

A LOW DIMENSIONAL EXAMPLES

In this section we provide a detailed comparison of the toy example discussed in Section 2.2. In addition to the results of 10 steps, we also provide the outcomes of 2 steps and 5 steps for each method. As shown in Figure 8, both RF and CFM perform poorly with only 2 steps and 5 steps, whereas our Bi-DPM successfully learns the transformation between the two sets and preserves the relationships of the paired data in the meanwhile.

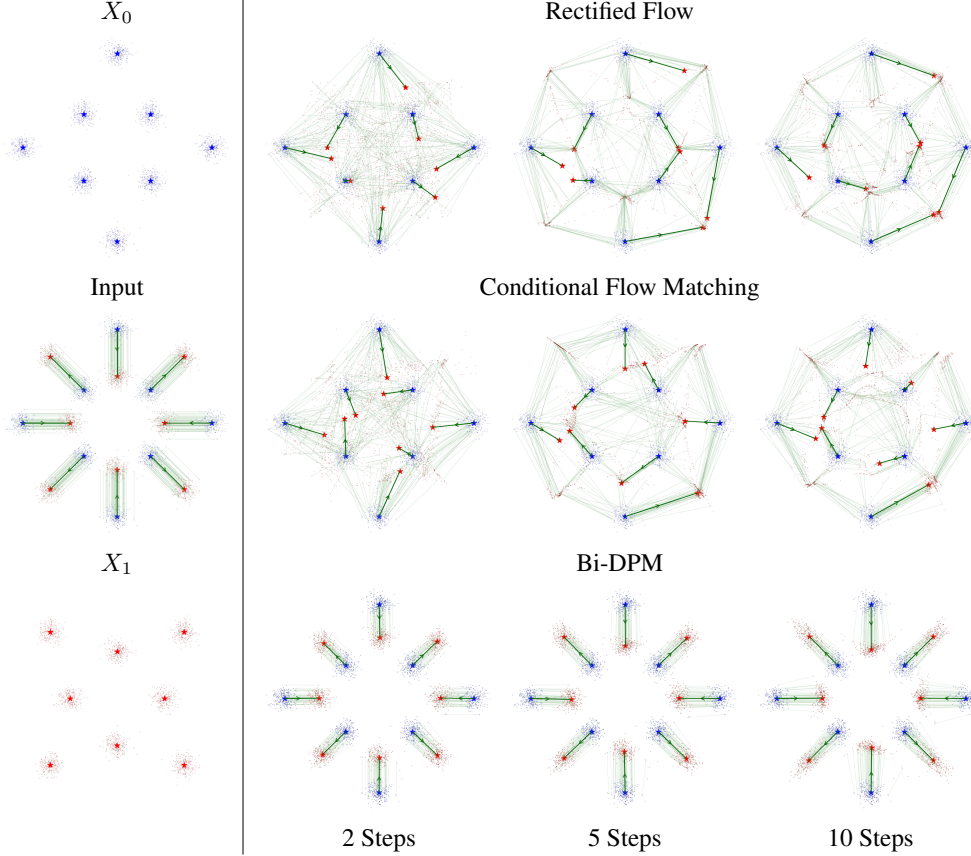


Figure 8: The performance of RF, CFM and Bi-DPM on the partially paired 8-Gaussian to 8-Gaussian toy example. For each methods, from left to right, the figure represents the results with ODE steps set to 2, 5, and 10.

Due to computation and memory constraints, here we test different number of steps on the toy examples using the datasets in Figure 8, and the L_2 distance between the generated and true data is as follows:

Table 6: The L_2 error of different step sizes on the totally paired 8-Gaussian to 8-Gaussian toy example.

	1-step	2-step	5-step	10-step
L_2 error (forward/backward)	0.015/0.015	0.009/0.008	0.011/0.012	0.013/0.019

As shown, 2-step achieves the best performance, while 1-step also performs comparably well compared to 5-step and 10-step. This partially justifies our choice of using only 1-step and 2-step in the image experiments. One intuition to use less steps is that the introduction of many intermediate steps may lead to unstable approximation, which may degrade the performance. A thorough stability analysis will be conduct in an ongoing work.

B BI-MODALITY MEDICAL IMAGE SYNTHESIS

B.1 RESULTS WITH TOTALLY PAIRED DATA

We present more visualized comparisons on CT/MRI Brain and CT/MRI Pelvis datasets, which are presented in Figure 9 and Figure 10. Moreover, the quantitative results of the comparison to other baselines on CT/MRI Pelvis with totally paired data are displayed in Table 8.

Table 7: Quantitative comparison on FID, SSIM and PSNR with 100% paired data. The **bold** data represent the best results and the underlined ones indicates the second best.

			RF	I-CFM	OT-CFM	VP-CFM	CycleGAN	SynDiff	Reg-GAN	Bi-DPM (1-step)	Bi-DPM (2-step)	
MRI T1/T2	T1	SSIM $\uparrow$	0.841 ± 0.038	0.837 ± 0.040	0.729 ± 0.070	0.710 ± 0.042	0.652 ± 0.025	0.832 ± 0.041	0.809 ± 0.038	0.869 ± 0.038	0.862 ± 0.038	
		FID $\downarrow$	49.004	85.069	168.269	63.826	80.771	97.494	85.767	40.611	57.800	
		PSNR $\uparrow$	22.175 ± 2.385	21.843 ± 2.392	19.131 ± 2.294	18.648 ± 2.124	17.588 ± 1.716	20.377 ± 2.491	20.676 ± 1.926	23.117 ± 2.471	22.886 ± 2.449	
	T2	SSIM $\uparrow$	0.833 ± 0.040	0.822 ± 0.040	0.666 ± 0.063	0.660 ± 0.044	0.633 ± 0.033	0.823 ± 0.043	0.805 ± 0.039	0.866 ± 0.041	0.857 ± 0.039	
		FID $\downarrow$	37.825	41.780	163.686	50.477	68.749	49.568	88.087	32.366	38.960	
		PSNR $\uparrow$	23.307 ± 1.839	22.743 ± 1.829	19.182 ± 1.857	18.843 ± 1.686	19.121 ± 1.257	21.965 ± 1.945	21.589 ± 1.451	24.845 ± 1.994	24.302 ± 1.944	
	CT/MRI Brain	CT	SSIM $\uparrow$	0.828 ± 0.046	0.828 ± 0.046	0.672 ± 0.081	0.694 ± 0.058	0.650 ± 0.057	0.796 ± 0.056	0.817 ± 0.045	<u>0.831</u> ± 0.046	0.832 ± 0.046
			FID $\downarrow$	62.425	70.691	136.472	81.583	108.584	70.172	38.606	33.426	34.557
			PSNR $\uparrow$	23.817 ± 1.856	24.122 ± 1.951	19.252 ± 1.956	21.366 ± 1.351	18.832 ± 0.985	22.344 ± 1.966	23.148 ± 1.790	23.656 ± 2.124	23.853 ± 2.080
		MRI	SSIM $\uparrow$	0.678 ± 0.044	0.685 ± 0.046	0.467 ± 0.071	0.597 ± 0.047	0.445 ± 0.037	0.503 ± 0.052	0.683 ± 0.050	<u>0.723</u> ± 0.047	0.726 ± 0.044
			FID $\downarrow$	56.972	85.528	147.784	72.462	116.487	71.952	31.856	29.991	31.452
			PSNR $\uparrow$	21.121 ± 1.174	21.131 ± 1.159	18.451 ± 1.521	19.650 ± 1.175	16.744 ± 0.927	17.172 ± 0.713	20.436 ± 1.339	<u>21.140</u> ± 1.364	21.158 ± 1.340

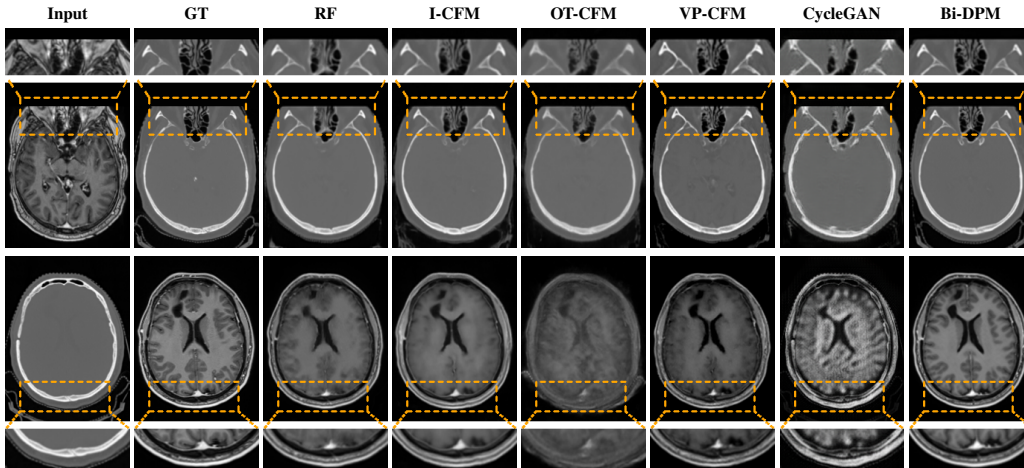


Figure 9: The synthetic images of **CT/MRI Brain** dataset for different methods.

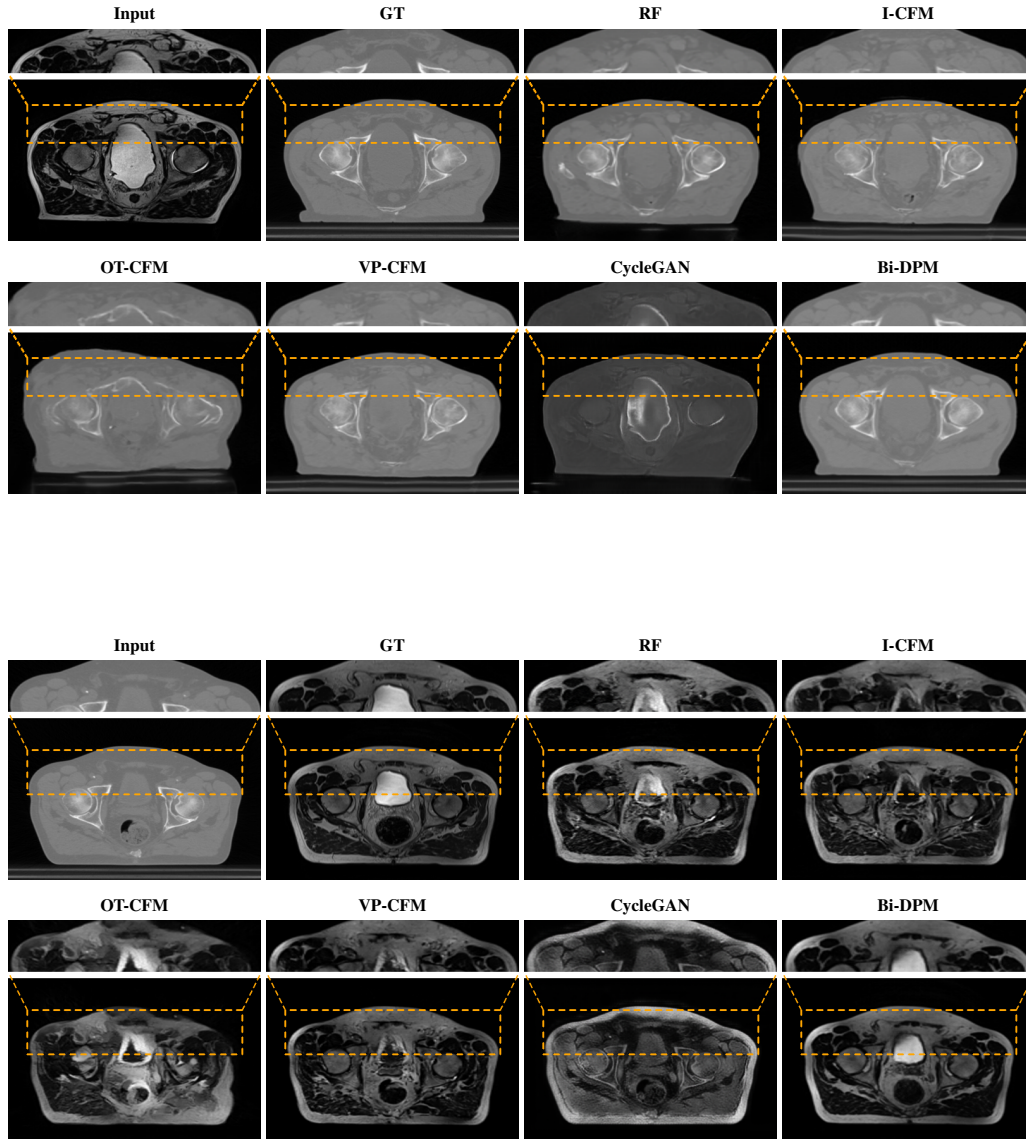


Figure 10: The synthetic images of **CT/MRI Pelvis** dataset for different methods.

Table 8: Quantitative comparison on FID, SSIM and PSNR with 100% paired data. The **bold** data represent the best results and the underlined ones indicates the second best.

		RF	I-CFM	OT-CFM	VP-CFM	CycleGAN	Bi-DPM (1-step)	Bi-DPM (2-step)	
CT/MRI Pelvis	CT	SSIM $\uparrow$	0.815	0.810	0.788	0.710	0.642	0.806	<u>0.812</u>
			$\pm$ 0.046	$\pm$ 0.042	$\pm$ 0.048	$\pm$ 0.058	$\pm$ 0.054	$\pm$ 0.052	$\pm$ 0.049
		FID $\downarrow$	71.626	72.556	107.349	68.122	99.398	46.600	<u>48.363</u>
			23.591	23.779	22.104	22.010	19.672	<u>23.796</u>	24.033
	MRI	PSNR $\uparrow$	$\pm$ 2.598	$\pm$ 2.279	$\pm$ 2.538	$\pm$ 1.932	$\pm$ 2.128	$\pm$ <u>2.313</u>	$\pm$ 2.596
			0.535	0.532	0.471	0.415	0.231	<u>0.571</u>	0.577
			$\pm$ 0.052	$\pm$ 0.054	$\pm$ 0.081	$\pm$ 0.038	$\pm$ 0.042	$\pm$ <u>0.051</u>	$\pm$ 0.052
		FID $\downarrow$	73.998	71.110	108.017	83.008	138.037	<u>68.381</u>	68.209
	17.892	17.641	16.090	16.364	14.428	<u>18.675</u>	18.930		
	$\pm$ 1.568	$\pm$ 1.497	$\pm$ 2.460	$\pm$ 1.512	$\pm$ 1.338	$\pm$ <u>1.687</u>	$\pm$ 1.668		

B.2 RESULTS WITH PARTIALLY PAIRED DATA

The quantitative results of the comparison to other baselines on CT/MRI Pelvis with 0.1 ratio of paired data are displayed in Table 9. Besides, more results about the tendency of the quality evaluation indices with regard to paired ratio are shown in Figure 11 and Figure 12.

Table 9: Quantitative comparison on FID, SSIM and PSNR with 10% paired data.

		RF	I-CFM	Bi-DPM (1-step)	Bi-DPM (2-step)
CT/MRI Pelvis	CT	SSIM $\uparrow$ ± 0.038 (0.815) ± 0.046	0.684 (0.810) ± 0.057 (± 0.042)	0.783 (0.806) ± 0.053 (± 0.052)	0.785 (0.812) ± 0.057 (± 0.049)
		FID $\downarrow$	147.101 (71.626)	102.730 (72.556)	52.517 (46.600)
		PSNR $\uparrow$	18.518 (23.591) ± 1.692 (± 2.598)	19.381 (23.779) ± 2.085 (± 2.279)	22.531 (23.796) ± 2.306 (± 2.313)
	MRI	SSIM $\uparrow$ ± 0.078 (± 0.052)	0.339 (0.535) ± 0.063 (± 0.054)	0.523 (0.571) ± 0.048 (± 0.051)	0.532 (0.577) ± 0.056 (± 0.052)
		FID $\downarrow$	138.121 (73.998)	166.533 (71.110)	76.436 (68.381)
		PSNR $\uparrow$	13.696 (17.892) ± 2.662 (± 1.568)	12.062 (17.641) ± 2.566 (± 1.497)	17.718 (18.675) ± 1.559 (± 1.687)

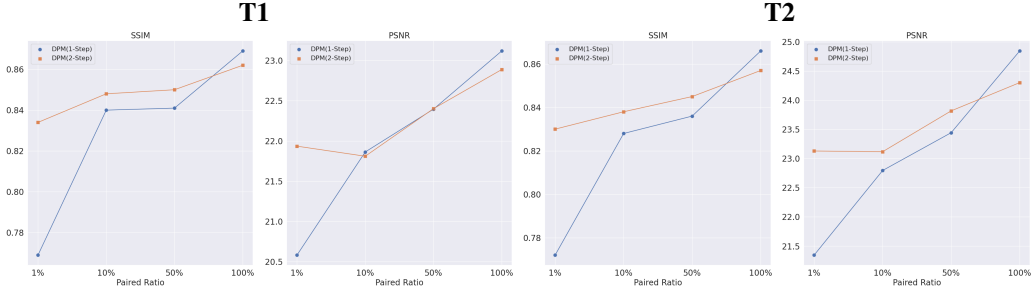


Figure 11: The quantitative comparison results on MRI T1/T2 dataset between different methods with various discrete ODE steps. From the top to the bottom, the figures show the results of synthetic CT and synthetic MRI. The indices are SSIM, and PSNR from left to the right.

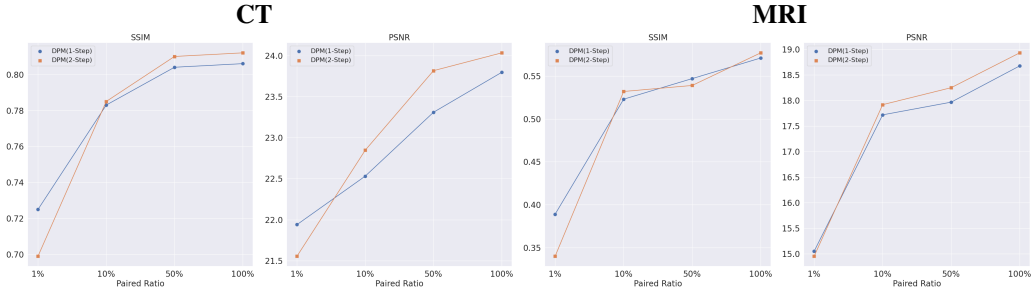


Figure 12: The quantitative comparison results on CT/MRI Pelvis dataset between different methods with various discrete ODE steps. From the top to the bottom, the figures show the results of synthetic CT and synthetic MRI. The indices are SSIM, and PSNR from left to the right.

B.3 ITERATION STEPS OF ODE

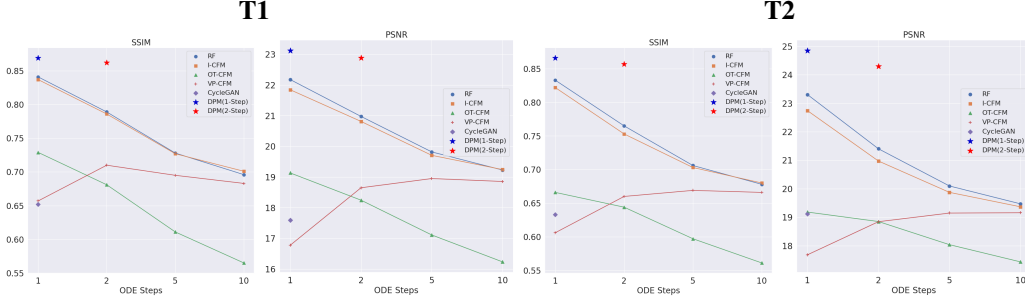


Figure 13: The quantitative comparison results on MRI T1/T2 dataset between different methods with various discrete ODE steps.

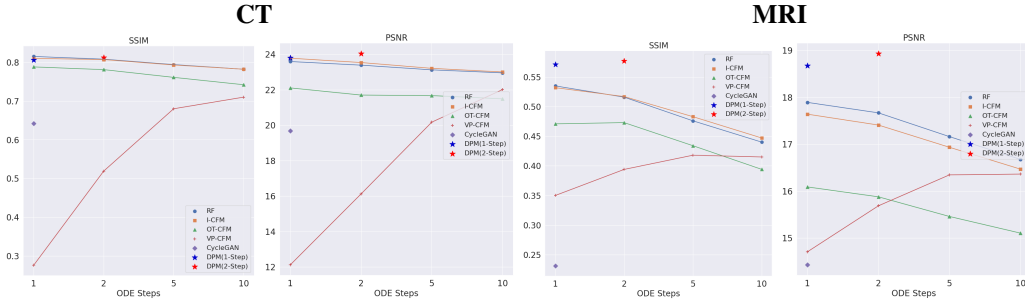


Figure 14: The quantitative comparison results on CT/MRI Pelvis dataset between different methods with various discrete ODE steps.

C TIME AND MEMORY COST

We also compare the memory cost in training process and the synthesis time cost for different methods. In training process, the batch size is set to 10. In testing, based on the MRI T1/T2 dataset, which contains 251×2 test images totally, we calculate the synthesis time cost of the whole dataset. For each method, we synthesize one image at a time and here are the results:

Table 10: Inference time cost comparison

	RF (RK45)	CFM (RK45)	CycleGAN	Reg-GAN	SynDiff	DPM (1-step)	DPM (2-step)
Time Cost	834s	834s	31s	10s	394s	20s	38s

Table 11: Training memory cost comparison

	RF (RK45)	CFM (RK45)	CycleGAN	Reg-GAN	SynDiff	DPM (1-step)	DPM (2-step)
Memory Cost	26G	36G	22G	12G	56G	48G	63G

D SEGMENTATION RESULTS

To assess the quality of the synthetic images, we evaluate their performance in a downstream segmentation task, representing a practical application of image synthesis. Using the MRI T1/T2 dataset from BraTS2021, we first train an nnUNet on real T1 and T2 training dataset. Subsequently, we test various combinations of synthetic and real data: baseline (T1 real and T2 real), T1 fake + T2 real and T1 real + T2 fake. The corresponding dice scores predicted by nnUNet are as follows:

972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025

Table 12: The segmentation results on MRI T1/T2 datasets.

	RF	CFM	CycleGAN	Reg-GAN	SynDiff	DPM
T1 Fake T2 Real	0.814	0.812	0.758	0.774	0.781	0.816
T2 Fake T1 Real	0.690	0.662	0.519	0.642	0.619	0.716
Baseline Both Real	0.818					