Supplementary Material

The supplemental material is structured as follows:
1. We first provide a detailed introduction of the various diffusion models of the text and medical image modalities (CT, MRI, X-ray) in Appendix A.
2. We list all the pre-train tasks with corresponding datasets in Appendix B.
3. We also give more detailed introductions of the 5 fine-tuning tasks with 10 datasets in Appendix C.
4. We then provide more details of the LDM model architecture and configuration of four medical modalities in Appendix D.
5. To elucidate the effectiveness of the multi-flow training with the central alignment, we conduct more ablation studies of the multi-flow training strategy in Appendix E.
6. We also showcase the efficient-cost performance of our model across various training settings in Appendix F.
7. Lastly, we discuss the influence of different hyperparameters of the training process in Appendix G.

A. Introduction of the Diffusion Model

Following Versatile Diffusion [24], we employ the widely embraced UNet [19] incorporating cross attentions as the primary architecture for our diffusion model. A portion of the UNet aligns with Stable Diffusion [18], utilizing residual blocks for image data layers and incorporating cross-attention for contextual layers handling both text and image information.

A.1. Text Diffusion Model

The autoencoder of the text diffusion model is OPTIMUS [13] with the BERT [5] encoder and GPT-2 [16] decoder. It can transform sentences bidirectionally, generating 768-dimensional latent vectors from them, which follow a normal distribution. For the denoising UNet module, we adopt the 1D convolution in the residual blocks [24]. We use both CLIP [17] encoders as the prompt encoder C_T and context encoder V_T of the text modality.

A.2. Image Diffusion Model

The diffusion models for the three medical image modalities (CT, X-ray, MRI) employ the same structure which follows the Stable Diffusion 1.5 [18] and is initialized with the same weights. This way can transfer the knowledge and outstanding generation fidelity trained on extensive high-quality image datasets from Stable Diffusion [18] to our models. Same as the text diffuser, we also adopt the CLIP [17] as the prompt encoder and context encoder of the CT, MRI, and X-ray modalities.

B. Pre-train Tasks with Datasets

Tasks	Datasets	Sample Numbers
Text→Xray, X-ray→Text, Contrastive	MIMIC-CXR [12]	227k
Text→CT, CT→Text, Contrastive	MedICat [21]	131k
CT→MRI, MRI→CT, Contrastive	Brain tumor MRI and CT scan [2]	4.5k

Table A1. The pre-training tasks with corresponding datasets and the training total numbers of the samples. Contrastive: the contrastive learning for alignment of the prompt encoders.

In Table A1, we outline the training objectives for MedM2G, encompassing tasks such as medical chest X-ray report generation, medical MRI synthesis, medical multi-modal translation, and contrastive learning for aligning prompt encoders. Table A1 furnishes a summary of the datasets, tasks, and sample numbers. The pre-training datasets are collected for the

following domains: the medical image-text, text-Xray, Xray-CT, CT-MRI, which all follow the central alignment strategy for pre-training.

MIMIC-CXR MIMIC-CXR [12] is an extensive dataset containing $377,110$ chest X-rays linked to $227,827$ imaging studies. The data is collected from the Beth Israel Deaconess Medical Center. The images come with 14 labels generated through the application of two natural language processing tools to the corresponding free-text radiology reports. We adopt the MIMIC-CXR to align the text and X-ray modalities, as well as the text→X-ray and X-ray→text generation tasks.

MedICat MedICat [21] is a collection of medical images presented in context, comprising $217,000$ images sourced from $131,000$ open-access biomedical papers. The dataset encompasses captions, and inline references for 74% of the figures, and includes manually annotated subfigures and subcaptions for a subset of the figures. This dataset includes CT scans with text reports, adopted for the aligning of the text and the CT modalities (Text→CT and CT→Text generation tasks).

Brain tumor MRI and CT scan Brain tumor MRI and CT scan [2] is a novel brain tumor dataset containing $4,500$ 2D MRI-CT slices. Paired for MRI and CT scans, the dataset comprises scan data from 41 patients, with 2D slices extracted from the 3D volume. After registration, the 3D MRI and CT scans can be represented as a $237 \times 197 \times 189$ matrix. To ensure compatibility between training models and inputs, each 3D image is sliced, and $4,500$ pairs of 2D MRI-CT images are selected as the final training data. It is adopted for the alignment and generation between the CT and MRI modalities.

By adopting the multi-flow central alignment training approach, this alignment method leads to a natural and effective alignment even with limited paired data across all modalities. Significantly, it enables the implicit alignment of medical multi-modalities (CT, MRI, X-Ray) within the same space, facilitating versatile generation capabilities even in the absence of well-paired data.

C. Fine-tuning Tasks with Datasets

Tasks	Datasets	Modality	Sample Numbers
Medical Report Generation	MIMIX-CXR [12]	Text, X-ray	2,227
	IU X-ray [4]	Text, X-ray	3,955
Medical Image Generation	Chest X-ray [23]	X-ray, Text	112,120
	SLIVER07 [9]	CT	4,159
	ACDC [1]	MRI	1,902
MRI synthesis	BraTS 2020 [2]	MRI	4000
	IXI [10]	MRI	5500
MRI-CT translation	Gold Atlas male pelvis [15]	MRI,CT	1350
Chest X-ray generation	MIMIC-CXR [12]	X-ray, Text	2,227
	Chest X-ray [23]	X-ray, Text	108,948

Table A2. The 5 medical fine-tuning tasks with 10 corresponding datasets, modality, and the total sample numbers.

C.1. Medical Report Generation

MIMIX-CXR The MIMIC-CXR dataset [12], contains a comprehensive set of X-ray images, consisting of 377,100 radiology images focused on the chest region, along with 227,835 accompanying reports from patients. Following RoentGen [3], we adopt the official MIMIC split of the test set for the Chest X-ray generation task, which includes $2,227$ Xray-report samples. We employ the training subset of MIMIC-CXR for fine-tuning the medical report generation task and evaluate the test subset. The batch size for MIMIC-CXR is set to 64 and the maximum output report length is set to 50.

IU X-ray The IU X-ray [4] is the pre-dominant medical dataset employed for the medical report generation task, which contains $7,470$ chest X-ray images and $3,955$ related clinic reports from $3,955$ patients. Radiologists have provided annotations for MeSH in this dataset. The dataset comprises free-text radiology reports from clinical practices, encompassing multiple sections. We follow the original data split rates and set the batch size to 16 for the IU X-ray training. We set the maximum output report length of the IU X-ray [4] to 45.

C.2. Medical Image Generation

Chest X-ray The ChestX-ray [23] dataset consists of $112,120$ chest X-ray images in PNG format, each with a resolution of 1024×1024 pixels. We follow the original data splits of $70\%/10\%/20\%$ train/val/test for the medical image generation tasks.

SLIVER07 For the SLIVER07 [9]dataset, we utilized 20 scans available in the training dataset. Each slice was converted to a PNG image without any additional preprocessing. The dataset comprises a total of $4,159$ images, each with a resolution of 512×512 pixels.

ACDC The ACDC [1] dataset consists of 150 cardiac cine-magnetic resonance imaging (MRI) exams. We utilized the training dataset, which includes 100 exams. The images were rescaled to the range $[0, 255]$ using SimpleITK and zero-padded. Each slice was then converted into a 2D PNG image. In total, this dataset comprises $1,902$ images, each with a resolution of 512×512 pixels.

We all follow the original fine-tuning settings which undergoes end-to-end training with Adam using standard parameters (β_1 = 0.9 and β_2 = 0.999). Training occurs in mini-batches of size 16, with an initial learning rate set at 0.001. The learning rate is decayed by a factor of 10 whenever the validation loss reaches a plateau after an epoch.

C.3. MRI Synthesis Task

BraTS The BraTS [2] dataset analyzed T1, T2, and Fluid Attenuated Inversion Recovery (FLAIR) weighted brain MR images from 55 patients with gliomas, partitioned into training, validation, and test sets with 25, 10, and 20 subjects, respectively. The T2 and FLAIR volumes were registered to the T1 volume in the validation/test set. For each subject, 100 axial cross-sectional slices containing brain tissue were selected. Different scanning protocols were employed by multiple institutions.

IXI The IXI [10] dataset analyzed T1, T2, and Proton Density (PD) weighted images from 40 healthy subjects, with $(25, 5, 10)$ individuals retained for (training, validation, testing). T2 and PD volumes were registered to the T1 volume in the validation/test set. For each subject, 100 axial cross-sectional slices containing brain tissue were selected. The scanning parameters for T1 were TE=$4.6ms$, TR=$9.81ms$, for T2, TE=$100ms$, TR=$8178.34ms$, and for PD images, TE=$8ms$, TR=$8178.34ms$. The common spatial resolution was $0.94 \times 0.94 \times 1.2mm^3$.

The batch size is set to 8, and the learning rate is set to $9.6e-5$. Noise variances, ranging from β_1 = $10e-4$ to β_T = 0.02, are employed.

C.4. MRI-CT Translation

Gold Atlas male pelvis The pelvic [15] dataset analyzed T1 and T2-weighted MRI as well as CT images of 15 subjects, divided into $(9, 2, 4)$ individuals for (training, validation, testing). T1 and CT volumes were registered to the T2 volume in the validation/test set. For each subject, 90 axial cross-sectional slices were selected. For T1 scans, specifications included TE=$7.2ms$, TR=$500-600ms$, with a resolution of $0.88 \times 0.88 \times 3mm^3$, or TE=$4.77ms$, TR=$7.46ms$, with a resolution of $1.10 \times 1.10 \times 2mm^3$. For T2 scans, specifications included TE=97ms, TR=6000-6600ms, with a resolution of $0.88 \times 0.88 \times 2.50mm^3$, or TE=91-102ms, TR=12000-16000ms, with a resolution of $0.88-1.10 \times 0.88-1.10 \times 2.50mm^3$. For CT scans, specifications included a resolution of $0.10 \times 0.10 \times 3mm^3$ with Kernel=B30f or a resolution of $0.10 \times 0.10 \times 2mm^3$ with Kernel=FC17. To accelerate the synthesis task for MRI scans, $4\times$ retrospective undersampling was performed on fully sampled MRI data in 2D to obtain low-resolution images with a $16x$ acceleration rate. The training batch size is set to 64.

C.5. Chest X-ray Generation Task

MIMIC-CXR We assess the quality and clinical effectiveness of the generated chest X-rays and reports across various dimensions. Standard evaluation metrics for generative models, including FID and BLEU, are employed. A total of 208,534 studies, each containing a maximum of 3 chest X-rays with common views (PA, AP, and LATERAL3), are selected for evaluation. The dataset follows the official split of MIMIC-CXR (training set: 204,102, validation set: 1,659, test set: 2,773).

Chest X-ray The ChestX-ray [23] includes $108,948$ frontal-view X-ray images belonging to 32,717 distinct patients. The dataset is annotated with eight disease labels extracted from radiological reports using natural language processing. Each image can have multiple labels. We follow the original data splits of $70\%/10\%/20\%$ train/val/test for the chest X-ray generation task. The batch size is set to 16 with an initial learning rate of 0.001. We adopt other Adam optimizer with β_1 = 0.9 and β_2 = 0.999.

D. Model Architecture and Configuration

We provide more details of the model architecture and configuration in Table A3. Following CoDi [22], the diffusion models of four modalities (Text, CT, MRI, X-ray) are all based on the UNet structure with specific settings. The λ_1 in Eq. 4 is set to $5e-3$. The experiment for the influence of λ_1 is conducted in Appendix G.

More comparisons of MRI synthesis tasks (T1→T2, PD→T1) are listed in Table A4. Detailed comparative experiments demonstrate that our model excels in generating intricate brain sulci and tumor boundaries, effectively preserving anatomical structure.

Modality	Text LDM	X-ray LDM	CT LDM	MRI LDM
Hyperparameter				
Architecture	LDM	LDM	LDM	LDM
z-shape	768 ×1×1	4×64×64	4×64×64	4×64×64
Channels	320	320	320	320
Depth	2	4	4	4
Channel multiplier	1,2,4,4	1,2,4,4	1,2,4,4	1,2,4,4
Attention resolutions	64,32,16	64,32,16	64,32,16	64,32,16
Head channels	32	32	32	32
Number of heads	8	8	8	8
CA embed dim	768	768	768	768
Embedding Layer dim	768	768	768	768
CA resolutions	64,32,16	64,32,16	64,32,16	64,32,16
Autoencoders	Optimus	AutoKL	AutoKL	AutoKL
Weight initialization	Versatile Diffusion	SD-1.5	SD-1.5	SD-1.5
Parameterization	ϵ	ϵ	ϵ	ϵ
Learning rate	5.e-05	2e-05	1e-06	1e-06
Total batch size	1024	256	128	128
Diffusion Setup				
Diffusion steps	1000	1000	1000	1000
Noise schedule	Linear	Linear	Linear	Linear
β_0	0.00085	0.00085	0.00085	0.00085
β_T	0.012	0.012	0.012	0.012
Sampling Parameters				
Sampler	DDIM	DDIM	DDIM	DDIM
Steps	50	50	50	50
η	1.0	1.0	1.0	1.0
Guidance scale	2.0	2.0	2.0	2.0

Table A3. The architecture and configuration of different diffusion models. SD: Stable Diffusion. CA: Cross-attention layer. Embedding Layer: the embedding layer $\mathbb{F}_{emb}$.

Methods	BRATS				IXI	
	T2+T1+FLAIR→T1ce		T2+T1ce+T1→FLAIR		T2+T1→PD	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
MM-GAN [20]	$26.30_{\pm 1.91}$	$91.22_{\pm 2.08}$	$24.09_{\pm 2.14}$	$88.32_{\pm 1.98}$	$30.61_{\pm 1.25}$	$95.42_{\pm 1.78}$
Hi-Net [27]	$27.02_{\pm 1.26}$	$93.35_{\pm 1.34}$	$25.87_{\pm 2.82}$	$91.22_{\pm 2.13}$	$31.79_{\pm 1.66}$	$96.51_{\pm 2.23}$
ProvoGAN [25]	$29.26_{\pm 2.50}$	$93.96_{\pm 2.34}$	$25.64_{\pm 2.77}$	$90.42_{\pm 3.13}$	$29.93_{\pm 2.13}$	$94.62_{\pm 2.46}$
LDM [18]	$25.61_{\pm 2.48}$	$89.18_{\pm 2.55}$	$23.12_{\pm 3.16}$	$86.90_{\pm 3.24}$	$27.36_{\pm 1.96}$	$91.52_{\pm 2.16}$
CoLa-Diff [11]	$29.35_{\pm 2.40}$	$94.18_{\pm 2.46}$	$26.68_{\pm 2.74}$	$91.89_{\pm 3.11}$	$32.24_{\pm 1.86}$	$96.95_{\pm 2.61}$
ours	$\mathbf{30.12}_{\pm 1.78}$	$\mathbf{95.32}_{\pm 2.64}$	$\mathbf{27.89}_{\pm 2.84}$	$\mathbf{93.01}_{\pm 1.68}$	$\mathbf{34.12}_{\pm 1.82}$	$\mathbf{97.88}_{\pm 2.82}$

Table A4. The comparisons between our model MedM2G and advanced MRI synthesis models on BRATS and IXI datasets.

E. Ablation Study of Multi-flow Training Strategy

We conduct more ablation studies of multi-flow training in Table A5. It can be observed that models pre-trained on MIMIC-CXR [12] achieve a significant improvement in medical image-text generation tasks. Additionally, with the incorporation of the MedICat [21] pre-training dataset, accompanied by efficient-cost computational resources, the results of 5 generation tasks, including X-ray, MRI, and CT, have seen further enhancement. Furthermore, the inclusion of paired MRI-CT data has advanced the performance of unified generation across modalities, accompanied by a modest increase in computational resources.

F. More computation Costs of Different Training Settings

We provide more detailed computation costs of different training settings in Table A8. We separately computed the pretraining time and the added model parameters for the three pretraining tasks in Table A1 and each multi-flow configuration. The

Pre-train Dataset	MIMIC-CXR			ACDC	MIMIC-CXR (X-Ray generation)	BraTS2020 (T2+T1→PD)		Pelvic T2→CT		Pre-training time /epoch	Add Parameter
	BLEU-1	BLEU-4	ROUGE_L	Fid(↓)	Fid(↓)	PSNR	SSIM	PSNR	SSIM	/h	/M
MIMIC	$0.389_{\pm 0.009}$	$0.129_{\pm 0.011}$	$0.283_{\pm 0.012}$	20.13	3.1	$33.76_{\pm 2.12}$	$97.41_{\pm 1.87}$	$27.22_{\pm 0.23}$	$88.68_{\pm 1.49}$	0.7	46.4
MIMIC+MedICat	$0.399_{\pm 0.008}$	$0.136_{\pm 0.012}$	$0.298_{\pm 0.011}$	16.68	2.2	$33.98_{\pm 1.68}$	$97.67_{\pm 1.72}$	$27.38_{\pm 0.37}$	$88.99_{\pm 1.47}$	1.4	85.3
MIMIC+MedICat+MRI-CT	$\mathbf{0.412}_{\pm 0.007}$	$\mathbf{0.142}_{\pm 0.010}$	$\mathbf{0.309}_{\pm 0.009}$	**15.89**	**1.7**	$\mathbf{34.12}_{\pm 1.98}$	$\mathbf{97.88}_{\pm 1.89}$	$\mathbf{27.45}_{\pm 0.19}$	$\mathbf{89.23}_{\pm 1.54}$	**1.8**	**96.6**

Table A5. The ablation study of the pre-training datasets. MRI-CT: Brain tumor MRI and CT scan dataset [2].

Methods	Pre-train Datasets	Pre-train samples	MIMIC-CXR			ACDC	MIMIC-CXR	BraTS		Pelvic	
			BLEU-1	BLEU-4	ROUGE_L	Fid(↓)	Fid	PSNR	SSIM	PSNR	SSIM
VD [24]	original	700M	$0.356_{\pm 0.008}$	$0.008_{\pm 0.009}$	$0.254_{\pm 0.006}$	30.12	12.7	$28.97_{\pm 2.12}$	$78.45_{\pm 2.33}$	$17.87_{\pm 0.98}$	$71.43_{\pm 1.34}$
	M+M+MC	598K	$\mathbf{0.368}_{\pm 0.010}$	$\mathbf{0.112}_{\pm 0.008}$	$\mathbf{0.262}_{\pm 0.006}$	**26.67**	**9.8**	$\mathbf{29.88}_{\pm 2.13}$	$\mathbf{80.12}_{\pm 2.54}$	$\mathbf{19.21}_{\pm 1.12}$	$\mathbf{75.67}_{\pm 1.18}$
BIND [8]	original	2270K	$0.362_{\pm 0.006}$	$0.101_{\pm 0.009}$	$0.259_{\pm 0.011}$	32.14	14.6	$27.66_{\pm 1.45}$	$71.12_{\pm 1.87}$	$15.43_{\pm 1.13}$	$65.38_{\pm 1.88}$
	M+M+MC	598K	$\mathbf{0.373}_{\pm 0.005}$	$\mathbf{0.109}_{\pm 0.011}$	$\mathbf{0.265}_{\pm 0.007}$	**28.34**	**11.2**	$\mathbf{28.34}_{\pm 2.22}$	$\mathbf{75.34}_{\pm 2.32}$	$\mathbf{18.78}_{\pm 1.05}$	$\mathbf{69.23}_{\pm 1.97}$
CoDi [22]	original	512M	$0.369_{\pm 0.006}$	$0.106_{\pm 0.011}$	$0.266_{\pm 0.005}$	25.12	10.9	$29.12_{\pm 2.11}$	$80.68_{\pm 1.86}$	$19.12_{\pm 0.88}$	$73.23_{\pm 1.22}$
	M+M+MC	598K	$\mathbf{0.381}_{\pm 0.008}$	$\mathbf{0.119}_{\pm 0.009}$	$\mathbf{0.273}_{\pm 0.010}$	**22.32**	**7.8**	$\mathbf{30.78}_{\pm 2.45}$	$\mathbf{84.44}_{\pm 2.01}$	$\mathbf{22.32}_{\pm 1.43}$	$\mathbf{78.86}_{\pm 1.89}$
Ours	**M+M+MC**	**598K**	$\mathbf{0.412}_{\pm 0.007}$	$\mathbf{0.142}_{\pm 0.010}$	$\mathbf{0.309}_{\pm 0.009}$	**15.89**	**2.7**	$\mathbf{34.12}_{\pm 1.98}$	$\mathbf{97.88}_{\pm 1.89}$	$\mathbf{27.45}_{\pm 0.19}$	$\mathbf{89.23}_{\pm 1.54}$

Table A6. The comparison between MedM2G and advanced general multi-modal generative models. M+M+MC: Pre-training datasets of MIMIC-CXR, MedICat and Brain tumor MRI and CT scan.

Text-to-Image Method	Dataset Fid(↓)		
	ChestXray14	ACDC	SLIVER07
Stable Diffusion-1.4 [18]	20.13	35.32	38.76
CogView [6]	16.45	31.23	30.17
Versatile Diffusion [24]	11.43	26.67	24.39
LDM [18]	10.33	26.02	21.72
CoDi [18]	8.68	22.32	15.21
Make-a-Scene [7]	5.33	21.17	10.78
GLIDE [14]	2.89	20.19	8.45
Ours	**1.84**	**15.89**	**6.89**

Table A7. The comparison between MedM2G and advanced general text-to-image models across ChestXray, ACDC, and SLIVER07 datasets.

computation results demonstrate the superior efficiency of our model. Benefiting from the proposed central alignment strategy, our model can achieve the unification of multiple medical modalities through multi-flow training, with a linear increase in computing cost, avoiding significant computational resource consumption like others.

G. Influence of Hyperparameters

As shown in Table A9, A10 and A11, we demonstrate the influence of the various hyperparameters, including the depth of the UNet, the cross-attention embedding dimensions in the UNet, the dimension of the embedding layer $\mathbb{F}_{emb}$ in Section 3.4, the scaling size of the CLIP prompt encoder for alignment, and the balancing hyperparameter λ_1 in Eq. 4.

UNet Depth In Table A9, we separately investigate the influence of the depth of the UNet network on the experimental results for the four modalities on the MIMIC-CXR [12] and ACDC [1] datasets. We set the depth to be $2-5$ layers, where the text UNet achieved the best performance at a depth of 2, while CT, MRI, and X-ray all performed best when the depth was set to 4.

Cross-attention Embedding In Table A10, we investigate the impact of the cross-attention dimension in UNet. We conduct experiments with three settings for the embedding dimension of cross-attention: 512, 768, and 1024. It is important to note that, to align the four modalities (Text, CT, MRI, X-ray), the embedding dimension of UNET is uniformly set for all modalities. The results indicate that the optimal performance is achieved when the embedding dimension of cross-attention is set to 768.

Dimension of Embedding Layer In Table A10, we vary the dimension of the embedding layer $\mathbb{F}_{emb}$ in Section 3.4 with three settings. The best performance on downstream tasks is achieved when the encoding dimension for all four modalities

Training Settings		Pre-training time /epoch /h	Add Parameter /M
Task1	Text→X-ray	0.2	12.1
	X-ray→Text	0.2	11.7
	Contrastive	0.5	16.8
	Total	0.8	33.8
Task2	Text→CT	0.2	13.2
	CT→Text	0.2	11.9
	Contrastive	0.5	18.6
	Total	0.7	38.4
Task3	CT→MRI	0.3	19.8
	MRI→CT	0.3	18.7
	Contrastive	0.4	25.4
	Total	0.7	41.2
Single-flow(Task1)		0.8	33.8
Two-flow(Tasks1+2)		1.4	55.9
Three-flow(Task1+2+3)		1.8	96.6

Table A8. The computation costs of different training settings, including the pre-training tasks and the multi-flow strategies.

Modality	Hyper	MIMIC-CXR			ACDC
		BLEU-1	BLEU-4	ROUGE_L	Fid(↓)
Text	**2**	**0.410**$_{\pm 0.009}$	**0.141**$_{\pm 0.009}$	**0.310**$_{\pm 0.011}$	**15.86**
	3	0.407$_{\pm 0.007}$	0.135$_{\pm 0.010}$	0.303$_{\pm 0.011}$	16.67
	4	0.405$_{\pm 0.008}$	0.133$_{\pm 0.011}$	0.306$_{\pm 0.009}$	16.45
	5	0.404$_{\pm 0.009}$	0.132$_{\pm 0.010}$	0.304$_{\pm 0.008}$	16.16
CT	2	0.408$_{\pm 0.006}$	0.139$_{\pm 0.007}$	0.309$_{\pm 0.011}$	16.03
	3	0.410$_{\pm 0.009}$	0.142$_{\pm 0.008}$	0.311$_{\pm 0.011}$	15.98
	4	**0.411**$_{\pm 0.011}$	**0.144**$_{\pm 0.010}$	**0.313**$_{\pm 0.009}$	**15.91**
	5	0.408$_{\pm 0.008}$	0.143$_{\pm 0.006}$	0.312$_{\pm 0.010}$	15.99
MRI	2	0.405$_{\pm 0.006}$	0.138$_{\pm 0.008}$	0.309$_{\pm 0.007}$	16.13
	3	0.408$_{\pm 0.009}$	0.141$_{\pm 0.011}$	0.311$_{\pm 0.009}$	16.02
	4	**0.412**$_{\pm 0.008}$	**0.142**$_{\pm 0.008}$	**0.312**$_{\pm 0.011}$	**15.93**
	5	0.411$_{\pm 0.008}$	0.137$_{\pm 0.009}$	0.307$_{\pm 0.009}$	16.32
X-ray	2	0.412$_{\pm 0.012}$	0.141$_{\pm 0.014}$	0.308$_{\pm 0.011}$	16.28
	3	0.415$_{\pm 0.011}$	0.143$_{\pm 0.011}$	0.308$_{\pm 0.009}$	16.07
	4	**0.416**$_{\pm 0.010}$	**0.147**$_{\pm 0.009}$	**0.315**$_{\pm 0.007}$	**15.82**
	5	0.415$_{\pm 0.011}$	0.145$_{\pm 0.009}$	0.312$_{\pm 0.008}$	15.96

Table A9. The influence of the depth of the four different UNet hyperparameters for text, CT, MRI, and X-ray modalities.

Hyperparameter	Settings	MIMIC-CXR			ACDC
		BLEU-1	BLEU-4	ROUGE_L	Fid(↓)
CA embed	512	0.405$_{\pm 0.008}$	0.139$_{\pm 0.010}$	0.306$_{\pm 0.007}$	16.32
	768	**0.411**$_{\pm 0.009}$	**0.144**$_{\pm 0.011}$	**0.313**$_{\pm 0.012}$	**15.91**
	1024	0.408$_{\pm 0.008}$	0.141$_{\pm 0.007}$	0.309$_{\pm 0.009}$	16.12
Embedding layer	512	0.406$_{\pm 0.010}$	0.141$_{\pm 0.009}$	0.311$_{\pm 0.008}$	15.99
	768	**0.413**$_{\pm 0.013}$	**0.143**$_{\pm 0.011}$	**0.314**$_{\pm 0.008}$	**15.89**
	1024	0.409$_{\pm 0.007}$	0.138$_{\pm 0.006}$	0.312$_{\pm 0.008}$	16.04
CLIP scale size	ViT-B	0.412$_{\pm 0.011}$	0.142$_{\pm 0.009}$	0.312$_{\pm 0.010}$	15.93
	Vit-L	**0.419**$_{\pm 0.008}$	**0.151**$_{\pm 0.009}$	**0.324**$_{\pm 0.011}$	**14.89**

Table A10. The influence of the dimension of the cross-attention embedding and the embedding layer $\mathbb{F}_{emb}$, and the CLIP scale size of the prompt encoders.

Hyperparameter	Dataset Fid(↓)		
λ_1	ChestXray14	ACDC	SLIVER07
5e-04	2.89	16.78	7.32
5e-03	**1.84**	**15.89**	**6.89**
1e-03	1.99	16.12	7.02
1e-02	2.13	16.34	7.11

Table A11. The influence of the non-negative balancing hyperparameter λ_1 in Eq. 4.

is set to 768.

CLIP Scale Size As shown in Table A10, we pre-train the prompt encoder of four modalities with the ViT-Base, ViT-Large, and ViT-Huge settings. Our results demonstrate that the deeper and larger ViT model provides stronger improvements on the corresponding fine-tuning datasets. In our paper, to maintain the same settings as other models for fair comparison, we adopt the results of ViT-Base for comparison with other state-of-the-art models.

Non-negative Balancing Hyperparameter In Table A11, we explore the influence of the non-negative balancing hyperparameter λ_1 in Eq. 4 across three medical image generation datasets ChestXray [23], ACDC [1], and SLIVER07 [9]. We ran the experiment settings from $5e-4$ to $1e-2$ and found the best results for $\lambda_1 = 5e-3$, which is the same as the Barlow Twins [26].

References

[1] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018. 2, 3, 5, 7

[2] Brain tumor MRI and CT scan. https://www.kaggle.com/datasets/chenghanpu/brain-tumor-mri-and-ct-scan, 2022. Accessed: 2022-03-18. 1, 2, 3, 5

[3] Pierre Chambon, Christian Bluethgen, Jean-Benoit Delbrouck, Rogier Van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P Langlotz, and Akshay Chaudhari. Roentgen: vision-language foundation model for chest x-ray generation. *arXiv preprint arXiv:2211.12737*, 2022. 2

[4] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, pages 304–310, 2016. 2

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[6] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34: 19822–19835, 2021. 5

[7] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022. 5

[8] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 5

[9] Tobias Heimann, Bram Van Ginneken, Martin A Styner, Yulia Arzhaeva, Volker Aurich, Christian Bauer, Andreas Beck, Christoph Becker, Reinhard Beichel, György Bekes, et al. Comparison and evaluation of methods for liver segmentation from ct datasets. *IEEE transactions on medical imaging*, 28(8):1251–1265, 2009. 2, 3, 7

[10] IXI dataset. https://brain-development.org/ixi-dataset/, 2023. Accessed: 2023-02-14. 2, 3

[11] Lan Jiang, Ye Mao, Xiangfeng Wang, Xi Chen, and Chao Li. Cola-diff: Conditional latent diffusion model for multi-modal mri synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 398–408. Springer, 2023. 4

[12] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs, 2019. 1, 2, 4, 5

[13] Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. Optimus: Organizing sentences via pre-trained modeling of a latent space, 2020. 1

[14] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 5

[15] Tufve Nyholm, Stina Svensson, Sebastian Andersson, Joakim Jonsson, Maja Sohlin, Christian Gustafsson, Elisabeth Kjellén, Karin Söderström, Per Albertsson, Lennart Blomqvist, et al. Mr and ct data with multiobserver delineations of organs in the pelvic area—part of the gold atlas project. *Medical physics*, 45(3):1295–1300, 2018. 2, 3

[16] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 4, 5

[19] 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. 1

[20] Anmol Sharma and Ghassan Hamarneh. Missing mri pulse sequence synthesis using multi-modal generative adversarial network. *IEEE transactions on medical imaging*, 39(4):1170–1183, 2019. 4

[21] Sanjay Subramanian, Lucy Lu Wang, Sachin Mehta, Ben Bogin, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. Medicat: A dataset of medical images, captions, and textual references. *CoRR*, abs/2010.06000, 2020. 1, 2, 4

[22] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *arXiv preprint arXiv:2305.11846*, 2023. 3, 5

[23] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 2, 3, 7

[24] Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7754–7765, 2023. 1, 5

[25] Mahmut Yurt, Muzaffer Özbey, Salman UH Dar, Berk Tinaz, Kader K Oguz, and Tolga Çukur. Progressively volumetrized deep generative models for data-efficient contextual learning of mr image recovery. *Medical Image Analysis*, 78:102429, 2022. 4

[26] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 7

[27] Tao Zhou, Huazhu Fu, Geng Chen, Jianbing Shen, and Ling Shao. Hi-net: hybrid-fusion network for multi-modal mr image synthesis. *IEEE transactions on medical imaging*, 39(9):2772–2781, 2020. 4