This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

OT-Coupled Flow: Towards Realistic TEE Simulation via Unpaired Cross-Modality Translation with Anatomical Consistency

Yunhao Li[1], Yidan Feng[1⋆], Ho-On Alston Conrad Chiu[2], Simon Cheung-Chi Lam[2], Yan Pang[3], Qiong Wang[3], Chun-Ka Wong[2], and Jing Qin[1,4,5]

[1] The Center for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong, China
yunhaol.li@connect.polyu.hk, yidan2.feng@polyu.edu.hk, harry.qin@polyu.edu.hk

[2] Queen Mary Hospital, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China
{alstonchiu,drsimonlam}@gmail.com, wongeck@hku.hk

[3] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
{yan.pang,wangqiong}@siat.ac.cn

[4] CAS SIAT-PolyU Multi-modal Medical Molecular Imaging Joint Laboratory, The Hong Kong Polytechnic University, Hong Kong, China

[5] PolyU-Qianhai Technology and Innovation Research Centre, The Hong Kong Polytechnic University, Hong Kong, China

Abstract. CT-based transesophageal echocardiography (TEE) simulation provides a vital environment for procedural training and autonomous robotic navigation. While recent advances approach realistic simulation via conditional generative models, they typically require strictly paired cross-modal data or costly expert annotations. Acquiring such supervision is practically infeasible due to the intrinsic mismatch between static 3D CT and dynamic operator-dependent 2D TEE, compounded by disparate imaging physics and non-rigid tissue deformations. To address this, we present the first generative TEE simulation engine capable of synthesizing high-fidelity TEE images directly from local CT slices without paired supervision. Our framework operates in a latent space constrained by an adversarial VQ-VAE, pre-trained on a large-scale TEE database to capture the modality-consistent prior. To tackle the central challenge of unpaired translation, we introduce Optimal Transport Coupled Flow Matching (OTC-Flow) as our core technical design, which establishes reliable pseudo-correspondences between CT-conditioned features and target TEE latents, enabling robust flow matching regression. Furthermore, as an indispensible complement, we integrate a pseudo-supervised anatomical consistency loop that leverages robust CT geometry to enforce strict spatial alignment in a completely label-free manner. Extensive experiments demonstrate that our framework significantly outperforms representative unpaired translation models in both fidelity and

⋆ Corresponding author.

anatomical consistency. Code, preprocessing scripts, and split details are available at yunhaoli24/OTC-Flow.

Keywords: Transesophageal Echocardiograph · Cross-Modality Translation · Optimal Transport · Conditional Flow Matching

1 Introduction

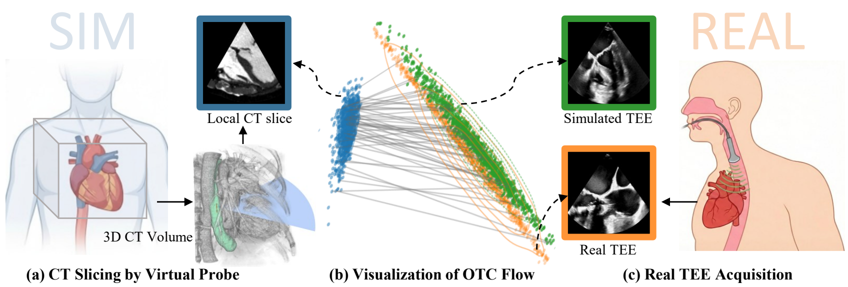

Fig. 1. Illustration of our realistic TEE simulation system. OTC Flow utilizes abundant unpaired real TEE data and guarantees the simulated TEE within the real domain.

While transthoracic echocardiography (TTE) remains the primary cardiovascular screening tool, it is frequently compromised by limited acoustic windows and signal attenuation through the chest wall [12]. In contrast, Transesophageal echocardiography (TEE) is performed from within the esophagus, directly adjacent to the heart, providing high-fidelity visualization of posterior cardiac structures. This anatomical advantage drives its rapidly expanding role in modern cardiology, especially as an indispensable guide for complex, minimally invasive catheter procedures [22,15]. However, blind navigation, complex probe steering, and the risk of esophageal injury makes TEE acquisition technically demanding, creating a critical global shortage of proficient operators [15]. Patient-specific TEE simulation derived from CT has emerged as a scalable solution to this critical bottleneck. Unlike generic phantoms, CT-based simulation provides a diverse, risk-free anatomical environment for procedural planning and operator training [23]. Furthermore, realistic TEE simulation serves as a foundational platform for developing autonomous robotic navigation systems [14,2,3], and bridges the data gap by enabling the transfer of abundant anatomical annotations from CT to the label-scarce TEE domain. Current TEE simulators typically render ultrasound by ray-casting over segmented anatomical labels rather than leveraging the fine-grained tissue textures in the original CT volumes [24,1]. As a result, the simulated images lack the characteristic acoustic texture and visual complexity of real TEE, leading to a domain mismatch with real clinical data that degrades the transfer of downstream models [14,2]. Recent work [3] introduced CycleGAN as a data engine for robotic ultrasound simulation, attempting to bridge the domain gap by training on unpaired real ultrasound data. However, CycleGAN is known to suffer from mode collapse, anatomical hallucinations, and distribution shifts [4]. These limitations have been substantially alleviated by the rapid progress of deep generative modeling, with diffusion probabilistic models [11,26] and flow matching [16,17,6] emerging as the state-of-the-art for high-fidelity synthesis. These advanced architectures have been adapted for conditional ultrasound generation, employing techniques such as ControlNet [7] or conditional flow matching [25] to guide synthesis. Crucially, however, these methods rely on paired masks as condition for real ultrasound frames to explicitly model the joint distribution. This strict supervised requirement poses a severe limitation for CT-based TEE simulation, as expert annotations are exceptionally scarce. More importantly, we aim to leverage the rich anatomical details from CT, yet acquiring perfectly paired CT-TEE data is practically infeasible. As illustrated in Fig. 1, the intrinsic disparity between a one-time 3D CT scan and a dynamic, operator-dependent 2D TEE process, compounded by disparate imaging physics and complex intraoperative deformations, severely hinders accurate cross-modal alignment.

In this paper, we present the first generative TEE simulation engine that produces realistic TEE images directly from local CT slices. We first pre-train an adversarial vector-quantized variational autoencoder (aVQ-VAE) on over 81,510 unlabeled clinical TEE frames to learn a modality-consistent latent prior. Building on this latent representation, we tackle the central challenge of unpaired CT-conditioned synthesis: the model must preserve CT-consistent anatomy while avoiding artefacts from CT-specific structures unobservable in ultrasound, entirely without paired data or TEE masks. We address this with a novel translation framework driven by Optimal Transport Coupled Flow Matching (OTC-Flow). As our core technical contribution, OTC-Flow leverages a global optimal transport coupling to provide pseudo-correspondences between CT-conditioned features and real TEE latent codes. This enables robust conditional flow matching in an unpaired setting, ensuring the generated distribution faithfully matches the real TEE manifold. Building upon this generative foundation, we further introduce a pseudo-supervised anatomical consistency loop as a supplementary structural constraint. Leveraging mature CT analysis tools, this mechanism forces the generated TEE to align with the robustly segmented structures of the source CT, fully utilizing CT's inherent structural information for precise, label-free alignment. Finally, extensive experiments demonstrate that our framework achieves superior fidelity in anatomical consistency and TEE realism compared to representative unpaired translation alternatives, supported by in-depth ablations validating our key design choices.

2 Method

2.1 Framework Overview

As illustrated in Fig. 2, our framework consists of two stages. Stage I pre-trains an adversarial VQ-VAE on unlabeled TEE frames and freezes it to provide a

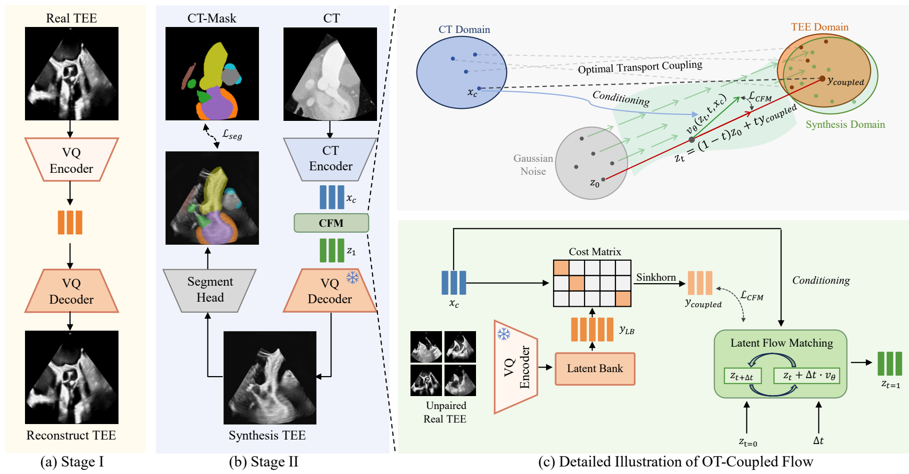

Fig. 2. Framework overview. Stage I: Pre-training of aVQ-VAE. Stage II: Unpaired cross-modality translation with a pseudo-supervised anatomical consistency loop.

modality-consistent latent prior. Stage II tackles the unpaired cross-modality translation from the CT to the TEE domain. To achieve this, we introduce Optimal Transport Coupled Flow Matching (OTC-Flow) as our key technical design to synthesize realistic TEE images from unaligned CT fan slices. During synthesis, a structural encoder E_{ct} extracts a semantic condition $\mathbf{x}_c$ from an input CT fan slice x_{fan}^{CT}. To overcome the challenge of unpaired training, our OTC-Flow mechanism maintains a cross-iteration latent bank storing unpaired real TEE encodings. By employing a global optimal transport coupling, OTC-Flow establishes a pseudo-correspondence between the CT condition $\mathbf{x}_c$ and a target TEE latent code $\mathbf{y}_{\text{coupled}}$. This dynamically matched endpoint guides a time-dependent vector field $v_\theta(\cdot, t, \mathbf{x}_c)$, transporting sampled noise $\mathbf{z}_0 \sim \mathcal{N}(0, \mathbf{I})$ along a trajectory over $t \in [0, 1]$. The terminal latent state $\mathbf{z}_{t=1}$ is quantized and decoded by the frozen aVQ-VAE to yield the synthesized image $\tilde{x}^{TEE}$. To enforce strict spatial alignment alongside this generation process, the framework incorporates a pseudo-supervised anatomical consistency loop. Specifically, an auxiliary segmentation network S processes the synthesized output $\tilde{x}^{TEE}$ to predict semantic anatomical maps $\hat{y}^{TEE}$. These predictions are explicitly supervised by structural pseudo-labels derived directly from the corresponding CT. By providing this explicit anatomical feedback, the consistency loop imposes a strong low-frequency geometric constraint, ensuring that the synthesized TEE strictly preserves subject-specific cardiac shapes without requiring manual ultrasound annotations.

2.2 Optimal Transport Coupling

Standard conditional flow matching learns a time-dependent velocity field by regressing toward targets specified by paired endpoints. However, in unpaired CT-to-TEE synthesis, stochastic endpoint pairing would induce semantically inconsistent targets and inevitably increase the variance of velocity regression. We address this challenge by computing a global optimal transport plan to dynamically assign each CT condition a semantically consistent pseudo-correspondence. Specifically, our method maintains a latent bank (LB) that aggregates empirical TEE latent representations across training iterations. Let $\{\mathbf{x}_c^{(i)}\}_{i=1}^{B}$ and $\{\mathbf{y}_{\mathrm{LB}}^{(j)}\}_{j=1}^{Q}$ denote the condition embeddings extracted from the CT inputs for the current mini-batch and the candidate latent embeddings of TEE retrieved from the LB, respectively. An optimal transport cost matrix $C \in \mathbb{R}^{B\times Q}$ quantifies the pairwise distances between features, establishing a reliable metric for the semantic alignment between these cross-modal representations:

$$C_{ij} = \left\| \mathbf{x}_c^{(i)} - \mathbf{y}_{\mathrm{LB}}^{(j)} \right\|_2^2 . \tag{1}$$

Given the cost matrix, the assignment across the mini-batch is determined by an entropic optimal transport plan that minimizes the aggregated semantic discrepancy. Specifically, the optimal assignment is determined by the entropic optimal transport plan $P^\star \in \mathbb{R}_+^{B\times Q}$, obtained by solving the following objective under uniform marginal constraints $a = \frac{1}{B}\mathbf{1}_B$ and $b = \frac{1}{Q}\mathbf{1}_Q$:

$$P^\star = \arg \min_{P\mathbf{1}_Q = a,\ P^\top \mathbf{1}_B = b} \langle P, C\rangle - \epsilon \sum_{i,j} P_{ij} \left(\log P_{ij} - 1\right). \tag{2}$$

The L2 cost operates in the learned latent space rather than on raw images, providing a target preference between CT-conditioned embeddings and TEE latent-bank candidates. OT then imposes a balanced global assignment across the mini-batch, whereas nearest-neighbor pairing retains only local L2 preference. We solve Eq. 2 via log-domain Sinkhorn-Knopp iterations [5]. While barycentric projection provides a soft interpolation, it can map samples off the empirical support. To preserve support fidelity of TEE latents, we extract a discrete pairing from $P^\star$ by $j^\star(i) = \arg\max_j P^\star_{ij}$ and set $\mathbf{y}_{\mathrm{coupled}}^{(i)} = \mathbf{y}_{\mathrm{LB}}^{(j^\star(i))}$. These matched target endpoints $\{\mathbf{y}_{\mathrm{coupled}}^{(i)}\}_{i=1}^{B}$ subsequently serve as the semantically consistent targets for the flow matching process. Finally, to maintain a continuous and diverse supply of candidate representations, the latent codes of the real TEE images from the current mini-batch are enqueued into the LB. OT coupling and the LB are used only during training to construct stable latent supervision. At inference, synthesis requires the CT encoder, the learned latent transport, and the frozen VQ-VAE decoder.

2.3 Conditional Latent Flow Matching

Driven by the optimal transport coupling, the flow matching component of OTC-Flow provides a deterministic generative path learning a continuous mapping from a simple prior to a target distribution. Unlike diffusion models relying on a stochastic reverse process, Flow Matching directly trains a neural network to approximate a velocity field defining an interpolation path.

Let $\mathbf{y}_{\text{coupled}}$ denote the target endpoint established via our OT pseudo-correspondence, and $\mathbf{x}_c$ the CT-derived condition embedding. With Gaussian noise $\mathbf{z}_0 \sim \mathcal{N}(0, \mathbf{I})$, we define a probability path where $t = 0$ represents the noise state and $t = 1$ the target data. The conditional linear interpolation path:

$$\begin{aligned} \mathbf{z}_t &= (1-t)\mathbf{z}_0 + t\mathbf{y}_{\text{coupled}}, \\ u_t(\mathbf{z}_0, \mathbf{y}_{\text{coupled}}) &= \frac{d\mathbf{z}_t}{dt} = \mathbf{y}_{\text{coupled}} - \mathbf{z}_0, \end{aligned} \tag{3}$$

where $u_t(\mathbf{z}_0, \mathbf{y}_{\text{coupled}})$ represents the instantaneous ground-truth velocity vector that drives the interpolation from the noise $\mathbf{z}_0$ to the paired sample $\mathbf{y}_{\text{coupled}}$. Training involves learning a time-dependent conditional vector field $v_\theta(\mathbf{z}_t, t, \mathbf{x}_c)$ by minimizing the flow matching regression loss:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t,\mathbf{z}_0,\mathbf{y}_{\text{coupled}},\mathbf{x}_c}\left[\left|v_\theta(\mathbf{z}_t, t, \mathbf{x}_c) - (\mathbf{y}_{\text{coupled}} - \mathbf{z}_0)\right|^2\right]. \tag{4}$$

During inference, given a sampled noise $\mathbf{z}_0 \sim \mathcal{N}(0, \mathbf{I})$ and a condition $\mathbf{x}_c$, the target sample is generated by solving the ordinary differential equation:

$$\frac{d\mathbf{z}}{dt} = v_\theta(\mathbf{z}, t, \mathbf{x}_c), \quad \text{with } \mathbf{z}(0) = \mathbf{z}_0, \tag{5}$$

over the interval $t \in [0, 1]$. We employ an Euler solver to discretize $[0, 1]$ into N steps with step size $\Delta t = 1/N$, updating the state as follows:

$$\mathbf{z}_{t+\Delta t} = \mathbf{z}_t + \Delta t \cdot v_\theta(\mathbf{z}_t, t, \mathbf{x}_c). \tag{6}$$

After N steps, the final output $\mathbf{z}_{t=1}$ is obtained as the approximation of $\mathbf{y}_{\text{coupled}}$, guaranteeing that generated features lie robustly on the real TEE manifold.

2.4 Training Objective

The overall loss function involves a composite of $\mathcal{L}_{seg}$ and $\mathcal{L}_{CFM}$ alongside auxiliary regularizers to stabilize synthesis and enhance visual quality. Specifically, $\mathcal{L}_{seg}$ explicitly drives the pseudo-supervised anatomical consistency loop by supervising the anatomical prediction $\hat{y}^{TEE} = S(\tilde{x}^{TEE})$ against the condition-projected pseudo label $\tilde{y}^{TEE}$ via a Dice objective. While the OTC-Flow targets the realistic texture distribution of $\mathbf{y}_{\text{coupled}}$, $\mathcal{L}_{seg}$ imposes a critical geometric constraint. This ensures the synthesized anatomy remains faithful to the CT condition $\mathbf{x}_c$. The conditional flow matching loss $\mathcal{L}_{CFM}$ is computed using endpoints $(\mathbf{z}_0, \mathbf{y}_{\text{coupled}})$ and condition $\mathbf{x}_c$. To mitigate quantization drift and enforce consistency, we employ a latent consistency loss penalizing the discrepancy between the transported latent vector $\mathbf{z}_{t=1}$ and its re-encoded representation

Table 1. Quantitative comparison results with unpaired image translation methods.

Method	Generative Metrics			Perceptual & Anatomical Metrics		
	KID↓	FID↓	FRD↓	LPIPS↓	NIQE↓	Dice↑
CycleGAN [29]	10.74 (+2.35)	110.50 (+31.22)	86.82 (+59.49)	43.65 (+6.18)	10.57 (+3.05)	91.20 (-6.80)
CUT [21]	9.82 (+1.43)	102.99 (+23.71)	52.13 (+24.80)	47.87 (+10.40)	8.19 (+0.67)	93.10 (-4.90)
OTCS [9]	10.43 (+2.04)	93.11 (+13.83)	43.38 (+16.05)	40.12 (+2.65)	8.65 (+1.13)	95.30 (-2.70)
SynDiff [20]	10.55 (+2.16)	93.84 (+14.56)	38.90 (+11.57)	41.23 (+3.76)	7.88 (+0.36)	96.15 (-1.85)
UNSB [27]	8.66 (+0.27)	88.03 (+8.75)	47.28 (+19.95)	39.21 (+1.74)	8.42 (+0.90)	96.50 (-1.50)
OTC-Flow	**8.39**	**79.28**	**27.33**	**37.47**	**7.52**	**98.00**

$\hat{\mathbf{z}}_{t=1} = E_{vq}(\tilde{x}^{TEE})$ from the VQ encoder. Furthermore, a radial power spectral density loss aligns texture statistics by matching frequency-domain profiles of real and synthesized TEE. Finally, a lightweight path-space discriminator provides an adversarial signal to enhance perceptual realism.

3 Experiments

3.1 Datasets

We utilized the Multi-Modality Whole Heart Segmentation (MMWHS) dataset [30] as the source domain, using 20 CT volumes with seven key anatomical labels (LV, LABC, LVBC, RABC, RVBC, AO, and PA) to supervise anatomical consistency. For the target domain, we collected 81,510 real TEE images from 377 patients undergoing LAA closure or MitraClip procedures at two hospitals, acquired via a Philips EPIQ CVx system with an X8-2t probe. These frames were extracted from complete intraoperative videos after de-identification, with long-dwell repeated frames removed. To simulate TEE views, we segmented the esophagus from CT volumes to constrain the probe action space. We then sampled virtual probe poses and depths to extract 60,000 corresponding fan-shaped 2D CT slices. All comparisons and ablations used the same data construction procedure and an 8:2 train/validation split.

3.2 Implementation Details and Evaluation Metrics

Our method is implemented in PyTorch Lightning on two NVIDIA 3090 GPUs using AdamW [18] with an initial learning rate of 1×10^{-4}. Training 20 epochs with a batch size of 32. The CT encoder and segment head use a UNet backbone. The latent velocity network for flow matching comprises CNN blocks, four time-step integration steps, and a 0.1 output scale. For optimal transport, we maintained a 4096-size queue; the Sinkhorn plan was computed over 50 iterations with 0.1 adaptive epsilon scaling. The VQ-VAE encoder used 16,384 codebook embeddings of dimension 8. For assessment, we employ generative metrics to measure the distance between the synthesized and real TEE data distributions. Specifically, we utilize FID [10] to evaluate global realism, KID [8] as an unbiased estimator, and FRD [13] for robust feature-level comparison. Complementarily,

Table 2. Ablation results. NN: Nearest-neighbor. ADM: Adversarial domain matching.

	Latent alignment				Supervision	Metrics					
Variant	FM	OT	NN	ADM	Seg	KID↓	FID↓	FRD↓	LPIPS↓	NIQE↓	Dice↑
OTC-Flow	✓	✓	×	×	✓	8.39	79.28	27.33	37.47	7.52	98.00
w/o OT	✓	×	×	×	✓	11.25	105.62	41.50	45.10	8.35	88.42
NN pair	✓	×	✓	×	✓	9.35	88.90	33.65	40.12	7.81	93.50
ADM	×	—	—	✓	✓	13.40	118.33	45.80	48.90	8.65	82.19
w/o FM	×	—	—	×	✓	9.88	92.15	35.40	42.30	8.05	94.15
w/o Seg	✓	✓	×	×	×	9.10	86.45	48.22	46.55	7.89	-

LPIPS [28] and NIQE [19] are used to assess perceptual fidelity. Dice evaluates CT-conditioned anatomical consistency: CT labels are projected through the same fan grid into pseudo-TEE labels, and segmentations predicted from synthesized TEE images are compared with the projected CT labels.

3.3 Comparison Study

We compared OTC-Flow against representative SOTA conditional generation models capable of unpaired modality translation. Tab. 1 reports quantitative results for TEE synthesis. OTC-Flow achieves top performance across distributional, perceptual and anatomical metrics, substantially better than the strongest baseline UNSB based on Schrödinger Bridge. This validates the effectiveness of our design in achieving robust domain matching and anatomical consistency. This is further supported by visual comparison results Fig. 3. Our method preserves coherent cardiac structures and generates modality-consistent ultrasound textures. Conversely, existing baselines frequently exhibit over-smoothed textures, structural distortions, or inconsistent appearances.

3.4 Ablation study

Results in Tab. 2 validate our design in two key aspects: 1) Latent distribution alignment: Our OT-based coupling (OTC-Flow) significantly outperforms both nearest-neighbor (NN pair) and random pairing (w/o OT) within the flow matching framework. Random pairing removes cost-aware target selection, whereas NN pair preserves local L2 preference without the global marginal constraints of OT; their inferior FRD values support the complementary roles of latent-space target preference and balanced OT assignment. Notably, ineffective latent matching strategies (random pairing or adversarial domain matching at latent space (ADM)) prove detrimental, yielding results inferior to the baseline without flow matching (w/o FM). 2) Anatomical consistency: Removing the segmentation supervision causes a marked degradation in FRD and LPIPS. Since FRD quantifies radiomic discrepancies in texture and morphology, this drop confirms that our pseudo-supervised loop is vital for synthesizing realistic, ultrasound-specific

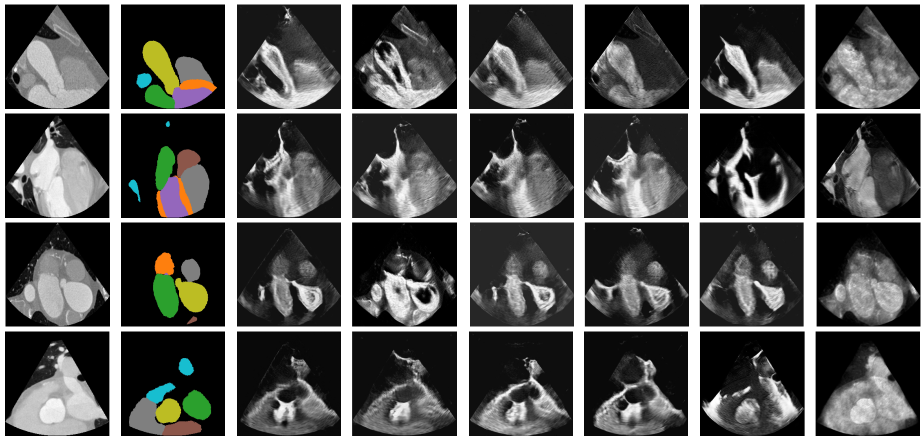

Fig. 3. Qualitative comparison with generative alternatives.

anatomy, while the LPIPS results highlight its role in suppressing generative artifacts.

4 Conclusion

We presented OTC-Flow, the first generative engine for unpaired CT-to-TEE simulation. By integrating optimal transport coupling into flow matching, our framework establishes robust cross-modal correspondences to overcome intrinsic CT-TEE misalignment. Complemented by a pseudo-supervised anatomical consistency loop, it ensures strict structural fidelity leveraging CT geometry. Experiments confirm that OTC-Flow synthesizes realistic ultrasound textures with high anatomical precision, offering a scalable, risk-free platform for procedural training and autonomous robotic navigation.

Acknowledgements This work was supported by a grant under the Hong Kong RGC Collaborative Research Fund (project no. C5055-24G), the Natural Science Foundation of China under Grant U23A20391 & 62576335, in part by the Guangdong Natural Science Foundation under Grant 2025A1515010276, and in part by the Shenzhen Science and Technology Program under Grant JCYJ20241202152803005 & KJZD20240903102717023.

Disclosure of Interests The authors have no competing interests to declare.

References

1. Amadou, A.A., Peralta, L., Dryburgh, P., Klein, P., Petkov, K., Housden, R.J., Singh, V., Liao, R., Kim, Y.H., Ghesu, F.C., et al.: Cardiac ultrasound simulation

for autonomous ultrasound navigation. Frontiers in Cardiovascular Medicine **11**, 1384421 (2024)
2. Amadou, A.A., Singh, V., Ghesu, F.C., Kim, Y.H., Stanciulescu, L., Sai, H.P., Sharma, P., Young, A., Rajani, R., Rhode, K.: Goal-conditioned reinforcement learning for ultrasound navigation guidance. In: International conference on medical image computing and computer-assisted intervention. pp. 319–329. Springer (2024)
3. Ao, Y., Moghani, M., Mittal, M., Prajapat, M., Wu, L., Giraud, F., Carrillo, F., Krause, A., Fürnstahl, P.: Sonogym: High performance simulation for challenging surgical tasks with robotic ultrasound. In: Advances in Neural Information Processing Systems (2025)
4. Cohen, J.P., Luck, M., Honari, S.: Distribution matching losses can hallucinate features in medical image translation. In: International conference on medical image computing and computer-assisted intervention. pp. 529–536. Springer (2018)
5. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013)
6. Dao, Q., Phung, H., Nguyen, B., Tran, A.: Flow matching in latent space. arXiv preprint arXiv:2307.08698 (2023)
7. Freiche, B., El-Khoury, A., Nasiri-Sarvi, A., Hosseini, M.S., Garcia, D., Basarab, A., Boily, M., Rivaz, H.: Ultrasound image generation using latent diffusion models. In: Medical Imaging 2025: Ultrasonic Imaging and Tomography. vol. 13412, pp. 287–292. SPIE (2025)
8. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. The journal of machine learning research **13**(1), 723–773 (2012)
9. Gu, X., Yang, L., Sun, J., Xu, Z.: Optimal transport-guided conditional score-based diffusion model. Advances in Neural Information Processing Systems **36**, 36540–36552 (2023)
10. 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)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
12. Huang, S.J., McLean, A.S.: Appreciating the strengths and weaknesses of transthoracic echocardiography in hemodynamic assessments. Cardiology research and practice **2012**(1), 894308 (2012)
13. Konz, N., Osuala, R., Verma, P., Chen, Y., Gu, H., Dong, H., Chen, Y., Marshall, A., Garrucho, L., Kushibar, K., et al.: Fréchet radiomic distance (frd): A versatile metric for comparing medical imaging datasets. Medical Image Analysis p. 103943 (2026)
14. Li, K., Li, A., Xu, Y., Xiong, H., Meng, M.Q.H.: Rl-tee: Autonomous probe guidance for transesophageal echocardiography based on attention-augmented deep reinforcement learning. IEEE Transactions on Automation Science and Engineering **21**(2), 1526–1538 (2023)
15. Liang, J., Ma, X., Liang, G.: Transesophageal echocardiography: Revolutionizing perioperative cardiac care. Biomolecules and Biomedicine **25**(2), 314 (2024)
16. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2022)
17. Liu, X., Gong, C., et al.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: The Eleventh International Conference on Learning Representations (2022)

18. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations
19. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal processing letters **20**(3), 209–212 (2012)
20. Özbey, M., Dalmaz, O., Dar, S.U., Bedel, H.A., Özturk, Ş., Güngör, A., Cukur, T.: Unsupervised medical image translation with adversarial diffusion models. IEEE Transactions on Medical Imaging **42**(12), 3524–3539 (2023)
21. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: European conference on computer vision. pp. 319–345. Springer (2020)
22. Perk, G., Lang, R.M., Garcia-Fernandez, M.A., Lodato, J., Sugeng, L., Lopez, J., Knight, B.P., Messika-Zeitoun, D., Shah, S., Slater, J., et al.: Use of real time three-dimensional transesophageal echocardiography in intracardiac catheter based interventions. Journal of the American Society of Echocardiography **22**(8), 865–882 (2009)
23. Pezel, T., Dreyfus, J., Mouhat, B., Thébaut, C., Audureau, E., Bernard, A., Badie, Y.L., Bohbot, Y., Fard, D., Nguyen, L.S., et al.: Effectiveness of simulation-based training on transesophageal echocardiography learning: the simulator randomized clinical trial. JAMA cardiology **8**(3), 248–256 (2023)
24. Piórkowski, A., Kempny, A.: The transesophageal echocardiography simulator based on computed tomography images. IEEE Transactions on Biomedical Engineering **60**(2), 292–299 (2012)
25. Reynaud, H., Gomez, A., Leeson, P., Meng, Q., Kainz, B.: Echoflow: A foundation model for cardiac ultrasound image and video generation. arXiv preprint arXiv:2503.22357 (2025)
26. 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)
27. Ye, J.C., Kim, B., KWON, G., et al.: Unpaired image-to-image translation via neural schrödinger bridge. In: The Twelfth International Conference on Learning Representations, ICLR 2024. International Conference on Learning Representations (2024)
28. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
29. 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)
30. Zhuang, X., Li, L., Payer, C., Štern, D., Urschler, M., Heinrich, M.P., Oster, J., Wang, C., Smedby, Ö., Bian, C., et al.: Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. Medical image analysis **58**, 101537 (2019)