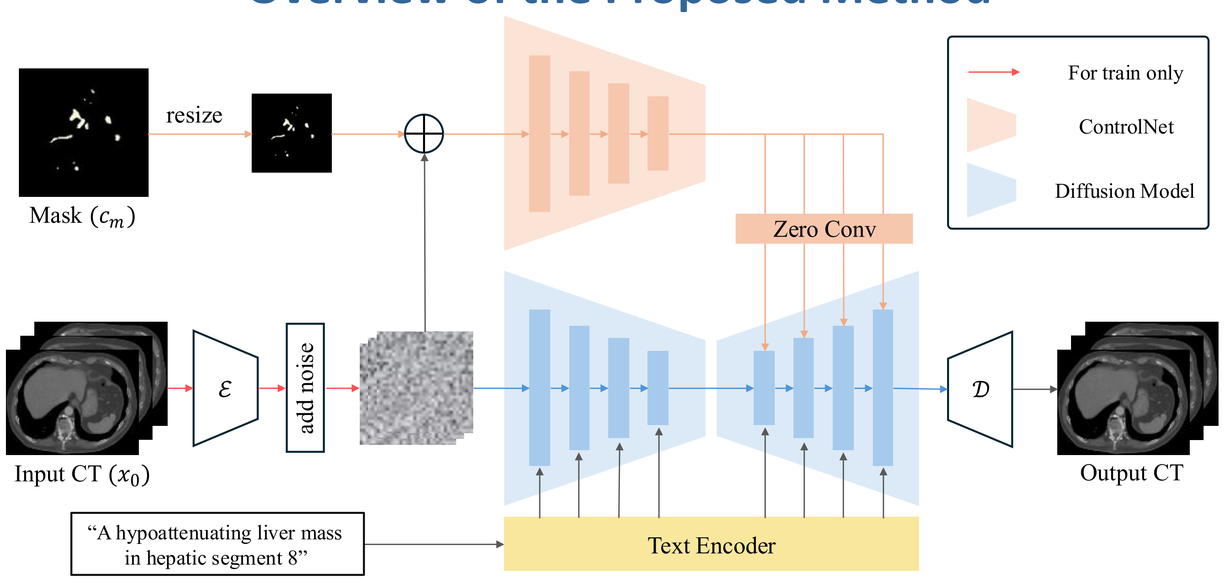
Zhizhen Song[1, 2], Jingke Zhu[1, 2], Wenyuan Huang[1], Wenjian Qin[1]
1 Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
2 University of Chinese Academy of Sciences, Beijing, China

Introduction

- The acquisition and annotation of medical imaging data face challenges, including privacy concerns, high human-annotation costs, and data imbalance.
- Current medical image generation techniques have reliability and control lability issues. The generated images often contain anatomical errors and lack precise control over essential details, resulting in untrustworthy synthesized medical images.

We propose a novel generation methodology that augments control capabilities in the generation process by integrating text prompts with anatomical masks.

Overview of the Proposed Method



The input CT image is encoded into the latent space by the encoder. Then, diffusion model running on the latent space is trained with the feature extracted by the text encoder. Finally, we create a copy of the diffusion model as the ControlNet, freeze the parameters of the original diffusion model, and train the ControlNet using liver segmentation mask. The generated feature will be decoded into the output CT.

Methodology

Autoencoder Model

- We utilize a VAE to compress high-resolution images into a lower-dimensional latent space.

$$\min_{\mathcal{E},\mathcal{D}} \max_{\mathcal{C}} \left(\mathcal{L}_{\text{recon}}\left(x, \mathcal{D}(\mathcal{E}(x))\right) + \mathcal{L}_{\text{lpips}}\left(x, \mathcal{D}(\mathcal{E}(x))\right) + \mathcal{L}_{\text{reg}}(\mathcal{E}(x)) + \mathcal{L}_{\text{adv}} \right)$$

$$\mathcal{L}_{\text{adv}} = \log\, \mathcal{C}(x) + \log\left(1 - \mathcal{C}(\mathcal{D}(\mathcal{E}(x)))\right)$$

Text-Conditioned Diffusion Model

- we employ a text encoder to extract features from radiology reports, converting natural language from these reports into high-dimensional vectors f_{text}.
- And use a 3D U-Net with cross-attention modules as the denoising model.

The training objective of our diffusion model is as follows:

$$\mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t,f_{text}}\left[\| \epsilon - \epsilon_\theta(z_t, t, f_{text}) \|_1\right]$$

Anatomical Control

- we use ControlNet to dynamically integrate anatomical masks, enabling adaptation to diverse downstream tasks.
- Given the segmentation mask c_m, The overall learning objective of the entire diffusion algorithm, which incorporates the ControlNet, is formulated as follows:

$$\mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t,f_{text},c_m}\left[\| \epsilon - \epsilon_\theta(z_t, t, f_{text}, c_m) \|_1\right]$$

Result

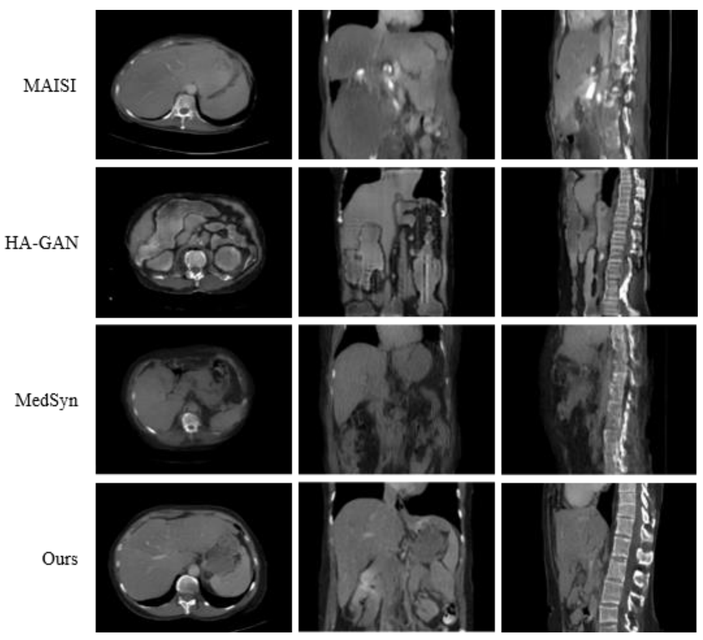
Quantitative analysis

- **Better performance on synthetic metrics**

Method	FID↓ (Axial)	FID↓ (Sagittal)	FID ↓(Coronal)	FID↓ (Avg.)
MAISI	**3.973**	6.052	9.208	6.277
HA-GAN	10.815	10.907	11.753	11.159
MedSyn	13.594	9.886	12.057	11.846
Ours	4.605	**4.865**	**5.725**	**5.064**

Qualitative Evaluation

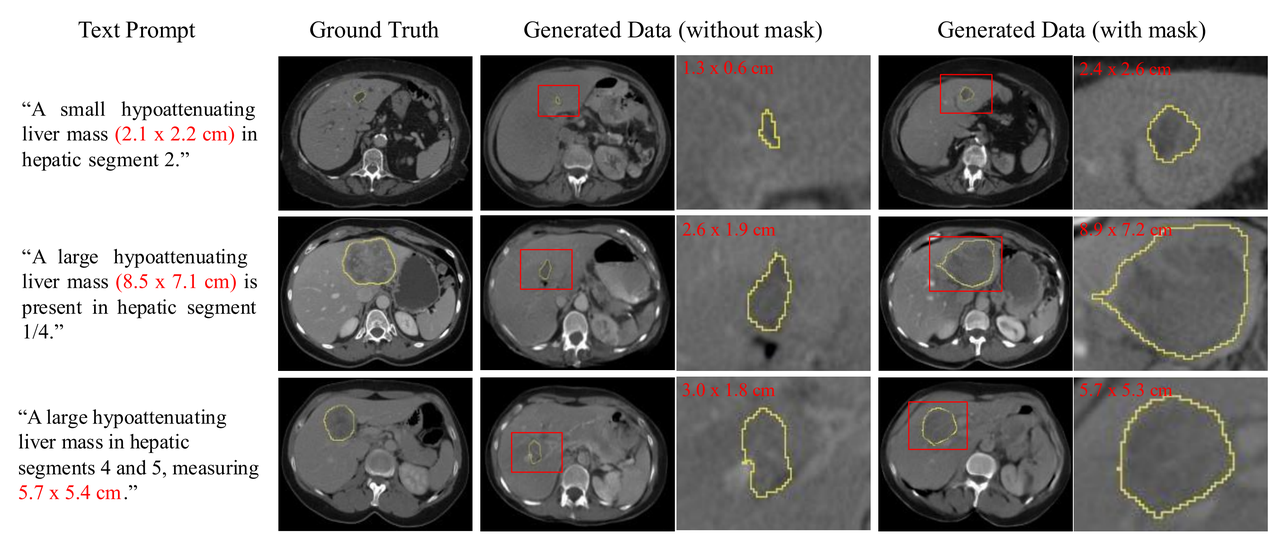
- **Superior anatomical fidelity in abdominal details**



Our method achieves superior anatomical fidelity with finer structural details compared to the base line methods. Especially in liver vessels and abdominal details

Compared to other methods, our approach can achieve higher synthesis quality.

- **Images conditionally generated**

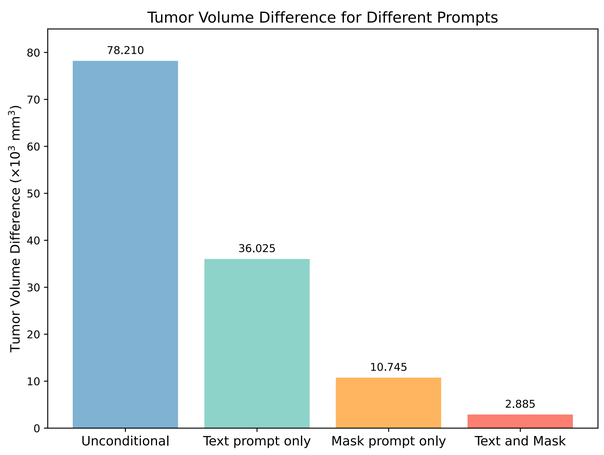


We used text prompts and masks as conditions for liver CT generation. The first column presents the text prompts. The second column shows the ground truth. The remaining columns show the images generated using different prompts. The red text in the upper left corner of each image indicates the physical size of the tumor as measured in the generated image.

Data Augmentation in Downstream Tasks

Method	DSC	NSD
Real tumors	64.7	65.9
Augmented data	**69.5**	**68.4**

Ablation



We segment the liver tumor in the synthetic image, and measure the tumor volume. We calculated the absolute value of the difference between the synthetic image and the ground truth.