Text-to-Conditioned Abdominal CT Image Generation with ControlNet

Rina Oh[1[0000-0002-7412-1249]] and Tad Gonsalves[1[0000-0001-9424-3078]]

[1] Sophia University, 7-1, Kioicho, Chiyoda-ku, Tokyo, Japan
rina_oh@sophia.ac.jp
t-gonsal@sophia.ac.jp

Abstract. Deep learning models for generative imaging have made impressive strides, yet their application to medical imaging remains limited because of data scarcity. We investigate the potential of diffusion models for generating abdominal Computed Tomography (CT) slice images conditioned on textual prompts and structural guidance. We fine-tuned ControlNet on the RAOS abdominal CT dataset, aiming to generate CT slices in three anatomical planes (axial, coronal, and sagittal) while accurately reflecting the organs specified in the prompt. During training, we optimized only the ControlNet parameters responsible for encoding the conditional inputs, keeping the pre-trained Stable Diffusion backbone (including the denoising U-Net) frozen to preserve its generative prior. We compared two forms of structural conditioning: Canny edge maps and organ-wise semantic segmentation masks. Quantitative evaluation using FID, FD-DINOv2, SSIM, LPIPS, and MedCLIP, together with qualitative visualization, showed that structural conditioning improved generation controllability and anatomical plausibility. ControlNet with Canny edge achieved the best SSIM and LPIPS across all views. Prompt-ablation experiments further showed that structural conditions alone can preserve spatial layout, but text prompts remain important for stabilizing semantic composition. These results highlight the complementary roles of textual and structural conditioning in medical image synthesis.

Keywords: Deep Learning, Image Generation, Computer Vision, ControlNet, Diffusion Models, Medical Image Processing

1 Introduction

1.1 Background

Recent advances in deep learning-based image generation have significantly improved the quality of synthesized images. In particular, diffusion models have demonstrated performance superior to earlier generative approaches, such as Generative Adversarial Networks (GANs). Systems based on Stable Diffusion [1]—a latent diffusion model comprising a text encoder for prompt conditioning and a denoising network for image synthesis—have rapidly proliferated as consumer-oriented services.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_22

However, applying such models to high-stakes domains such as medical imaging remains challenging. Developing reliable medical imaging models generally requires large, diverse, well-curated, and accurately annotated datasets, yet constructing such datasets is difficult because expert annotation is time-consuming, cost-prohibitive, and highly dependent on domain expertise [2]. In addition, medical datasets are often imperfect, containing scarce, sparse, or noisy annotations, and their collection and sharing are further constrained by patient privacy, governance, and ethical considerations [3].

These limitations motivate the development of robust medical image augmentation and synthesis methods that can complement limited real-world data and reduce dependence on exhaustive manual annotation. Rather than merely producing visually realistic images, clinically useful synthesis methods should also preserve anatomically consistent structures and plausible spatial relationships among organs [4].

Inspired by the recent success of diffusion models, this study explores their potential for generating professional-grade medical images guided by textual prompts specifying imaging modalities (e.g., Computed Tomography (CT)). Beyond visual fidelity, we prioritize anatomical consistency—specifically, the correct spatial positioning and structural relationships of organs—which are critical for medical applicability.

1.2 The Objectives of Our Study

To investigate the applicability of diffusion models to high-quality medical image generation, we employed ControlNet [5], an extension of diffusion models that enables explicit control over the spatial structure of generated images via conditional inputs (e.g., Canny edge maps and semantic segmentation masks).

In our experiments, we generated abdominal CT slices conditioned on both structural inputs and textual prompts describing the scan view and the organs present. By incorporating this conditional information, we evaluated whether diffusion-based models could produce anatomically coherent and prompt-aligned medical images.

Through qualitative visualization and quantitative evaluations using FID, FD-DINOv2, SSIM, LPIPS and MedCLIP scores, we derived the following findings:

- ControlNet with structural conditioning improved generation controllability, particularly in preserving the specified scan view and the coarse anatomical layout
- ControlNet with Canny edge provided the most balanced performance overall, achieving the best structural fidelity and perceptual similarity across views as reflected by SSIM and LPIPS
- Canny edge conditioning was more effective for preserving local boundaries and fine structural details, whereas segmentation conditioning was more useful for maintaining coarse organ placement; meanwhile, text prompts remained important for stabilizing the overall semantic composition

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_22

2 Related Studies

2.1 Image Generation via Deep Learning

Generative Adversarial Networks (GANs) [6] were long regarded as the state of the art for image synthesis. More recently, however, diffusion models have shown superior performance in image synthesis tasks [7]. Diffusion models consist of a forward process that progressively adds Gaussian noise and a reverse process that restores data through denoising [8]. Ho et al. [9] improved their practicality by introducing a simplified noise-prediction objective.

Among diffusion-based methods, ControlNet [5] enables controllable generation by encoding conditional inputs and injecting the resulting features into a pre-trained, frozen text-to-image diffusion model. This design allows fine-grained spatial control without modifying the original backbone weights. Because ControlNet can work effectively even with relatively small datasets, it is a promising approach for medical imaging, where annotated data is limited.

2.2 Medical Image Augmentation

Generating medical images remains challenging because of anatomical complexity and the scarcity of high-quality annotated datasets.

GAN-based approaches have been applied to medical image generation, such as synthetic brain CT generation from low-field MR images [10] and skin disease image synthesis [11]. However, GAN-based methods often suffer from unstable training and mode collapse [12]. To avoid these limitations, we focus on diffusion models, which generate images through iterative denoising rather than adversarial training.

Recent studies have applied diffusion models to medical image synthesis, including brain MRI generation [13], CT generation guided by cone-beam CT [14], and abdominal CT generation conditioned on semantic segmentation maps [15]. Most of these studies focused on fixed acquisition views or specific conditioning settings. In contrast, our study targets flexible abdominal CT slice generation across multiple anatomical planes by conditioning on both textual prompts and structural inputs. To support prompt-based generation while preserving spatial constraints under limited training data, we adopt ControlNet as our core method.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_22

3 Our Method

3.1 Model Architecture

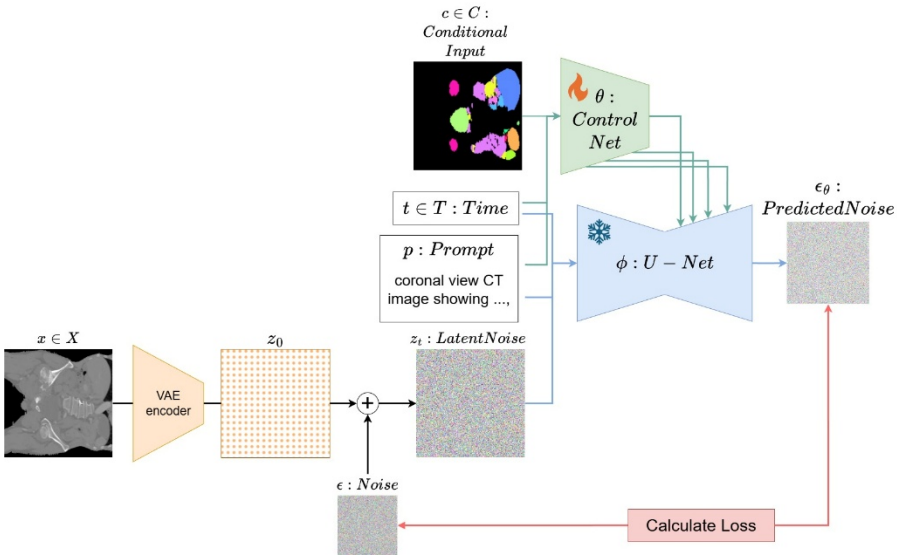

Fig. 1. Overview of training construction for ControlNet

We trained a ControlNet-based latent diffusion model to generate abdominal CT slice images conditioned on (i) a structural condition image that provides a coarse spatial layout and (ii) a text prompt describing the CT view and the set of organs present in the slice.

We implemented our model using the Diffusers library [16], which provides pre-trained latent diffusion models (e.g., Stable Diffusion) and their controllable variants (e.g., ControlNet). Following the standard latent diffusion formulation adopted in Stable Diffusion, images are encoded into a lower-dimensional latent space using a pre-trained variational autoencoder (VAE). Specifically, a target image $x \in X$ is mapped to a latent representation z_0 via the VAE encoder, and generation is performed in latent space to reduce computational cost; the final output image is obtained by decoding the predicted latent using the VAE decoder.

In accordance with the original ControlNet design [5], we used a pre-trained Stable Diffusion model as the backbone and trained a ControlNet branch to inject conditioning information while keeping the backbone weights fixed. Concretely, we used the Stable Diffusion v1.5 checkpoint and its VAE from *runwayml/stable-diffusion-v1-5*. For improved training stability and faster convergence, we initialized the ControlNet weights from publicly available pre-trained ControlNet checkpoints in Diffusers: a Canny-edge–conditioned model (*lllyasviel/control_v11p_sd15_canny*) and a segmentation-conditioned model (*lllyasviel/sd-controlnet-seg*). Then we further fine-tuned them on our abdominal CT dataset.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_22

In our framework, two types of conditional inputs were provided to the model: a structural condition image $c \in C$ and a text prompt p. The structural condition image c was derived from each real CT slice and consisted of either a Canny edge map or a semantic segmentation map, depending on the experimental setting. These condition images were used to provide explicit anatomical or boundary information to the model.

For conditional generation, the condition image c was fed into the trainable ControlNet branch, which produced condition-dependent feature maps. These feature maps were injected into the intermediate layers of the pre-trained Stable Diffusion U-Net, allowing the denoising process to be guided by the structural information while preserving the generative prior of the original diffusion model.

In parallel, the text prompt p was encoded by the pretrained Stable Diffusion text encoder from [16] to obtain the text embedding $e(p)$, which was used to condition the U-Net through the standard cross-attention mechanism. The prompts were designed to explicitly describe the imaging modality, scan orientation, and major visible abdominal organs. The model was jointly conditioned on both the structural condition image and the modality-and-anatomy-aware text prompt. In our implementation, the semantic segmentation map was first generated from the slice annotation, and the Canny edge condition was then derived from the corresponding segmentation image so that organ-wise contours could be emphasized in a controlled and anatomically consistent manner.

During training, the conditional image c was processed by the ControlNet conditioning encoder to produce multi-scale feature maps. In the Diffusers implementation, ControlNet outputs additional residual feature tensors aligned with the U-Net's internal resolutions (i.e., residuals for each down-sampling block and the middle block). These residuals are then added to the corresponding activations of the frozen denoising U-Net, enabling spatial control without modifying the original U-Net weights (Fig. 1 shows an overview of the training phase, where ControlNet receives conditional inputs and prompts.)

3.2 Training Objective

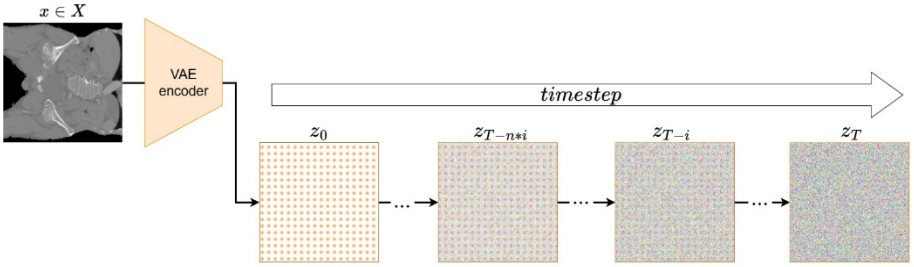

Fig. 2. Forward process in our ControlNet.

Our training follows the standard diffusion noise-prediction objective. In each training iteration, we first sampled a CT volume, selected an anatomical plane and slice position, and extracted a target slice. From the same sample, we constructed the corresponding structural condition image and text prompt. The target slice was then encoded

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_22

into latent space by the frozen VAE encoder, after which Gaussian noise was added at a randomly sampled diffusion step. The noisy latent, together with the prompt embedding and structural condition, was used to train ControlNet while keeping the Stable Diffusion backbone frozen. Given a latent z_0 encoded from a real image x, we sample a time-step t and generate latent noise z_t by Gaussian noise $\epsilon \sim N(0, I)$ with DDIM scheduler via the forward process (Fig. 2 shows an overview of the forward process):

$$z_t = \sqrt{\overline{\alpha_t}} z_0 + \sqrt{1 - \overline{\alpha_t}} \epsilon \qquad (1)$$

where $\overline{\alpha_t}$ is the cumulative product of the noise schedule gradually up to time-step t.

The denoising U-Net (with kept frozen weights) ϕ receives z_t, t, and the text embedding $e(p)$, while ControlNet θ receives $z_t, t, e(p)$, and c, and produces control residuals that are injected into the U-Net. The U-Net then predicts the added noise:

$$\epsilon_\theta = \phi(z_t, t, e(p), \theta(z_t, t, e(p), c)) \qquad (2)$$

Then we optimize the mean squared error between the predicted and actual added noise:

$$L = E_{z_0, t, c, p, \epsilon \sim N(0,I)} [\|\epsilon - \epsilon_\theta\|_2^2] \qquad (3)$$

This formulation allows the model to learn how to translate structural constraints from c and semantic constraints from the prompt p into anatomically consistent CT-like image generations, while leveraging the strong generative prior of the pre-trained Stable Diffusion backbone.

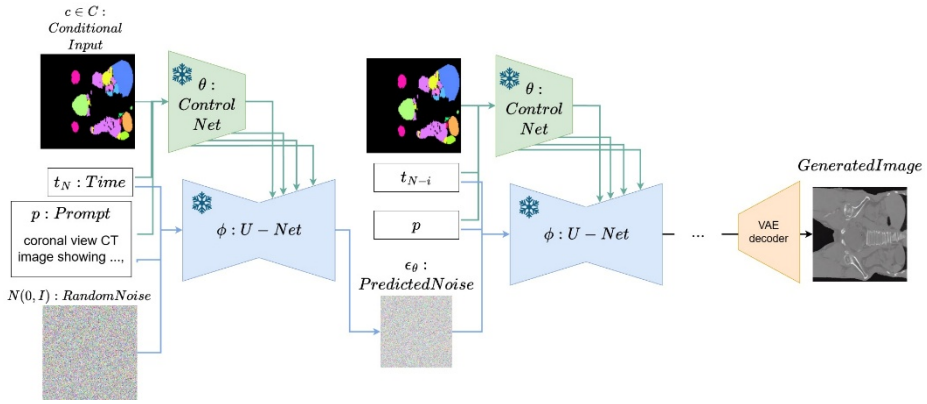

Fig. 3. Reverse process in our ControlNet

In the reverse process, conditional input and prompts are input to the trained ControlNet and multi-scale feature maps are obtained. Generation starts from a Gaussian noise latent, and at each denoising step the same structural condition image and prompt embedding are provided to ControlNet and the frozen U-Net, respectively. ControlNet produces multi-scale residual features that are injected into the corresponding U-Net

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_22

blocks to guide denoising. After the final denoising step, the resulting latent is decoded by the VAE decoder to obtain the synthesized CT slice.

4 Experimental Setup

4.1 Training Setup

Dataset. We employed the RAOS CT dataset [17][18], which consists of 413 real clinical scans annotated for 19 different organs. From this dataset, 250 scans were used for training, and the remaining scans were reserved for evaluation. For preprocessing, we extracted 512×512 slice sizes from the CT volumes after clipping Hounsfield Unit (HU) between -1000 and 1000. Semantic segmentation images were generated by coloring the masks of each organ according to the corresponding annotations. Slices were extracted in one of three randomly selected anatomical planes: axial, sagittal, or coronal. Additionally, the slice positions were randomly sampled within the central 1/4th range of each volume's dimension. As a data augmentation step, random horizontal and vertical flipping was applied. During evaluation, only the central slice of each scan was used. To construct text prompts corresponding to each slice, we referenced the annotations and slicing direction, generating prompts in the format: "*{slicing} view CT image showing {organ1}, {organ2}, {organ3}, …* ". All prompts followed the same template-based design so that the imaging modality, anatomical plane, and visible organs were described in a consistent order across samples, thereby reducing prompt ambiguity during both training and evaluation. For training ControlNet with semantic segmentation as the conditional input, we used these preprocessed semantic images so that each organ has a different coloring. When training ControlNet whose conditional input was the edge image, we applied Canny edge detection on the semantic segmentation images using the OpenCV library [19] and preprocessed Canny edge images whose colorings are different for each organ.

Training environment. Using the above preprocessing, we defined one epoch as 100 iterations, each consisting of a randomly sampled slice position and direction. Each ControlNet variant was trained for 10,000 epochs with a batch size = 4. We adopted the AdamW optimizer [20]. The learning rate was set to 2×10^{-6}.

4.2 Evaluation Metrics

Visual quality. To assess the similarity between generated images and real CT slices, we employed four evaluation metrics: Fréchet Inception Distance (FID) [21], FD-DINOv2 [22], Structural Similarity Index Measure (SSIM) [23], and Learned Perceptual Image Patch Similarity (LPIPS) [24]. FID measures the Fréchet distance between the feature distributions of generated and real images using an Inception network pre-trained on object classification. FD-DINOv2 modifies this approach by replacing the

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_22

Inception encoder with a DINOv2-ViT-L/14 encoder trained via self-supervised learning, aiming to better align with human perceptual judgments. SSIM evaluates image similarity in terms of luminance, contrast, and structural information, thereby reflecting whether local image structures are preserved. LPIPS measures perceptual similarity using deep visual features and is designed to better capture human perceptual differences than pixel-wise comparisons. For FID, FD-DINOv2, and LPIPS, lower scores indicate better similarity to real CT images, whereas for SSIM, higher scores indicate better structural fidelity.

Prompt alignment. To evaluate how well the generated images aligned with the given prompts, we used a MedCLIP-based similarity score [25] in *https://github.com/RyanWangZf/MedCLIP.git*. Unlike general-domain CLIP-based evaluation methods [26], which is trained mainly on general-domain image–text pairs from the internet, MedCLIP is a vision-language model specialized for medical images and medical texts, and is therefore better suited to assessing semantic consistency in the medical imaging domain. Specifically, we computed the cosine similarity between the image embedding and the prompt embedding obtained from MedCLIP. A higher score indicates stronger semantic alignment between the generated image and the conditioning prompt. By adopting MedCLIP, we aimed to evaluate prompt consistency using a representation space that better reflects clinically relevant semantics than a general-purpose CLIP model.

5 Results

After training Stable Diffusion and two ControlNet variants, we evaluated five generation settings, including prompt-conditioned and prompt-free variants of each ControlNet model, on 163 CT scans held out for validation. We report both qualitative comparisons (Fig. 4–5) and quantitative metrics (Table 1). Image fidelity was assessed using FID, FD-DINOv2, SSIM, and LPIPS, while prompt alignment was evaluated using MedCLIP.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_22

5.1 Qualitative Visualization

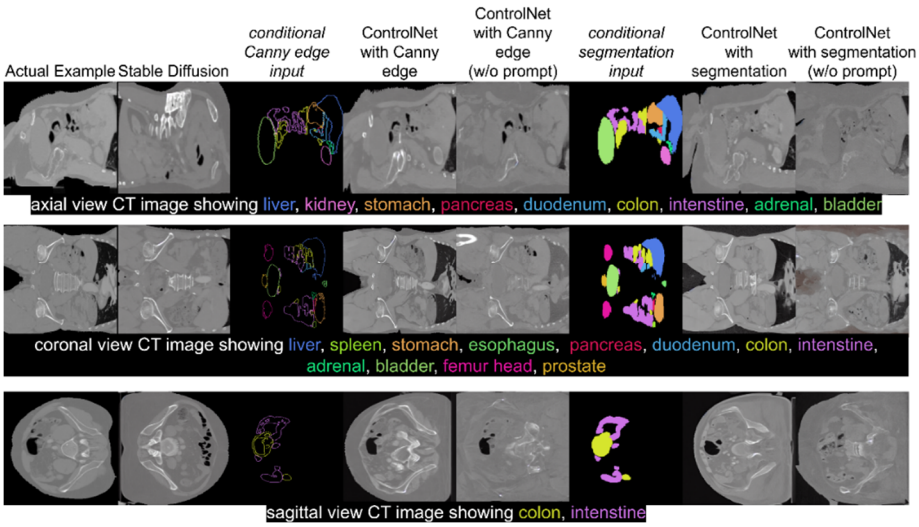

Fig. 4. Qualitative results of Stable Diffusion, ControlNet with Canny edge, ControlNet with Canny edge (w/o prompt), ControlNet with segmentation, and ControlNet with segmentation (w/o prompt). The colored text in each prompt corresponds to the colored structures in the conditional inputs. "w/o prompt" indicates that an empty prompt was used.

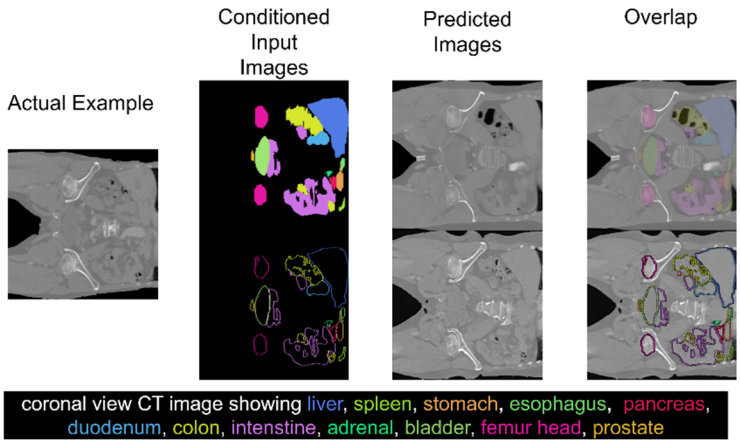

Fig. 5. Qualitative comparison of ControlNet with Canny edge and ControlNet with segmentation using overlap visualizations between the conditional inputs and the predicted images.

Fig. 4 compares representative outputs from Stable Diffusion, ControlNet with Canny edge, ControlNet with Canny edge (without prompt), ControlNet with segmentation, and ControlNet with segmentation (without prompt). Stable Diffusion reproduces the specified slicing plane reasonably well, indicating that text-only conditioning can provide a certain level of global view control in this setting. The generated images also

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_22

exhibit CT-like contrast differences among bone, air, and soft tissue. However, the boundaries and shapes of the prompted organs remain relatively blurred, and the organ-specific structures are not always clearly separated. This suggests that Stable Diffusion is driven primarily by the global attributes emphasized in the prompt, such as imaging modality and slicing plane, whereas finer anatomical details of the specified organs are not sufficiently enforced.

ControlNet with Canny edge and ControlNet with segmentation further improve structural controllability by incorporating explicit spatial conditions. When both a conditional image and a text prompt are provided, the generated images more consistently preserve the intended anatomical arrangement and produce organ layouts that are more compatible with the conditioning inputs. This indicates that the conditional image provides strong spatial guidance, while the text prompt helps constrain the global semantic content, including the intended view and the set of organs described in the prompt.

To isolate the contribution of the text prompt, we also evaluated ControlNet with Canny edge (without prompt) and ControlNet with segmentation (without prompt). In both settings, the generated images still follow the specified slicing plane reasonably well, showing that the conditional image alone already contains sufficient spatial information for coarse view control. However, these prompt-free results are noticeably blurrier and less semantically organized than those generated with text prompts. This suggests that structural conditions alone can preserve spatial layout, but textual conditioning remains important for stabilizing the overall semantic composition of the generated CT slice.

Fig. 5 further visualizes structural adherence by overlapping the conditional inputs with the predicted images. ControlNet with segmentation generally preserves the coarse positions of the organs indicated by the color-coded masks, showing that the model successfully learns the global spatial arrangement from the segmentation map. Nevertheless, some adjacent organs still appear fused or only weakly separated in the synthesized CT images, even though they are assigned different labels in the conditional input. This suggests that segmentation conditioning is used primarily to enforce organ placement rather than to guarantee sharp organ-wise boundaries.

In contrast, ControlNet with Canny edge often produces sharper local boundaries in the overlap visualization. Because the Canny input explicitly represents contour information as edges, it may provide clearer cues for local boundary reconstruction through the ControlNet encoder. This can lead to more distinct inter-organ separation in some regions than in the segmentation-conditioned case, although the Canny condition does not explicitly encode organ identity. Taken together, these qualitative results suggest complementary roles for the two types of conditioning: segmentation maps are more effective for controlling coarse anatomical layout, whereas Canny edge maps can better support local boundary preservation. In addition, text prompts remain important for stabilizing the global semantic interpretation of the generated CT slices.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_22

5.2 Quantitative Scores

Table 1. Quantitative results for axial, coronal, and sagittal views. FID, FD-DINOv2, and LPIPS are reported with ↓ values indicating better performance, whereas SSIM and MedCLIP are reported with ↑ values indicating better performance. "w/o prompt" indicates that an empty prompt was used. Bold values denote the best score for each metric. Values in parentheses for MedCLIP indicate the score computed using the real test dataset.

		Stable Diffusion	ControlNet with Canny edge	ControlNet with Canny edge (w/o prompt)	ControlNet with segmentation	ControlNet with segmentation (w/o prompt)
Axial	FID↓	109.842	**99.284**	183.864	106.229	168.518
	FD-DINOv2 ↓	**490.148**	512.143	742.904	552.679	702.569
	SSIM↑	0.475	**0.558**	0.527	0.546	0.503
	LPIPS ↓	0.624	**0.455**	0.534	0.500	0.566
	MedCLIP ↑(0.381)	0.266	0.264	**0.416**	0.381	0.392
Coronal	FID↓	**80.310**	83.535	110.789	94.437	123.402
	FD-DINOv2 ↓	**249.884**	283.174	384.218	440.590	507.791
	SSIM↑	0.594	**0.650**	0.639	0.627	0.618
	LPIPS ↓	0.529	**0.364**	0.389	0.413	0.436
	MedCLIP ↑(0.231)	0.175	0.187	0.276	0.241	**0.287**
Sagittal	FID↓	140.717	**111.881**	221.005	112.081	196.717
	FD-DINOv2 ↓	509.990	**497.855**	1052.801	532.399	1091.358
	SSIM↑	0.461	**0.491**	0.484	0.460	0.447
	LPIPS ↓	0.624	**0.490**	0.578	0.522	0.604
	MedCLIP ↑(0.167)	**0.219**	0.131	0.197	0.110	0.198

We evaluated generation quality using FID, FD-DINOv2, SSIM, and LPIPS, and assessed text-image semantic alignment using MedCLIP. FID and FD-DINOv2 measure distributional discrepancy between generated and real CT slices, whereas SSIM and LPIPS evaluate structural fidelity and perceptual similarity. Because distribution-based metrics can vary depending on the feature encoder and sample size, we interpret them

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_22

together with SSIM, LPIPS, and the qualitative results rather than relying on a single metric.

From the distribution-based evaluation, ControlNet with Canny edge achieved the lowest FID in the axial view (99.284) and sagittal view (111.881), while Stable Diffusion achieved the lowest FID in the coronal view (80.310). For FD-DINOv2, Stable Diffusion showed the lowest score in the axial (490.148) and coronal (249.884) views, whereas ControlNet with Canny edge achieved the lowest score in the sagittal view (497.855). These results indicate that the relative ranking depends on the feature encoder used for evaluation. In particular, Stable Diffusion appears to match the coarse global appearance of real CT slices reasonably well in some views, while ControlNet with Canny edge provides an advantage in other settings.

When evaluated using SSIM and LPIPS, ControlNet with Canny edge achieved the best scores in all three views (SSIM: 0.558, 0.650, and 0.491; LPIPS: 0.455, 0.364, and 0.490 for axial, coronal, and sagittal views, respectively). This consistent trend suggests that ControlNet with Canny edge is particularly effective for preserving local structures and perceptual similarity to the target abdominal CT slices. Compared with ControlNet with segmentation, the model conditioned on Canny edges produced images that were more faithful to the structural details of the reference slices, which is consistent with the qualitative observation that edge inputs provide clearer cues for local boundary reconstruction.

A clear degradation was observed when the text prompt was removed from ControlNet. In both the ControlNet with Canny edge (without prompt) and ControlNet with segmentation (without prompt) settings, the prompt-free variants showed markedly worse FID, FD-DINOv2, SSIM, and LPIPS than their corresponding prompt-conditioned counterparts across all views. This result supports the qualitative findings: although the conditional image alone provides strong spatial guidance, the text prompt remains important for stabilizing the overall semantic composition and improving anatomical realism.

MedCLIP displays a different pattern from the image-similarity metrics. The highest MedCLIP values are obtained by the Canny-conditioned model without prompt in the axial view (0.416), by the segmentation-conditioned model without prompt in the coronal view (0.287), and by Stable Diffusion in the sagittal view (0.219). This suggests that MedCLIP is more sensitive to coarse medical-semantic compatibility, such as whether an image broadly resembles an abdominal CT slice, than to fine anatomical fidelity. We therefore treat MedCLIP as a complementary semantic metric and rely on SSIM and LPIPS for assessing structural quality.

Overall, these quantitative results suggest that ControlNet with Canny edge provides the most balanced performance, especially in terms of structural fidelity and perceptual similarity, whereas Stable Diffusion can still achieve competitive scores in some distribution-based evaluations that reflect coarse global appearance. The divergence between MedCLIP and the image-similarity metrics also highlights the importance of evaluating medical image generation from multiple perspectives rather than relying on a single score.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_22

6 Conclusion

This study investigated diffusion-based conditional generation, particularly ControlNet, as a potential approach for alleviating data scarcity in medical imaging. Both qualitative and quantitative results showed that explicit structural conditioning improves the controllability of generated CT images and helps preserve anatomically plausible spatial layouts, although the most effective conditioning strategy depends on the evaluation criterion.

Among the evaluated models, ControlNet with Canny edge provided the most balanced performance overall. It achieved the best SSIM and LPIPS scores across all views and favorable FID results in the axial and sagittal views, suggesting that edge-based conditioning is effective for preserving local structures and perceptual similarity to real CT slices. In contrast, ControlNet with segmentation was useful for preserving coarse organ placement in the qualitative overlap analysis, although organ-wise boundaries were not always clearly separated. Stable Diffusion also remained competitive in some distribution-based metrics, indicating that text-only conditioning can still capture coarse global appearance, even when fine-grained organ structures remain blurred.

The prompt-ablation experiments further clarified the complementary roles of structural and textual conditioning. When the text prompt was removed, both ControlNet with Canny edge (without prompt) and ControlNet with segmentation (without prompt) still preserved the target slicing plane to some extent, showing that the conditional image alone provides strong spatial guidance. However, these prompt-free variants consistently degraded in FID, FD-DINOv2, SSIM, and LPIPS, and produced blurrier, less semantically organized images. These findings indicate that the conditional image mainly stabilizes spatial structure, whereas the text prompt remains important for constraining the overall semantic composition of the generated CT slice.

The MedCLIP results showed a different tendency from the image-similarity metrics. In some views, the highest MedCLIP scores were obtained by prompt-free ControlNet variants or by Stable Diffusion, rather than by the models that achieved the best SSIM or LPIPS. This suggests that MedCLIP reflects coarse medical-semantic compatibility, but does not necessarily capture fine-grained anatomical fidelity or organ-boundary accuracy.

These findings also reveal a practical limitation. Although conditional diffusion models can improve controllability and realism under limited data, they still require additional supervision such as structural condition images and carefully designed prompts, which may impose a substantial preparation cost in clinical settings. For more flexible cross-modality generation, methods such as SmartControl [27] may provide a promising direction. Related studies on unpaired medical image translation also offer useful insights: UNIT-DDPM [28] enables unpaired translation by exchanging intermediate denoising features across domains, and diffusion-based CT-to-MR translation methods [29] have been explored to address misalignment while preserving structural consistency. Future work should reduce dependence on extensive annotations and paired data while maintaining anatomical fidelity.

A further limitation of the present study is that we did not include anatomy-aware overlap metrics, such as Dice or IoU, nor did we evaluate whether the generated images

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_22

improve downstream tasks such as segmentation augmentation. In addition, the present study did not include evaluation by radiologists or other medical experts. These remain important directions for determining whether the proposed framework is useful not only for visual synthesis but also for practical medical-image analysis workflows.

References

1. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 10684-10695 (2022)
2. Huang, S. C., Pareek, A., Jensen, M., Lungren, M. P., Yeung, S., Chaudhari, A. S.: Self-supervised learning for medical image classification: a systematic review and implementation guidelines. NPJ Digital Medicine, 6(1), 74. (2023)
3. Price, W. N., Cohen, I. G.: Privacy in the Age of Medical Big Data. Nature medicine, 25(1), pp. 37-43 (2019).
4. Wang, T., Lei, Y., Fu, Y., Wynne, J. F., Curran, W. J., Liu, T., Yang, X.: A review on medical imaging synthesis using deep learning and its clinical applications. Journal of applied clinical medical physics, 22(1), pp. 11-36 (2021).
5. Zhang, L., Rao, A., Agrawala, M: Adding Conditional Control to Text-to-Image Diffusion Models. Proceedings of the IEEE/CVF international conference on computer vision, pp. 3836-3847 (2023)
6. Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. Advances in neural information processing systems, 27 (2014).
7. Luo, X., Li, Z., Zhang, S., Liao, W Dhariwal, P., Nichol, A.: Diffusion Models Beat GANs on Image Synthesis. Advances in neural information processing systems, 34, 8780-8794 (2021)
8. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep Unsupervised Learning using Nonequilibrium Thermodynamics. International conference on machine learning, pp. 2256-2265, pmlr (2015).
9. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. Advances in neural information processing systems, 33, 6840-6851 (2020)
10. Vellini, L., Quaranta, F., Menna, S., Pilloni, E., Catucci, F., Lenkowicz, J., Votta, C., Aquilano, M., D'Aviero, A., Iezzi, M., Preziosi, F., Re, A., Boschetti, A., Piccari, D., Piras, A., Dio, C. D., Bombini, A., Mattiucci, G. C., Cusumano, D.: A deep learning algorithm to generate synthetic computed tomography images for brain treatments from 0.35 T magnetic resonance imaging. Physics and Imaging in Radiation Oncology, 33, 100708 (2025).
11. Guo, K., Chen, J., Qiu, T., Guo, S., Luo, T., Chen, T., Ren, S.: MedGAN: An adaptive GAN approach for medical image generation. Computers in Biology and Medicine, 163, 107119 (2023).
12. Zhou, T., Li, Q., Lu, H., Cheng, Q., Zhang, X.: GAN review: Models and medical image fusion applications. Information Fusion, 91, 134-148 (2023).
13. Pinaya, W. H., Tudosiu, P. D., Dafflon, J., Da Costa, P. F., Fernandez, V., Nachev, P., Ourselin, S., Cardoso, M. J.: Brain Imaging Generation with Latent Diffusion Models. In MICCAI workshop on deep generative models, pp. 117-126, Cham: Springer Nature Switzerland (2022).

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_22

14. Peng, J., Qiu, R. L., Wynne, J. F., Chang, C. W., Pan, S., Wang, T., Roper, J., Liu, T., Patel, P. R., Yu, D. S., Yang, X: CBCT - Based synthetic CT image generation using conditional denoising diffusion probabilistic model. Medical physics, 51(3), 1847-1859 (2024).
15. Zhuang, Y., Hou, B., Mathai, T. S., Mukherjee, P., Kim, B., Summers, R. M.: Semantic Image Synthesis for Abdominal CT. International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 214-224, Cham: Springer Nature Switzerland (2023).
16. Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Nair, D., Paul, S., Berman, W., Xu, Y., Liu, S., Wolf, T.: Diffusers: State-of-the-art diffusion models, GitHub, GitHub repository, https://github.com/huggingface/diffusers, last accessed 2026/01/15
17. Wang, G.: Rethinking Abdominal Organ Segmentation (RAOS) in the clinical scenario: A robustness evaluation benchmark with challenging cases. International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 531-541. Cham: Springer Nature Switzerland (2024)
18. Luo, X., Liao, W., Xiao, J., Chen, J., Song, T., Zhang, X., Li, K., Metaxas, N. D., Wang, G., Zhang, S.: WORD: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from CT image. Medical Image Analysis, 82, 102642 (2022)
19. Bradski, G.: The OpenCV Library. Dr. Dobb's Journal: Software Tools for the Professional Programmer, 25(11), 120-123 (2000)
20. Loshchilov, I., Hutter, F: Decoupled Weight Decay Regularization, arXiv preprint arXiv:1711.05101 (2017)
21. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. Advances in neural information processing systems, 30 (2017)
22. Stein, G., Cresswell, J., Hosseinzadeh, R., Sui, Y., Ross, B., Villecroze, V., Liu, Z., Caterini, A. L., Taylor, J. E. T., Loaiza-Ganem, G.: Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. Advances in Neural Information Processing Systems, 36, 3732-3784 (2023)
23. Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P.: Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE transactions on image processing, 13(4), pp. 600-612 (2004).
24. Zhang, R., Isola, P., Efros, A. A., Shechtman, E., Wang, O.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 586-595 (2018).
25. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 3876-3887 (2022).
26. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: CLIPScore: A Reference-free Evaluation Metric for Image Captioning. Proceedings of the 2021 conference on empirical methods in natural language processing, pp. 7514-7528 (2021)
27. Liu, X., Wei, Y., Liu, M., Lin, X., Ren, P., Xie, X., Zuo, W.: SmartControl: Enhancing Controlnet for Handling Rough Visual Conditions. European Conference on Computer Vision (pp. 1-17). Cham: Springer Nature Switzerland, (2024)
28. Sasaki, H., Willcocks, C. G., & Breckon, T. P.: UNIT-DDPM: UNpaired Image Translation with Denoising Diffusion Probabilistic Models. arXiv preprint arXiv:2104.05358, (2021).
29. Jang, H., Han, N., Kwon, J., Seo, H., Park, B. J., Choi, K.: Cyclic Conditional Diffusion Models for CT-to-MR Synthetic Image Segmentation with Misaligned Image Pairs. Expert Systems with Applications, 130631, (2025)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_22