Volumetric Conditioning Module to Control Pretrained Diffusion Models for 3D Medical Images

Supplementray Material

Suhyun Ahn   Wonjung Park   Jihoon Cho   Jinah Park
KAIST
{ahn.ssu, fabiola, zinic, jinahpark}@kaist.ac.kr

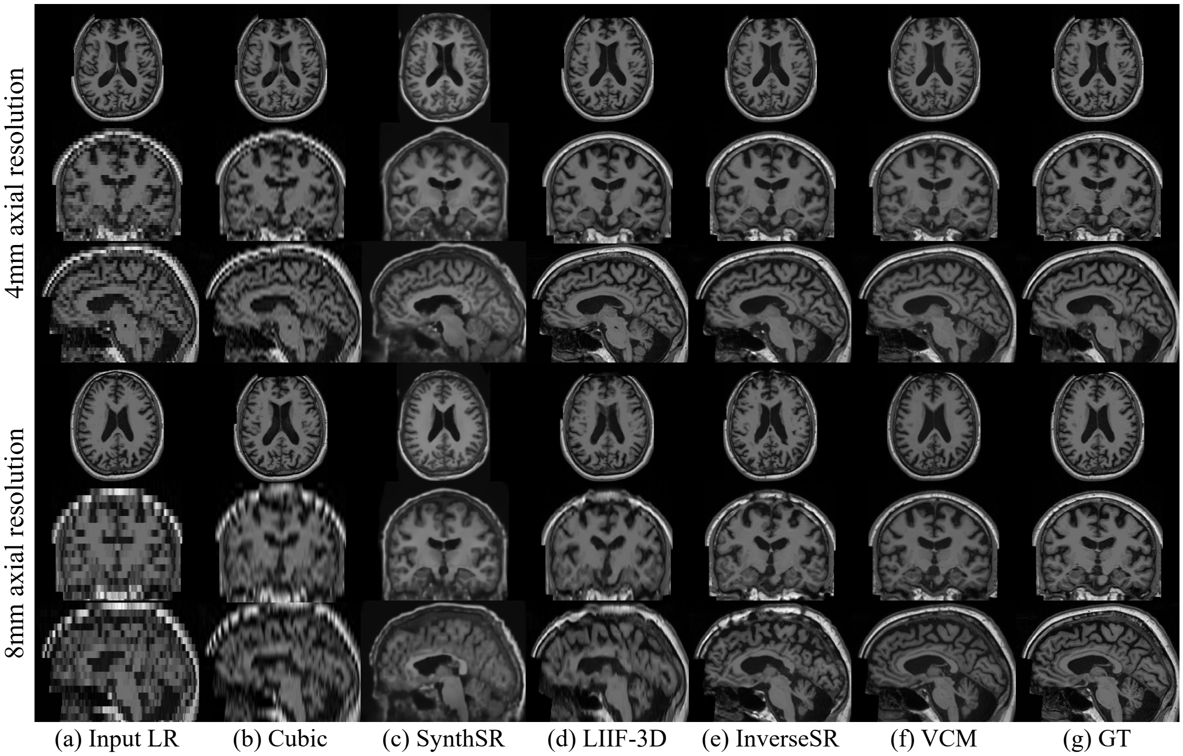

Figure 8. Qualitative results of 4 mm and 8 mm axial super-resolution from Sec. 4.3 and Fig. 7.

A. Additional experimental results

We provide the more qualitative results from Sec. 4.3 of the SR application and additional experiments and analysis with the LV masks Furthermore, as aforementioned in Secs. 1 and 2.2, we here suggest certain usage cases of spatial control methods in medical images such as image translation and precise synthesis to elaborate on the answers to the following questions: *"why are conditional generation approaches needed in the medical image domain?"* and *"what are the applications of spatial control methods for medical images?"*.

A.1. Super-resolution

For better comparability between VCM and the other methods for the super-resolution (SR) in Sec. 4.3 and Tab. 4, we visualize the more qualitative results of all methods.

In the low sparsity of 4 mm SR, as LIIF-3D [2] significantly outperforms in the metrics in Tab. 4, the model produces the most similar images with the ground truth. For example, the image appearance of LIIF-3D has not only correct semantics in the brain structures (*e.g.*, the lateral ventricles, the skull and white matter), but also high-frequency details such as brain folds (*e.g.*, the gray matter). The optimized outputs of InverseSR [17] illustrate comparable visual quality with that of VCM. Since VCM learns spatial controls for the 4 mm SR task with only 50 training samples, the results are promising in both data and computational efficiency. However, compared to LIIF-3D, InverseSR and our VCM show poor quality in the details of brain MRI scans.

In the higher sparsity of 8 mm SR, only VCM produces a normal brain MRI scan maintaining the shape of brain architectures such as the skull and the lateral ventricles, skull, and brain matters. The competitor in the 4 mm SR in both qualitative and quantitative results of Tab. 4 and Fig. 8, InverseSR [17] shows some destroyed appearances in the generated output. In addition, LIIF-3D also creates an abnormal image, failing to perform the 8 mm SR. On the other hand, SynthSR [7], a pretrained model for brain MRI SR, shows consistent results regardless of the degradation of the input image. However, the method produces the poor metric scores and the mispredicted outputs such as the skull thickness and brain structure details.

A.2. Image translation

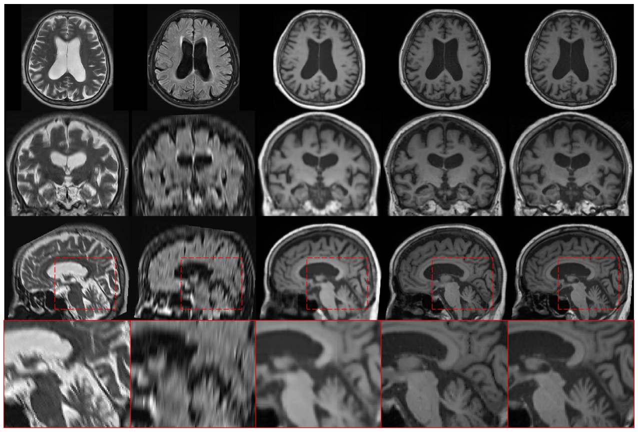

Figure 9. Comparison of generated T1w brain MRI results between Hyper-GAE [18] and our VCM for Image-to-Image translation. Used input conditions are T2 and FLAIR MRI scans.

Using the intrinsic magnetic resonance properties of tissues, we can acquire various MRI sequences such as T1, T2, and FLAIR, each offering exclusive information. A complete set of these sequences is ideal for an accurate diagnosis, but acquiring them in practice is often challenging due to the extended scanning time. Consequently, many studies have focused on predicting missing sequences from the available ones [18].

In the context of image to image translation tasks in medical images, VCM can guide the diffusion process to synthesize the missing MRI sequence using the other sequences as spatial controls. In this experiment, we used T2 and FLAIR MRI scans as conditions to generate a translated T1w MRI scan. VCM was trained with 50 training sets of T2, FLAIR, and T1w brain MRI sequences.

As illustrated in Fig. 9, VCM synthesizes a T1w MRI scan, reflecting detailed features such as folds from the T2 and FLAIR scans. Compared to HyperGAE [18], the state-of-the-art 3D-based translation method, our VCM produces clearer synthetic images leveraging the pretrained diffusion model. In addition, due to the high computational cost of 3D medical images, many 3D-based deep learning architectures employ patch-based scheme, which are prone to grid artifacts, as seen in the zoomed images of HyperGAE method (Fig. 9 (b)). In contrast, our VCM is free from grid artifacts, as it can handle the entire image even with an enterprise-level GPU (*e.g.*, 24GB VRAM).

A.3. Synthesis with precise semantics

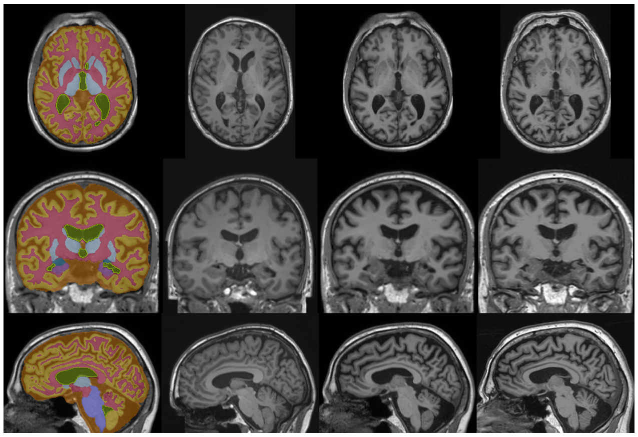

Figure 10. Samples of T1w brain MRI scans for uncontrolled output and controlled by our VCM with complex semantics.

As suggested in Sec. 1, the scarcity of available data in medical images can be addressed using high-quality synthetic data that are privacy-concern-free and precisely match the anatomical composition of real patient data. To generate these synthetic images, spatial control methods with the semantics of real patients can be a solution while also avoiding huge computational costs. To examine, we train our VCM with segmentation masks more complicated than those in Sec. 4.1. Specifically, we increase the number of semantics classes to include: 1) cerebrospinal fluid, 2) white matter, 3) gray matter, 4) all ventricles, 5) thalamus, caudate, putamen, and accumbens, 6) amygdala, 7) hippocampus, and 8) brain stem.

We visualize the samples generated by (a) BrainLDM [13] as uncontrolled and (b) controlled by our VCM with the semantics in Figs. 10 and 11. With the given input conditions, our VCM controls and generates the outputs of

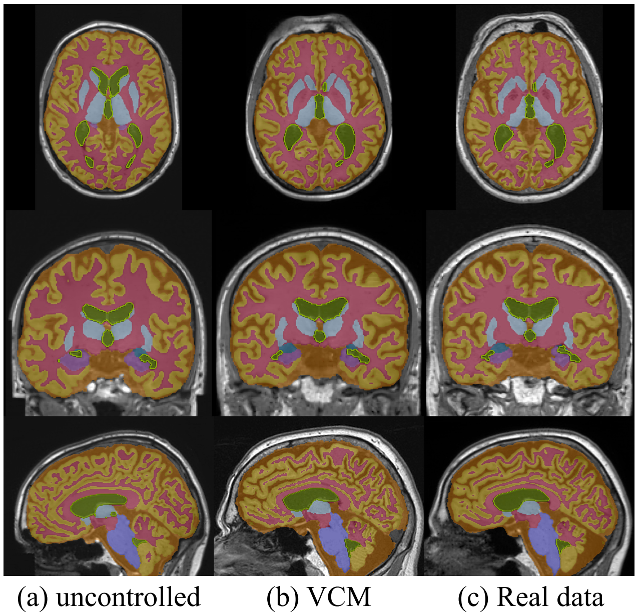

Figure 11. Comparison of the semantics of generated images between uncontrolled output and controlled by our VCM.

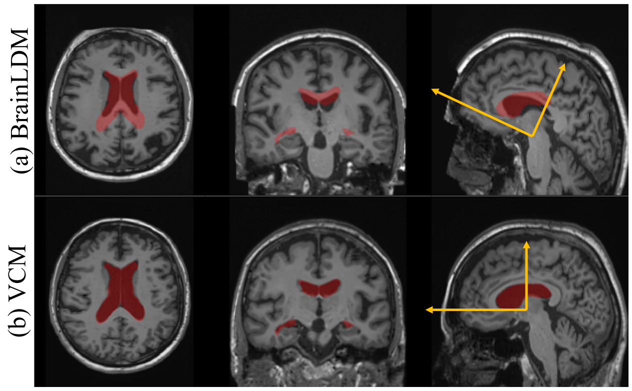

Figure 12. Spatially controlled output by VCM with the LV mask condition. (a) The original image synthesized by BrainLDM has a large displacement in terms of orientation with the given LV condition. (b) the guided output by VCM.

T1w brain MRI, precisely close to the real data, which is used to obtain the input semantics. Especially, not only the global appearance of the MRI scans (see Fig. 10), but also the generated semantic masks are highly homogeneous with the real data and the given condition (see Fig. 11).

A.4. Further analysis of spatial controls

We further analyze spatial controls through our VCM with lateral ventricle (LV) masks to generate T1w brain MRI scans. Through the following experiments and the corresponding analysis, we aim to understand what spatial control methods learn and act in guiding the generation process of pretrained diffusion models. To acquire an uncontrolled sample and a guided image with VCM, we sample both images copied from a single Gaussian noise. Subsequently, only one noise is provided the spatial controls by VCM as we illustrate in Fig. 3. Finally, the predicted noises are decoded by the BrainLDM decoder to obtain T1w brain MRI scans.

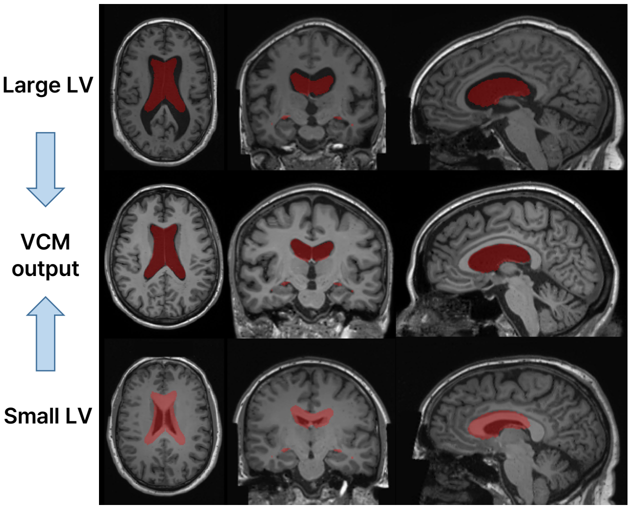

Figure 13. Examples of spatially controlled images by guiding outliers with VCM.

Appearance correction with given spatial conditions. Synthetic medical images require anatomical fidelity for plausible image synthesis. In this experiment, we assess whether VCM not only modifies the image based on given conditions, but also appropriately adjusts the overall appearance of the brain through orientation correction. Fig. 12 shows the samples generated by BrainLDM and VCM at the top of BrainLDM, respectively. If VCM learns spatial controls solely to create the LV in localized areas, the resulting output may appear unnatural, for instance, moving only the LV within the samples from BrainLDM. However, our VCM properly adjusts the orientation of the entire brain corresponding to the given LV mask, as well as accurately guides the LV in the desired area. This correction of orientation indicates that spatial control methods leverage the capabilities of the pretrained models to generate the appropriate output within the brain image distribution.

Spatial controls on outliers. We investigate the influence and strength of spatial control methods by controlling outliers synthesized from BrainLDM. Fig. 13 shows the samples from BrainLDM using large and small scalar values for LV volume and the guided image by VCM with LV mask. To make LV outlier images in the top and bottom rows of Fig. 13, we input two extrapolating values of the conditioning variables for small and large LV volumes. Although the LVs of the synthesized images from BrainLDM

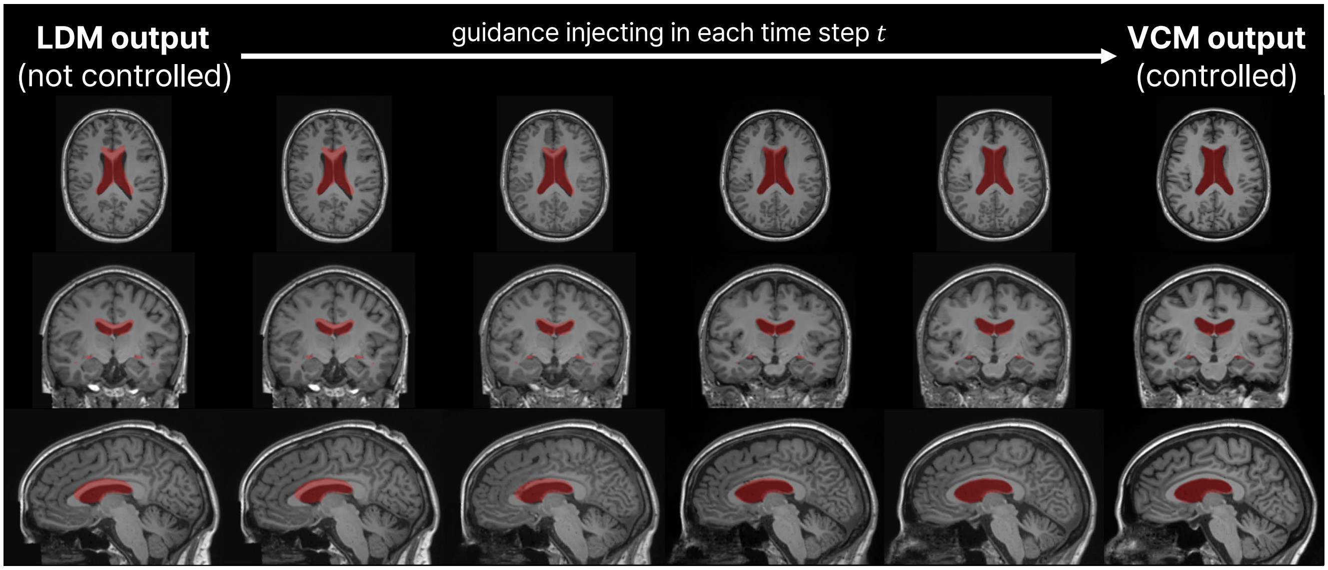

Figure 14. Visualization of spatial control guidance by VCM using the LV mask in each step of the diffusion process.

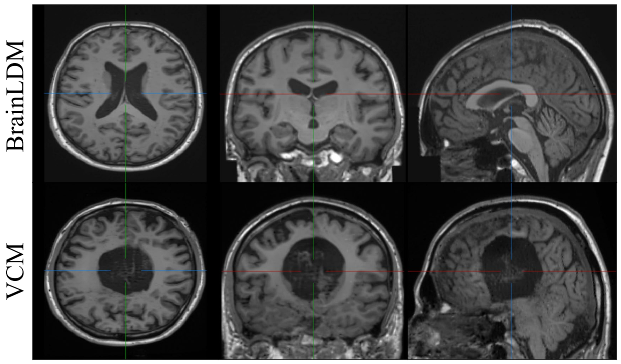

Figure 15. Synthetic image from an abnormal condition of the LV.

have significant differences with the given condition, the spatial control guides the generation process to match with the input mask, accurately.

Furthermore, we visualize the progress of spatial control of VCM by each time step in Fig. 14. In the early stage of generation, VCM mainly controls the generation steps to match the semantics of the LV and brain. After most of the semantics have been constructed by the diffusion process, the LV is not changed. Meanwhile, peripheral areas such as the cortical and subcortical areas are refined, appearing relevant structures with the given condition.

Extremely out-of-distribution conditions. Beyond various spatial control cases, we delve deeper into the investigation of robustness for VCM through abnormal input masks. To model the out-of-distribution condition, we replace the LV mask condition with a sphere-shaped mask. Since there is no training sample of the T1w brain MRI scan with the sphere-shaped LV, we can examine the robustness of our spatial control method.

Fig. 15 shows the generation results with VCM against an abnormal condition. Although the abnormal condition is unseen during VCM training, VCM controls the synthesized image to contain an LV structure localized with the given sphere mask. In addition, excluding the LV, the cortical area and skull shape show plausible appearances, as well as the background or the overall image intensity remains unaffected. However, the subcortical areas such as the corpus callosum and brain stem are destroyed. Unlike natural images whose image elements and compositions are highly diverse, medical images demand not only perceptually natural appearance but also anatomically correct structures. Consequently, the spatial control method fails to synthesize images under anatomically incorrect conditions. Thus, we suggest that researchers in the medical imaging field should be careful to use spatial control methods and properly evaluate the synthetic medical images.

B. Experimental configuration details

In this section, we provide a detailed description of the architecture and hyperparameters of VCM and the other methods for 3D implementations. In addition, we describe the details of the experiment setting.

B.1. Datasets

The data used in our study are a total of 604 healthy T1w Brain MRI scans obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [8]. All images were registered in MNI space using Advanced Normalization Tools [1]. For the 1D scalar input of BrainLDM,

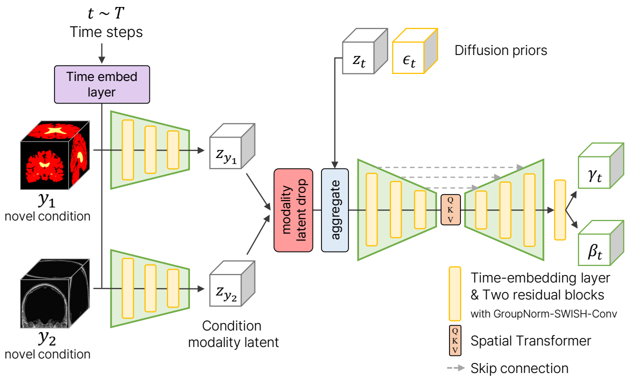

Figure 16. Details of VCM from Fig. 3. VCM takes diffusion priors, new conditions, and time steps as inputs, and outputs two modulation parameters $\{\gamma_t, \beta_t\}$ to modify the diffusion outputs.

the volume values of LV and Brain were provided through SynthSeg outputs and demographic information such as sex and age was obtained from ADNI Image and Data Archive [8]. We normalize all T1w MRI scans and skull images to have a zero-to-one value range by min-max normalization from 0 to 255 values with intensity clipping.

B.2. Implementation details

Network Construction. VCM consists of a time-conditioned asymmetric U-Net [14] with a deeper encoder than the decoder and is implemented by modifying the `DiffusionModelUNet` MONAI generative [12]. We use 16 base channels for the encoder, with a channel multiplier of [1,2,3,4,8,16], whereas the decoder is used with the last three elements of the encoder multiplier, with the same base channels. An MLP with two layers is used to embed each time step in the feature space (the purple box in Fig. 16), while a single linear layer is utilized to project time embedding in the convolutional blocks in every VCM layer (the yellow box in Fig. 16). Two residual blocks are used at every level, composed of a time embedding layer, GroupNorm, SWISH activation functions, and a convolution layer. Since the BrainLDM autoencoder maps a $1 \times 160 \times 224 \times 160$ T1w brain MRI scan to a latent vector with dimensions of $3 \times 20 \times 28 \times 20$, the last split layers of the VCM produce two modulation parameter tensors $\{\gamma_t, \beta_t\}$ in the same latent dimension.

Hyperparameters. For all experiments of Sec. 4, the linear beta scheduler for noise scheduling is used in the beta range of 0.0015 and 0.0205. We train every method using DDPM [6] with 1000 diffusion steps, while DDIM [15] is used for inference with 200 diffusion steps. Our VCM runs over 10,000 epochs with a minibatch of 16 using AdamW optimizer [9] and a base learning rate of 5×10^{-5}. The λ^{-1} of the loss for VCM in Eq. (4) is $16 \times 3 \times 20 \times 28 \times 20$. Axial flip augmentation is applied to improve generation performance according to Nichol, A. Q. and Dhariwal, P. [11], and linear learning rate annealing is employed to prevent unstable training in the early stage for all methods used in Sec. 4.

Table 5. Comparison of training time of conditional generation methods in 3D medical images.

Method	LDM (from scratch)	LDM (fine-tuning)	MCM	T2I-Adapter
# params	475.43M	553.19M	11.45M	447.26M
GPU days	2.431	2.778	1.250	4.375
Method	ControlNet	ControlNet-LITE	ControlNet-MLP	VCM
# params	193.88M	63.56M	9.03M	45.88M
GPU days	4.375	3.472	3.264	3.882

To analyze multimodal regularization techniques, we utilized two conditions: semantic map and partial image. During training VCM, the single semantic map condition, the single partial image condition, and the dual conditions appear with probability of 0.3, 0.3, and 0.4, respectively. We use separate asymmetric parts of the encoder for each condition (see Fig. 16) while keeping the number of parameters for VCM.

B.3. Comparison methods

In the experiments of Secs. 4.1 and 4.2, the *input hint* CNN encoder and `Pixelunshuffle` are also implemented in volumetric spaces for the input downsampling of ControlNet [20] and T2I-Adapter [10], respectively. In addition, we implement MCM-L [4] which has the same complexity as our VCM but utilizes the MCM architecture. In the cases of ControlNet-LITE and ControlNet-MLP, we refer to their repository [19] and replace all 2D learnable layers with 3D counterparts. For both variations of ControlNet, we employ the *input hint* CNN encoder to maintain the downsampling methods of the original paper [20]. For fine-tuning BrainLDM [13] directly, we adjust the first convolution layer to adapt additional spatial conditions, while we directly input the conditions into the model for training an LDM from scratch, referring to the Dorjsembe, Zolnamar, et al. approach [3].

In Tab. 5, we provide GPU days of the comparison methods from Sec. 4. Especially, the GPU days are analyzed by the same hyperparameters with 10,000 epochs with a minibatch of 16. For learning spatial controls within the 3D latent space, there is no significant difference among the competitive method ControlNet, lightweight variants of ControlNet (ControlNet-LITE and -MLP), and our VCM despite the differences in their model complexity.

Super-resolution For the super-resolution (SR) tasks in Sec. 4.3 and Appendix A.1, we re-implement LIIF-3D by replacing the 2D learnable operation with the 3D counterparts of LIIF [2]. Also, due to hardware limitations, we apply a patch-based approach to LIIF-3D to perform the SR in

Table 6. Quantitative evaluation of the synthetic images on the validation dataset and the results of VCM with abnormal LV mask.

Methods	# train data	FID 2D↓	FID 3D↓	LPIPS↑
BrainLDM	31,740 [16]	84.511	2.247	0.340
LDM (fine-tuning)	50 [8]	83.619	2.497	0.314
LDM (from scratch)	50 [8]	92.404	4.845	0.476
VCM (sphere masks)	-	81.444	7.515	0.331

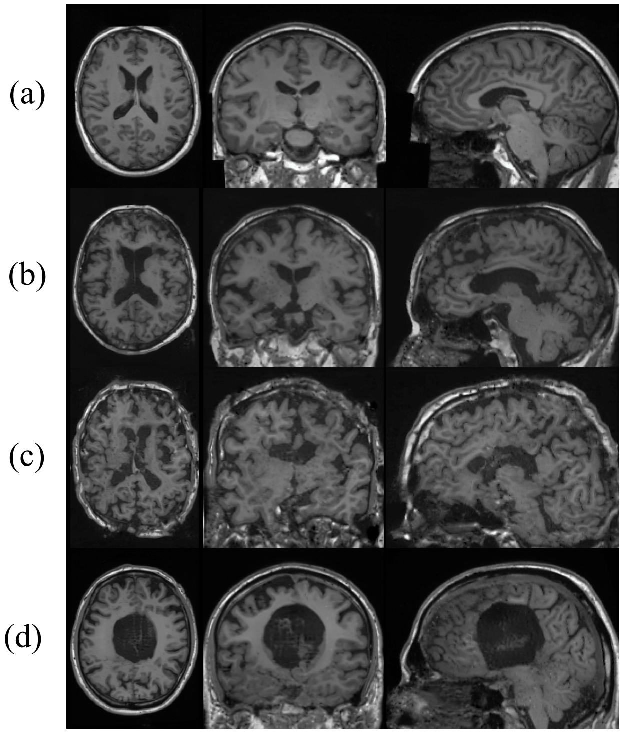

Figure 17. Examples of the generated output in Tab. 6. (a) the synthesized output of BrainLDM. (b) and (c) shows the samples from LDM trained by fine-tuning and from scratch, respectively. (d) shows the result of the abnormal condition of Appendix A.4.

volumetric space. To learn continuous image representation as in the original research [2], the model was trained with random scales in $\times 1 - \times 4$ and tested. In the case of InverseSR [17], we utilize the decoder method and their official repository.

C. Discussion for evaluation metrics

In the experiments (Sec. 4), we employ the FID [5] and LPIPS [21] to measure image quality and diversity, respectively, since they are widely used in both 2D synthetic natural images [4] and 3D medical images [3]. However, we discuss that these metrics are insufficient for analyzing synthetic medical images.

As illustrated in Tab. 6, samples from BrainLDM have 84.511 2D FID from real scans in the validation dataset of Sec. 4. However, even the wrongly synthesized images with noisy white matter in Fig. 17 (b) have a smaller 2D FID of 83.619, which means closer to the real image distribution. In addition, although 3D FID is larger than BrainLDM, it has a subtle difference, which could be misobserved in plausible synthetic images. Moreover, for the synthetic image conditioned with sphere LV masks (Fig. 17 (d)), the 2D FID is even smaller than the BrainLDM, while the 3D FID is large enough to judge that the synthetic image is far from the real image distribution. The evaluation results of the severely destroyed synthetic images (Fig. 17 (c)), which have a much more degraded and noisy appearance, finally show distinguishable large values in both 2D and 3D FID. In the case of LPIPS metrics, large values, which means high diversity in natural images, are insufficient to explain the diversity of generative models for medical images, since a severely abnormal appearance brings large LPIPS values, as shown in Fig. 17.

Based on the quantitative evaluation in Tab. 6, the following discussion points arise. First, qualitative evaluation is important in assessing the validity of synthetic medical images. Additionally, to evaluate the quality of 3D medical images, 2D FID which is widely used to measure the distance with 2D slices is less effective than 3D FID which refers to the entire 3D image. Finally, when large FID and LPIPS are measured, researchers should keep in mind that this means that the model synthesizes destroyed medical images beyond anatomically unnatural images or diverse images. Above all, innovative and effective evaluation metrics for medical images are essential for advancing research on medical image generative models.

D. Acknowledgment

This work was supported by Institute for Information communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.00223446, Development of object-oriented synthetic data generation and evaluation methods)

References

[1] Brian B Avants, Nick Tustison, Gang Song, et al. Advanced normalization tools (ants). *Insight j*, 2(365):1–35, 2009. 4

[2] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8628–8638, 2021. 1, 5, 6

[3] Zolnamar Dorjsembe, Hsing-Kuo Pao, Sodtavilan Odonchimed, and Furen Xiao. Conditional diffusion models for semantic 3d brain mri synthesis. *IEEE Journal of Biomedical and Health Informatics*, page 1–10, 2024. 5, 6

[4] Cusuh Ham, James Hays, Jingwan Lu, Krishna Kumar Singh, Zhifei Zhang, and Tobias Hinz. Modulating pretrained diffusion models for multimodal image synthesis. In

ACM SIGGRAPH 2023 Conference Proceedings, pages 1–11, 2023. 5, 6

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 5

[7] Juan Eugenio Iglesias, Benjamin Billot, Yaël Balbastre, Azadeh Tabari, John Conklin, R Gilberto González, Daniel C Alexander, Polina Golland, Brian L Edlow, Bruce Fischl, et al. Joint super-resolution and synthesis of 1 mm isotropic mp-rage volumes from clinical mri exams with scans of different orientation, resolution and contrast. *Neuroimage*, 237:118206, 2021. 2

[8] Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008. 4, 5, 6

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[10] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4296–4304, 2024. 5

[11] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 5

[12] Walter HL Pinaya, Mark S Graham, Eric Kerfoot, Petru-Daniel Tudosiu, Jessica Dafflon, Virginia Fernandez, Pedro Sanchez, Julia Wolleb, Pedro F Da Costa, Ashay Patel, et al. Generative ai for medical imaging: extending the monai framework. *arXiv preprint arXiv:2307.15208*, 2023. 5

[13] Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *MICCAI Workshop on Deep Generative Models*, pages 117–126. Springer, 2022. 2, 5

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 5

[15] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5

[16] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.*, 12(3):e1001779, Mar. 2015. 6

[17] Jueqi Wang, Jacob Levman, Walter Hugo Lopez Pinaya, Petru-Daniel Tudosiu, M Jorge Cardoso, and Razvan Marinescu. Inversesr: 3d brain mri super-resolution using a latent diffusion model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 438–447. Springer, 2023. 2, 6

[18] Heran Yang, Jian Sun, and Zongben Xu. Learning unified hyper-network for multi-modal mr image synthesis and tumor segmentation with missing modalities. *IEEE Transactions on Medical Imaging*, 42(12):3678–3689, 2023. 2

[19] Dzvinka Yarish. Controlnetlite — smaller and faster controlnet? - dzvinka yarish - medium. *Medium*, June 2023. 5

[20] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 5

[21] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6