Unsupervised Cardiac Video Translation Via Motion Feature Guided Diffusion Model

Swakshar Deb[1], Nian Wu[1], Frederick H. Epstein[2], and Miaomiao Zhang[1,3]

[1] Department of Electrical and Computer Engineering, University of Virginia, USA
[2] Department of Biomedical Engineering, University of Virginia, USA
[3] Department of Computer Science, University of Virginia, USA

Abstract. This paper presents a novel motion feature guided diffusion model for unpaired video-to-video translation (MFD-V2V), designed to synthesize dynamic, high-contrast cine cardiac magnetic resonance (CMR) from lower-contrast, artifact-prone displacement encoding with stimulated echoes (DENSE) CMR sequences. To achieve this, we first introduce a Latent Temporal Multi-Attention (LTMA) registration network that effectively learns more accurate and consistent cardiac motions from cine CMR image videos. A multi-level motion feature guided diffusion model, equipped with a specialized Spatio-Temporal Motion Encoder (STME) to extract hierarchical coarse-to-fine motion conditioning, is then developed to improve synthesis quality and fidelity. We evaluate our method, MFD-V2V, on a comprehensive cardiac dataset, demonstrating superior performance over the state-of-the-art in both quantitative metrics and qualitative assessments. Furthermore, we show the benefits of our synthesized cine CMRs improving downstream clinical and analytical tasks, underscoring the broader impact of our approach. Our code is publicly available at https://github.com/SwaksharDeb/MFD-V2V.

Keywords: Unsupervised video translation · Latent temporal motion · Generative diffusion · Cardiac MRI.

1 Introduction

Cardiac Magnetic Resonance (CMR) imaging plays an important role in assessing myocardial strain, which is a key indicator of cardiac dysfunction [6, 30]. Among various CMR techniques, displacement encoding with stimulated echoes (DENSE) CMR has demonstrated superior performance in capturing myocardial motion for highly accurate and reliable strain measurement [23, 27, 18]. In contrast to standard cine CMR, which relies on balanced steady-state free precession sequences optimized for high signal-to-noise ratio (SNR) and strong tissue contrast, DENSE CMR encodes myocardial displacement and deformation using stimulated echoes, allowing for high-resolution regional and segmental strain analysis. However, this process comes at the cost of lower SNR, as stimulated echoes retain only a fraction of the original magnetization, leading to reduced

signal intensity and noisier magnitude images [1]. Despite that DENSE CMR excels in providing highly detailed quantitative myocardial motion/strain data, its lower image quality presents challenges for tasks such as segmentation, feature extraction, and texture-based analysis, limiting its applicability in certain clinical applications [12].

Recent advancements in deep learning and generative models offer a promising new approach to synthesize high-quality cine CMRs from DENSE CMRs, which remains unexplored in the literature. Earlier research on video-to-video (V2V) translation using generative adversarial networks (GANs) has demonstrated their ability to transform video sequences across domains [4, 7]. However, GAN-based methods often suffer from limited diversity in their translations and are prone to mode collapse, restricting the variability of synthesized outputs [20]. More recently, diffusion-based generative models have emerged as a powerful alternative, offering superior diversity, fidelity, and stability over GANs [22]. Despite their advantages, most diffusion-based V2V models [16, 19, 9, 28] rely on a paired (supervised) training scenarios, where aligned source-target video pairs are required. However, cine and DENSE CMRs do not naturally form a paired dataset, as they are acquired using fundamentally different imaging sequences, leading to significant disparities in spatial and temporal resolution.

In this paper, we introduce a novel motion feature guided diffusion model for unpaired cardiac MRI V2V translation, termed MFD-V2V. Our method leverages the rich spatiotemporal motion information embedded in cine CMR videos by conditioning a generative diffusion model on learned motion features. Rather than using raw motion fields, we extract multi-level motion features that capture both coarse and fine-grained motion characteristics. Conditioning on these hierarchical features provides more informative guidance to the diffusion process, resulting in outputs that are both more realistic and temporally consistent. During inference, we leverage the displacement motion field provided by DENSE-CMR to generate realistic and temporally consistent cine CMR sequences; hence bridging the gap in unpaired cardiac MRI translation with improved anatomical and motion fidelity. Our contributions are threefold:

(i) We are the first to develop a generative diffusion V2V model to synthesize high quality cine CMR from DENSE CMR sequences.
(ii) Introduce a latent temporal multihead attention (LTMA) based registration network to effectively learn spatiotemporal motion from cardiac video sequences.
(iii) Develop a generative video diffusion model conditioned on multi-level motion features extracted by a specialized spatiotemporal motion encoder (STME) to enhance synthesis quality and fidelity.

We validate MFD-V2V on cardiac MR images collected from multiple sites [11, 18]. Experimental results demonstrate that MFD-V2V surpasses existing methods [15, 32, 4, 33, 8] by generating more realistic and temporally coherent cardiac MR videos. Furthermore, we highlight the advantages of our synthesized data in improving performance on a downstream CMR myocardium segmentation task.

2 Background: Video Diffusion Model

This section briefly reviews the concept of the video diffusion model (VDM) [15], which is an extension of the image-based diffusion model introduced in [14]. The VDM serves as a foundation for our proposed video translation model.

Given an image, x_0, sampled from the real data distribution $q(x)$, the forward process in a diffusion model is defined as a Markov chain that gradually adds Gaussian noise to x_0 over L timesteps. This process is governed by a variance scheduler $\{\beta_t \in (0,1)\}_{t=1}^{L}$, where $t \in \{1, \cdots, L\}$ denotes the diffusion timestep. The forward process is formally expressed as

$$q(x_{1:L}|x_0) = \prod_{t=1}^{L} q(x_t|x_{t-1}), \text{where } q(x_t|x_{t-1}) = \mathcal{N}(x_t, \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}),$$

with $\mathbf{I}$ representing an identity matrix. Using the notation $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, the direct formulation of the noising process from x_0 to x_t is $q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I})$. Similarly, in the reverse process, we model the joint distribution, $p_\theta(x_{0:L})$, as a markov chain starting from $p(x_L) = \mathcal{N}(x_L; 0, \mathbf{I})$, i.e.,

$$p_\theta(x_{0:L}) = p(x_L) \prod_{t=1}^{L} p_\theta(x_{t-1}|x_t), \text{where } p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \mathbf{\Sigma}_\theta(x_t, t)).$$

Here, μ_θ and $\mathbf{\Sigma}_\theta$ are the predicted mean and variance respectively. Following the prior work [14], we reparameterize the mean μ_θ using a noise prediction network ϵ_θ, while setting the variance to the identity matrix. The model is then trained by minimizing the objective, $L_\epsilon = \mathbb{E}_{x_t, \epsilon \sim \mathcal{N}(0,I)}[||\epsilon - \epsilon_\theta(x_t, t)||^2]$.

The VDM [15] builds upon the previously introduced image-based diffusion framework by employing a 3D U-Net [10], which factorizes both spatial and temporal dimensions by replacing standard 2D convolutions with 3D convolutions. Additionally, VDM incorporates spatial and temporal self-attention blocks to effectively capture structural details and motion dynamics across frames.

3 Our Method: MFD-V2V

This section introduces MFD-V2V, an unsupervised video translation model that for the first time synthesizes high-quality cine CMR from DENSE-CMR video sequences via a motion-guided diffusion model. Our method consists of two main components: (i) a latent temporal multihead attention registration network, LTMA, to learn temporally consistent and continuous motion from CMR video sequences; and (ii) a generative video diffusion model conditioned on multi-level motion features extracted by a spatiotemporal motion encoder, STME. An overview of the architecture is illustrated in Fig. 1.

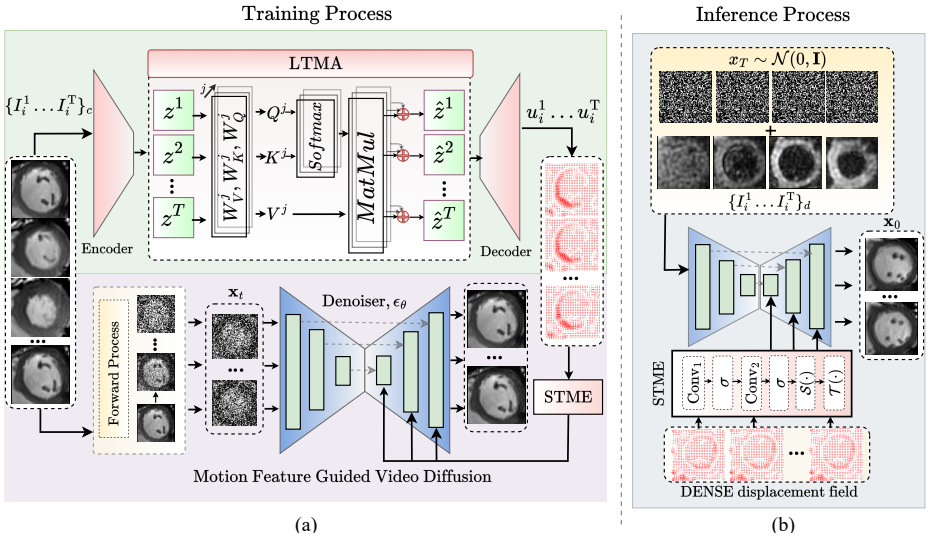

Fig. 1. An overview of our proposed MFD-V2V model. Left to right: (a) the training process with LTMA registration and multli-level motion feature guided video diffusion. (b) the inference process, where encoded DENSE motion features are used as condition.

3.1 LTMA Registration Network

Given a training dataset of N video sequences, where each sequence includes $T+1$ time frames, denoted as $\{I_i^\tau\}, i \in \{1, \cdots, N\}, \tau \in \{0, \cdots, T\}$. That is to say, for the ith training data, we have a sequence of image frames $\{I_i^0, \cdots, I_i^T\}$. By setting the first frame $\{I_i^0\}$ as a reference (source) image, there exists a number of T pairwise images, $\{(I_i^0, I_i^1), (I_i^0, I_i^2) \cdots, (I_i^0, I_i^T)\}$, to be aligned by their associated motion/displacement field $\{u_i^1, u_i^2 \cdots, u_i^T\}$. Similar to [3, 29], we employ U-Net architecture as the backbone of our registration encoder, $\mathcal{E}_{\theta_v}$, and decoder, $\mathcal{D}_{\theta_v}$, parameterized by θ_v. The encoder $\mathcal{E}_{\theta_v}$ projects the input image sequences into a latent velocity space $\mathbf{Z}_i = [z_i^1, z_i^2, \ldots, z_i^T] \in \mathbb{R}^{T \times H \times W \times C}$. Here C is the number of feature channels, and H, W represent the height and weight of the encoded motion, respectively. The decoder $\mathcal{D}_{\theta_v}$ is then used to project the latent features back to the input image space.

In contrast to existing methods that model temporal motion sequentially using recurrent residual networks [29] or long short-term memory [21], we leverage a self-attention mechanism to capture long-range temporal dependencies directly in the encoded velocity space across time. This enables a more expressive and scalable representation learning of cardiac dynamics by directly modeling global temporal interactions across all time frames.

For each encoded motion feature $\mathbf{Z}_i$, we define $\mathbf{Q}_i^{(j)}$, $\mathbf{K}_i^{(j)}$, $\mathbf{V}_i^{(j)}$ as the Query, Key, and Value matrices for the j-th attention head. Following similar principles

of [25], the output feature $\hat{\mathbf{Z}}_i$ of our proposed LTMA module is formulated as

$$\begin{aligned} \mathbf{Q}_i^{(j)} = \mathbf{W}_Q^{(j)}\mathbf{Z}_i, \mathbf{K}_i^{(j)} = \mathbf{W}_k^{(j)}\mathbf{Z}_i, \mathbf{V}_i^{(j)} = \mathbf{W}_V^{(j)}\mathbf{Z}_i, \\ \hat{\mathbf{Z}}_i = \oplus_{j=1}^{h}\text{Softmax}\big(\mathbf{Q}_i^{(j)}\mathbf{K}_i^{(j)T}\big)\mathbf{V}_i^{(j)}, \end{aligned} \tag{1}$$

where $\oplus$ denotes the concatenation operation, h is the total number of attention head and $\mathbf{W}_Q^{(j)}$, $\mathbf{W}_K^{(j)}$, $\mathbf{W}_V^{(j)}$ are the linear projection matrix associated with the j-th attention head.

Network loss. By defining Θ for all network parameters, we finally formulate the loss function of LTMA registration network as

$$l(\theta) = \sum_{i=1}^{N}\sum_{t=1}^{T}\lambda||I_i^1 \circ \phi_i^t(v_i^t(\hat{z}_i^t);\Theta) - I_i^t||_2^2 + ||\nabla v_i^t(\hat{z}_i^t;\Theta)||^2 + \text{reg}(\Theta), \tag{2}$$

where ϕ_i^t represents the transformation fields between pairwise frames, parameterized by a stationary velocity field v_i^t over time [26]. Please refer to [26] for further details. The motion or displacement field can then be computed as $u_i^t = \phi_i^t - id$, where id denotes an identity transformation (i.e., the original image grid). The λ is a positive weighting parameter, and reg$(\cdot)$ denotes the regularization function applied to the network.

3.2 Multi-Level Motion Feature Guided Video Diffusion

Given the learned motion fields, $\mathbf{u}_i = \{u_i^t\}_{t=1}^T$, from our previously introduced LTMA registration network, we first present STME - a spatiotemporal motion encoder designed to extract hierarchical motion features that guide the video diffusion model. Such a STME module captures both coarse and fine-grained dynamics from the motion fields to provide effective conditioning. Intuitively, we first apply two consecutive 3D convolution layers with nonlinear ReLU activations (σ) to the input motion sequence $\mathbf{u}_i$. These feature maps are then passed through spatial and temporal attention layers, $\mathcal{S}(\cdot)$ and $\mathcal{T}(\cdot)$, respectively, to extract fine-grained spatiotemporal features of the motion pattern [15]. We define the overall output of the STME block, $\mathbf{F} \in \mathbb{R}^{T\times H\times W\times C}$, as

$$\mathbf{F} = \mathcal{T} \circ \mathcal{S} \circ \sigma \circ \text{Conv}_2 \circ \sigma \circ \text{Conv}_1(\mathbf{u}_i), \tag{3}$$

where $\circ$ is the function composition. For motion conditioning, we employ cross attention [31, 32] within the decoder of the denoising network ϵ_θ (as illustrated in Fig. 1). Specifically, the learned motion feature map, $\mathbf{F}$, is treated as key and value, while the decoder's latent representation serves as the query. At the i-th decoder layer, given the latent feature, $\mathbf{h}_i \in \mathbb{R}^{T\times H'\times W'\times C'}$, we compute the cross-attention as $\text{CrossAttn}(\mathbf{h}_i, \mathbf{F}) = \text{Softmax}\left(\mathbf{Q}_{h_i}\mathbf{K}_F^T\right)\mathbf{V}_F$, where $\mathbf{Q}_{h_i} = \mathbf{W}_Q\mathbf{h}_i$, $\mathbf{K}_F = \mathbf{W}_K\mathbf{F}$ and $\mathbf{V}_F = \mathbf{W}_V\mathbf{F}$, with $\mathbf{W}_V$, $\mathbf{W}_K$, $\mathbf{W}_V$ being learnable projection matrices. To ensure compatibility across decoder layers, the motion feature map $\mathbf{F}$ is resized to the appropriate spatial dimensions via 1×1 convolution. Finally, the overall training objective is then defined as $L_\epsilon = \mathbb{E}_{x_t,\epsilon\sim\mathcal{N}(0,I)}[||\epsilon - \epsilon_\theta(x_t, t, \mathbf{F})||^2]$.

Inference Process. During inference, we initialize with random Gaussian noise, $x_T \sim \mathcal{N}(0, \mathbf{I})$, incorporated with DENSE, $\{I_i^1, \ldots, I_i^T\}_d$. The displacement field, directly provided with the DENSE CMR sequences [11], is used to extract motion conditions (see Fig. 1(b)). We employ the standard DDPM sampler [14] to iteratively synthesize Cine CMR.

4 Experimental Evaluation

We evaluate the effectiveness of our proposed model, MFD-V2V, on cardiac CMR video sequences by comparing it against five state-of-the-art generative models: CycleGAN [33], NiceGAN [8], RecycleGAN [4], VDM [15], and ControlNet [32]. The VDM model is trained without conditioning, while ControlNet is conditioned on the displacement field provided by our module LTMA. Since ControlNet is a fine-tuning approach, we use VDM as its backbone (in Tab. 1) and train only the ControlNet component, following the method in [32].

Dataset. We utilize 741 cine and DENSE CMR videos collected from 284 subject, including 124 healthy volunteers and 160 patients with various types of heart disease [11, 18]. All cine and DENSE CMR sequences were temporally and spatially aligned for efficient network training. In particular, the standard cine sequences were temporally resampled to 40 frames to match the DENSE temporal resolution. The average temporal resolutions of Cine and DENSE CMR sequences are 30ms and 17ms, respectively. In our experiments, all images were resampled at 1.0 mm^2 resolution and cropped to the left ventricular (LV) regions of the size 64×64. Both Cine and DENSE sequences have their corresponding LV segmentation masks manually annotated.

4.1 Experimental Design and Implementation Details

Similar to [15], we adopt the 3D-UNet as our denoiser. For the training of the denoiser, we use the Adam optimizer with a learning rate of 10^{-5}, and a batch

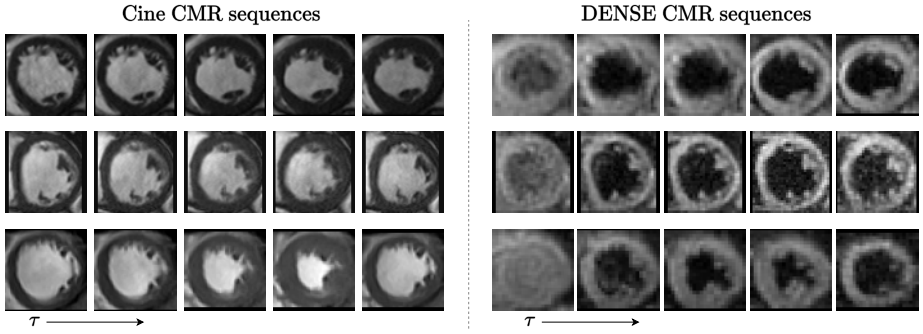

Fig. 2. Left to right: examples of unpaired Cine and DENSE cardiac sequences.

size of 20. We use the sigmoid-beta noise scheduler [17] with total of 1000 diffusion timestep. We used a 70/10/20 train/validation/test split at the subject level for all experiments. The model is trained on four NVIDIA A100 GPUs.

Evaluation Metrics. Since Cine and DENSE do not come with exact pair, this makes direct image-by-image comparison (e.g., PSNR, SSIM, etc) challenging. Therefore, we evaluate our model by assessing the distributional similarity between the synthesized and original cine CMR sequences. Specifically, we employ the following metrics: Fréchet Inception Distance (FID) [13], Kernel Inception Distance (KID) [5], Fréchet Video Distance (FVD) [24], and FID-VID [2]. In particular, FID and KID assess the appearance quality of individual frames, while FVD and FID-VID evaluate the temporal dynamics across the sequence.

4.2 Results

Tab. 1 presents the quantitative evaluation of our model vs. all baseline methods. Overall, our approach consistently outperforms the baselines across all evaluation metrics, demonstrating its ability to generate more realistic and higher-quality cine CMR videos from DENSE inputs. Among the baselines, VDM [15] surpasses GAN-based models, and its performance further improves when incorporating motion conditioning, as shown by ControlNet [32].

Fig. 3 displays example cine CMR sequences generated by all methods over time. Our approach clearly demonstrates superior spatial and temporal coherence throughout the entire video, producing more realistic and consistent CMR sequences compared to the baselines.

Table 1. Quantitative comparison of generated image sequences with baseline methods. The best and second best performances are highlighted with bold and underlined, respectively.

Category	Method	Condition	Reg. Net.	FID ↓	KID ↓	FVD ↓	FID-VID ↓
GAN	CycleGAN [33]	×	×	135.25	0.3071	141.71	68.567
	NiceGAN [8]	×	×	122.43	0.2081	131.77	77.013
	RecyleGAN [4]	×	×	97.996	0.1591	126.11	68.214
DMs	VDM [15]	×	×	<u>57.235</u>	0.0483	75.487	30.148
	ControlNet [32]	Motion	LTMA	61.452	<u>0.0449</u>	<u>70.217</u>	<u>27.479</u>
	Ours	Motion	LTMA	**43.432**	**0.0179**	**50.962**	**20.124**

Table 2. Ablation study on our proposed STME and LTMA modules.

STME	LTMA	FID ↓	KID ↓	FVD ↓	FID-VID ↓
×	×	55.103	0.0412	69.266	30.148
✓	×	53.244	0.0378	67.272	27.341
×	✓	47.432	0.0229	55.962	23.124
✓	✓	**43.432**	**0.0179**	**50.962**	**20.124**

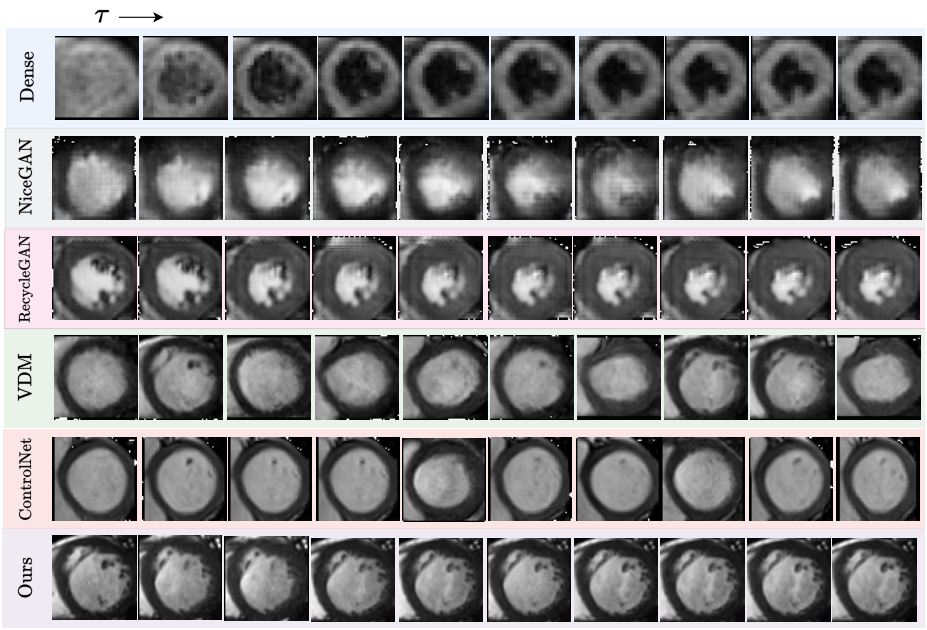

Fig. 3. Qualitative comparison of Cine CMR synthesis. Our approach (bottom row) preserves both spatial and temporal coherence, while VDM struggles with anatomical consistency during motion, and ControlNet shows improved temporal stability but slightly reduced spatial details.

Tab. 2 presents an ablation study evaluating the impact of our LTMA registration network and the STME block. Our experiments show that removing LTMA leads to a decline in the quality of synthesized CMR sequences. This indicates that incorporating LTMA improves our model performance by effectively capturing long-range temporal dependencies in the latent motion space. Meanwhile, adding the STME block further improves the results by extracting hierarchical motion features used as conditioning, thereby improving the quality of the synthesized video samples.

Downstream evaluation on segmentation task. We assess the utility of our synthesized cine CMR sequences by evaluating LV segmentation performance using a 3D U-Net [10] trained only on standard cine CMR. Direct inference on raw DENSE CMR sequences yields poor segmentation performance, with a dice score of 0.31, which indicates poor overlap between the predicted segmentation and the ground truth annotation. In contrast, applying the same model to our synthesized cine CMR significantly improves accuracy, achieving a dice score of 0.81, representing a 70% relative improvement. This highlights the effectiveness of our method in bridging the domain gap and enabling compatibility with existing cine-based clinical models.

5 Conclusion

In this paper, we introduced MFD-V2V, a novel motion feature guided diffusion model for unpaired cardiac video-to-video translation, which enables the synthesis of high-quality cine CMR from low-SNR DENSE CMR sequences. Our proposed method developed a new latent temporal multi-attention registration network to effectively learn accurate cardiac motion from cine CMR videos, followed by a motion-guided diffusion model enhanced with a spatio-temporal motion encoder to improve synthesis quality and fidelity. While we acknowledge that supervision from learned motion fields may introduce bias due to registration errors, the lack of ground truth cine-DENSE alignment makes this a practical compromise. In our experiments, we found that the benefits of using this supervision strategy outweigh the potential drawbacks.

Acknowledgments. This work was supported by NSF CAREER Grant 2239977 and NIH 1R21EB032597.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aletras, A.H., Ding, S., Balaban, R.S., Wen, H.: Dense: displacement encoding with stimulated echoes in cardiac functional mri. Journal of magnetic resonance (San Diego, Calif.: 1997) **137**(1), 247 (1999)
2. Balaji, Y., Min, M.R., Bai, B., Chellappa, R., Graf, H.P.: Conditional gan with discriminative filter generation for text-to-video synthesis. In: IJCAI. vol. 1 (2019)
3. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. IEEE transactions on medical imaging **38**(8), 1788–1800 (2019)
4. Bansal, A., Ma, S., Ramanan, D., Sheikh, Y.: Recycle-gan: Unsupervised video retargeting. In: Proceedings of the European conference on computer vision (ECCV). pp. 119–135 (2018)
5. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. In: International Conference on Learning Representations (2018)
6. Chadalavada, S., Fung, K., Rauseo, E., Lee, A.M., Khanji, M.Y., Amir-Khalili, A., Paiva, J., Naderi, H., Banik, S., Chirvasa, M., et al.: Myocardial strain measured by cardiac magnetic resonance predicts cardiovascular morbidity and death. Journal of the American College of Cardiology **84**, 648–659 (2024)
7. Chen, J., Li, Y., Ma, K., Zheng, Y.: Generative adversarial networks for video-to-video domain adaptation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 3462–3469 (2020)
8. Chen, R., Huang, W., Huang, B., Sun, F., Fang, B.: Reusing discriminators for encoding: Towards unsupervised image-to-image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8168–8177 (2020)

9. Chu, E., Huang, T., Lin, S.Y., Chen, J.C.: Medm: Mediating image diffusion models for video-to-video translation with temporal correspondence guidance. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38 (2024)
10. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention. pp. 424–432 (2016)
11. Ghadimi, S., Bivona, D., Bilchick, K., Epstein, F.: Deep learning-based prognostic model using cine dense mri for outcome prediction after cardiac resynchronization therapy. Journal of Cardiovascular Magnetic Resonance **26** (2024)
12. Gilliam, A.D., Epstein, F.H.: Automated motion estimation for 2-d cine dense mri. IEEE transactions on medical imaging **31**(9), 1669–1681 (2012)
13. 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)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
15. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in Neural Information Processing Systems **35** (2022)
16. Hu, Z., Xu, D.: Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. arXiv preprint (2023)
17. Jabri, A., Fleet, D.J., Chen, T.: Scalable adaptive computation for iterative generation. In: International Conference on Machine Learning. PMLR (2023)
18. Lei, P., Xing, J., Fang, F., Epstein, F.H., Zhang, M.: Dense-guided deep motion networks accounted by large rotations to improve myocardial strain analysis from routine cine mri. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2025)
19. Liang, F., Wu, B., Wang, J., Yu, L., Li, K., Zhao, Y., Misra, I., Huang, J.B., Zhang, P., Vajda, P., et al.: Flowvid: Taming imperfect optical flows for consistent video-to-video synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
20. Metz, L., Poole, B., Pfau, D., Sohl-Dickstein, J.: Unrolled generative adversarial networks. arXiv preprint arXiv:1611.02163 (2016)
21. Reinhardt, J.M., Ding, K., Cao, K., Christensen, G.E., Hoffman, E.A., Bodas, S.V.: Registration-based estimates of local lung tissue expansion compared to xenon ct measures of specific ventilation. Medical image analysis **12**(6), 752–763 (2008)
22. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
23. Sillanmäki, S., Vainio, H.L., Ylä-Herttuala, E., Husso, M., Hedman, M.: Measuring cardiac dyssynchrony with dense (displacement encoding with stimulated echoes)a systematic review. Reviews in Cardiovascular Medicine **24**(9), 261 (2023)
24. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint (2018)
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
26. Vercauteren, T., Pennec, X., Perchant, A., Ayache, N.: Symmetric log-domain diffeomorphic registration: A demons-based approach. In: International conference on medical image computing and computer-assisted intervention. Springer (2008)

27. Wang, Y., Zhang, M., Bilchick, K., Epstein, F.: Transstrainnet: Improved strain analysis of cine mri with long-range spatiotemporal relationship learning. Journal of Cardiovascular Magnetic Resonance **26** (2024)
28. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7623–7633 (2023)
29. Wu, N., Xing, J., Zhang, M.: Tlrn: Temporal latent residual networks for large deformation image registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 728–738. Springer (2024)
30. Xing, J., Ghadimi, S., Abdi, M., Bilchick, K.C., Epstein, F.H., Zhang, M.: Deep networks to automatically detect late-activating regions of the heart. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). IEEE (2021)
31. Xu, Z., Zhang, J., Liew, J.H., Yan, H., Liu, J.W., Zhang, C., Feng, J., Shou, M.Z.: Magicanimate: Temporally consistent human image animation using diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1481–1490 (2024)
32. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
33. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)