Shape-Guided Diffusion Model for Inverse Kinematics of Concentric Tube Robots

Haitao Gao, *Student, IEEE,* Yang Song, *Member, IEEE*, Liao Wu, *Member, IEEE*

***Abstract*—Autonomous minimally invasive surgery (MIS) with concentric tube robots (CTRs) requires fast, reliable inverse kinematics (IK), yet IK for CTRs is challenging due to nonlinear task to joint space mappings, multimodality from trigonometric parameterizations, and kinematic redundancy. We investigate how different task space representations affect learning-based IK for three-tube CTRs and propose a sparse $SE(3)$ shape feature-enhanced diffusion model that characterizes the full distribution of IK solutions. We show that deterministic regression with recorded joint configurations is fundamentally incompatible with the multimodal and redundant nature of the IK problem. Our diffusion-based approach instead approximates the multimodal joint space distribution, enabling sampling of diverse, valid solutions. We further introduce distribution aware metrics quantifying task space accuracy, joint solutions diversity, physical plausibility, and best-of-K quality. On the benchmark dataset, our method gives state-of-the-art, a tip position error of 0.083 ± 0.141 mm, tip orientation error of $0.218^\circ \pm 0.285^\circ$, backbone shape error of 0.122 ± 0.147 mm and orientation error of $0.173^\circ \pm 0.210^\circ$.**

***Index Terms*—Inverse kinematics, Generative models, Concentric tube robots**

I. Introduction

Concentric tube robots (CTRs) are built from nested, pre-curved, super-elastic tubes whose relative rotation and translation produce curvilinear shapes able to follow narrow anatomical pathways that rigid instruments cannot reach. This intrinsic dexterity makes them a leading candidate for robot-assisted minimally invasive surgery (MIS) [1], [2], and progress toward autonomy hinges on solving their inverse kinematics (IK) quickly and reliably. Unlike rigid-link manipulators, CTRs are governed by configuration-dependent continuum mechanics with strong inter-tube elastic coupling, and no closed-form IK exists. Model-based solvers [3], [4] are iterative, sensitive to initialization, and degrade near singular configurations and under hysteresis and snapping properties that are problematic for surgical deployment.

Data-driven IK is an attractive alternative, but it faces a structural obstacle: the IK map for CTRs is inherently redundant and multimodal under trigonometric representation. For three or more tubes, multiple distinct actuation configurations correspond to the same tip pose, and all are valid solutions. Most current learning-based CTR IK approaches use feed-forward neural networks (FFNs) to predict a single ground-truth actuation vector, optimizing it with a mean squared error (MSE) loss regressing with recorded ground-truth [5]–[8], which implicitly collapses the feasible set to its joint-space average generally not itself a feasible configuration. The benchmark study of [8] reports that no satisfactory learning-based IK model was obtained on their proposed public CTR dataset, identifying this as an open problem.

This work was supported by the Australian Research Council under Grant DP250102489.

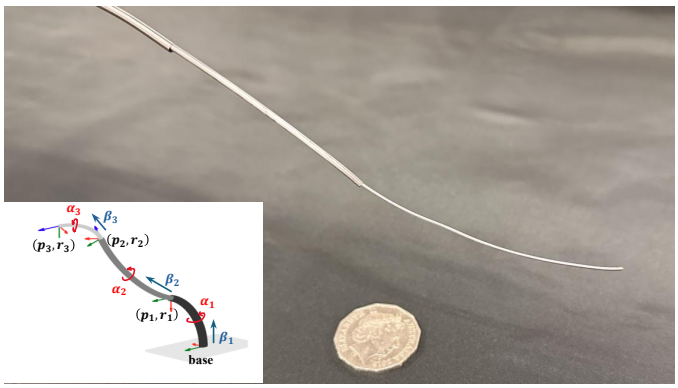

Fig. 1. Three-tube CTR prototype and notation. Joint actuation is the rotation α_i and translation β_i of each tube.

We address this by formulating CTR IK learning as a multimodal generative problem. We propose a conditional diffusion model that denoises both the joint actuation vector $\mathbf{q}$ and a sparse $SE(3)$ features $\boldsymbol{\Xi} \in \mathbb{R}^{18}$ conditioned on target representations $\mathbf{x}$, allowing the model to capture the full distribution of the solution space while embedding backbone shape information into the learning target. Our contributions are:

- A conditional diffusion framework for CTR IK that embeds sparse $SE(3)$ shape features and models IK as a multimodal conditional distribution.
- An analysis of how redundancy and the multimodal trigonometric joint representation break the FFN+MSE paradigm, rendering deterministic regression fundamentally unsuitable for CTR IK.
- A systematic study of four task-space conditioning settings, from tip pose to sparse per-tube observations, showing how geometric information richness shapes the accuracy and diversity of IK solutions.
- Multimodality-aware evaluation combining joint- and task-space metrics under a best-of-K protocol, yielding a more faithful assessment than MSE-based criteria.

II. Methodology

A. Problem Formulation

The raw CTR actuation $\mathbf{q}_{\text{raw}} = [\alpha_1, \beta_1, \alpha_2, \beta_2, \alpha_3, \beta_3]^\top$ encodes per-tube rotations and translations subject to nesting shown in Fig 1. Rotations are embedded as trigonometric pair

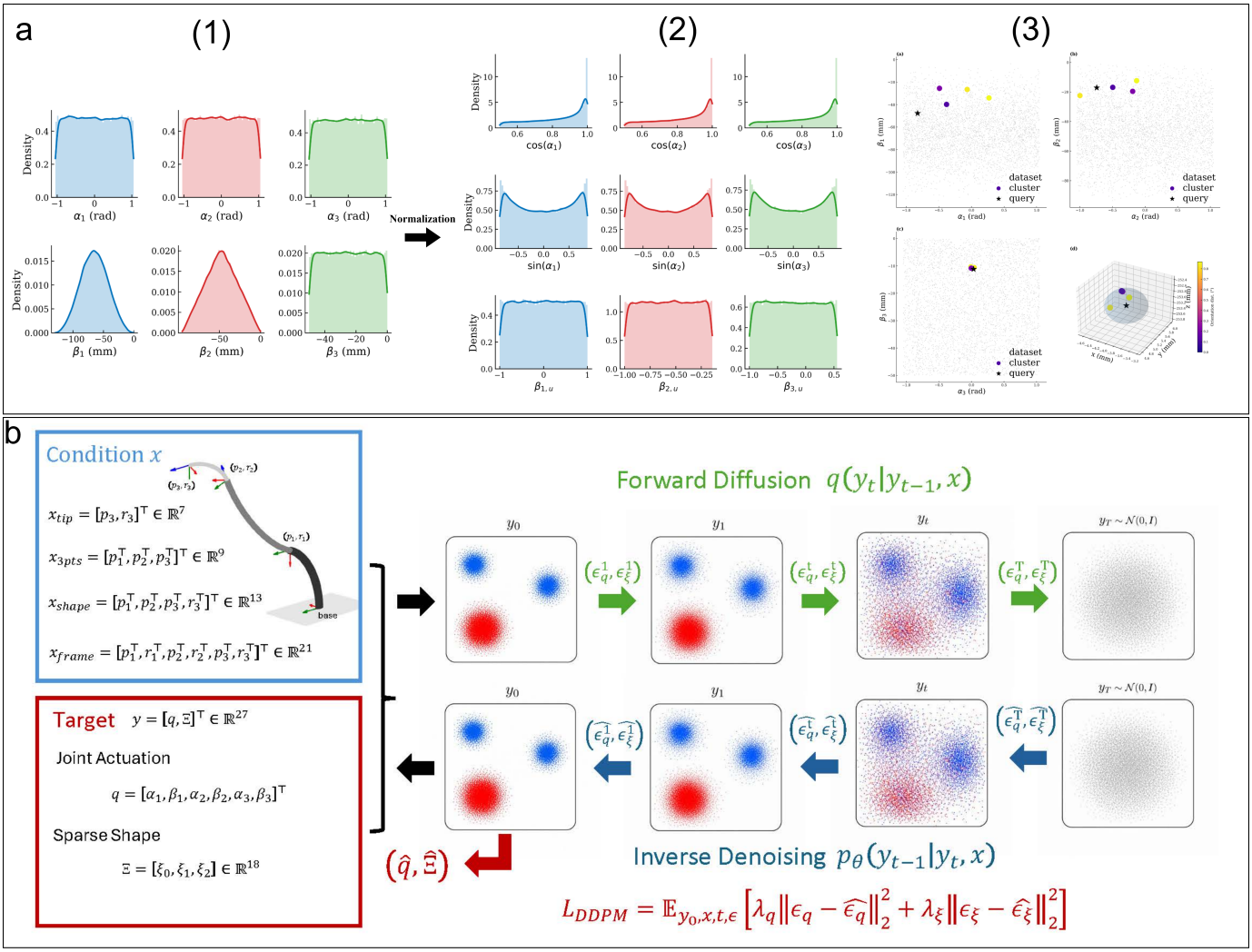

Fig. 2. **(a) Dataset characterisation.** (1) Raw joint-space distributions: rotation angles α_i are near-uniform over $[-1, 1]$ rad; translations β_i follow triangular marginals driven by mechanical coupling constraints. (2) After normalisation, the rotation marginals collapse to cosine-dominated peaks, revealing latent multimodality obscured in raw angle space. (3) Concrete illustration of IK redundancy: multiple distinct joint configurations (purple clusters) map to the same tip pose within a tolerance of ≤ 0.5 mm and $\leq 1^\circ$, confirming that the IK solution set is a one-to-many mapping unsuitable for deterministic regression. **(b) Conditional DDPM for CTR inverse kinematics.** *Left:* four conditioning variants of increasing geometric richness, from tip pose $\mathbf{c}_{\text{tip}} \in \mathbb{R}^7$ to the full sparse $SE(3)$ frame sequence $\mathbf{c}_{\text{frame}} \in \mathbb{R}^{21}$. *Centre:* learning target $\mathbf{y} = [\mathbf{q}^\top, \boldsymbol{\Xi}^\top]^\top \in \mathbb{R}^{27}$, jointly denoising joint actuation $\mathbf{q}$ and sparse shape features $\boldsymbol{\Xi}$. *Right:* forward process $q(\mathbf{y}_t \mid \mathbf{y}_{t-1}, \mathbf{x})$ corrupts $\mathbf{y}_0$ to $\mathcal{N}(\mathbf{0}, \mathbf{I})$; reverse process $p_\theta(\mathbf{y}_{t-1} \mid \mathbf{y}_t, \mathbf{x})$ recovers the full multimodal IK distribution.

$(\cos\alpha_i, \sin\alpha_i)$ and translations normalised to $[-1, 1]$, yielding a unit-less, scale-invariant joint vector $\mathbf{q} \in \mathbb{R}^9$ [9].

Shape information is critical for CTRs; conditioning on tip pose alone neglects backbone geometry. We therefore augment $\mathbf{q}$ with a sparse $SE(3)$ shape feature $\boldsymbol{\Xi} \in \mathbb{R}^{18}$ (Sec. II-B) and learn the conditional distribution $p_\theta(\mathbf{y}|\mathbf{x})$ where $\mathbf{y} = [\mathbf{q}^\top, \boldsymbol{\Xi}^\top]^\top \in \mathbb{R}^{27}$.

B. Sparse $SE(3)$ Shape Features

Motivated by Cosserat rod theory [10], continuous deformation is described by the kinematic ODE $\frac{d\mathbf{T}}{ds} = \mathbf{T}\hat{\boldsymbol{\xi}}$. Since the benchmark [8] provides only three distal sensor frames $\{\mathbf{T}_i\}_{i=1}^3 \subset SE(3)$, we adopt a **discrete** analogue. Frames are expressed relative to the base,

$$\hat{\mathbf{T}}_0 = \mathbf{I}, \quad \hat{\mathbf{T}}_1 = \mathbf{T}_0^{-1}\mathbf{T}_1, \quad \hat{\mathbf{T}}_2 = \mathbf{T}_0^{-1}\mathbf{T}_2, \quad \hat{\mathbf{T}}_3 = \mathbf{T}_0^{-1}\mathbf{T}_3. \tag{1}$$

and each consecutive pair yields a relative twist via the matrix logarithm:

$$\boldsymbol{\xi}_i = \log\left(\hat{\mathbf{T}}_i^{-1}\hat{\mathbf{T}}_{i+1}\right) \in \mathbb{R}^6, \quad i = 0, 1, 2. \tag{2}$$

Stacking these gives $\boldsymbol{\Xi} = [\boldsymbol{\xi}_0^\top, \boldsymbol{\xi}_1^\top, \boldsymbol{\xi}_2^\top]^\top \in \mathbb{R}^{18}$ indicating the coarse backbone information. At inference, the predicted $\hat{\boldsymbol{\Xi}}$ is integrated via $\hat{\mathbf{T}}_{i+1} = \hat{\mathbf{T}}_i \exp(\widehat{\hat{\boldsymbol{\xi}}_i})$ to recover the coarse backbone shape.

C. Denoising Diffusion Probabilistic Models

We model $p_\theta(\mathbf{y}|\mathbf{x})$ with a conditional denoising diffusion probabilistic models (DDPMs) [11]. Fig 2 (b) shows the full model architecture in which the forward process corrupts $\mathbf{y}_0$ with Gaussian noise ϵ under schedule $\{\bar{\alpha}_t\}$, and a denoising network $\boldsymbol{\epsilon}_\theta(\mathbf{y}_t, t, \mathbf{c})$ is predicting the noise introduced by

TABLE I
IK EVALUATION RESULTS (BEST-K, K=32). BEST RESULTS HIGHLIGHTED IN RED. LITERATURE VALUES REPORTED UNDER DIFFERENT EXPERIMENTAL SETUPS, SHOWN FOR REFERENCE ONLY.

Method	**Group**	**Tip Pos.** ↓ (mm)	**Tip Ori.** ↓ (°)	**Shape Pos.** ↓ (mm)	**Shape Ori.** ↓ (°)	α **Err.** ↓ (°)	β **Err.** ↓ (mm)
SE(3)-DDPM Frames	Ours	0.083 ± 0.141	0.218 ± 0.285	0.122 ± 0.147	0.173 ± 0.210	4.435 ± 4.005	0.771 ± 0.613
SE(3)-DDPM Shape	Ours	0.108 ± 0.094	0.291 ± 0.217	0.136 ± 0.100	1.359 ± 1.101	6.498 ± 7.324	0.742 ± 0.411
SE(3)-DDPM 3Pts	Ours	0.113 ± 0.078	0.976 ± 1.598	0.142 ± 0.080	1.589 ± 1.294	6.591 ± 7.003	0.670 ± 0.344
SE(3)-DDPM Tip	Ours	0.213 ± 0.132	0.311 ± 0.228	1.649 ± 0.946	4.623 ± 2.719	28.943 ± 22.459	2.920 ± 1.984
IK-MSE	Baseline	16.425 ± 6.150	27.912 ± 15.800	23.972 ± 8.551	–	57.364 ± 16.586	28.916 ± 14.838
IK-Cycle	Baseline	10.109 ± 6.595	17.927 ± 12.493	21.303 ± 9.059	–	58.729 ± 23.147	32.205 ± 17.596
DDPM-Tip	Baseline	0.267 ± 0.142	0.877 ± 1.263	1.754 ± 0.884	–	26.668 ± 21.714	3.132 ± 2.001
DDPM-3Pts	Baseline	0.224 ± 0.130	1.029 ± 1.630	0.417 ± 0.152	–	7.277 ± 6.922	0.588 ± 0.275
DDPM-Shape	Baseline	0.247 ± 0.168	0.885 ± 1.185	0.435 ± 0.213	–	6.954 ± 7.032	0.579 ± 0.309

forward process and trained with a weighted loss over the heterogeneous target:

$$\mathcal{L} = \mathbb{E}\big[\lambda_q \|\boldsymbol{\epsilon}_q - \hat{\boldsymbol{\epsilon}}_q\|_2^2 + \lambda_\xi \|\boldsymbol{\epsilon}_\xi - \hat{\boldsymbol{\epsilon}}_\xi\|_2^2\big]. \tag{3}$$

At inference we draw K candidates and select the best by minimising $\mathcal{S} = e_{\text{tip}} + e_{\text{shape}} + \lambda_R e_R$. Both single-sample and best-of-K results are reported.

Task-space conditions. To study the effect of geometric richness, we evaluate four conditioning signals of increasing information content:

$$\mathbf{x}_{\text{tip}} = [\mathbf{p}_{\text{tip}}^\top, \mathbf{r}_{\text{tip}}^\top]^\top \in \mathbb{R}^7, \quad \mathbf{x}_{\text{points}} = [\mathbf{p}_1^\top, \mathbf{p}_2^\top, \mathbf{p}_3^\top]^\top \in \mathbb{R}^9, \tag{4}$$

$$\mathbf{x}_{\text{shape}} = [\mathbf{p}_{1:3}^\top, \mathbf{r}_{\text{tip}}^\top]^\top \in \mathbb{R}^{13}, \quad \mathbf{x}_{\text{frames}} = [\mathbf{p}_i^\top, \mathbf{r}_i^\top]_{i=1}^3 \in \mathbb{R}^{21}. \tag{5}$$

D. Baselines

We compare against deterministic regressors (*IK-MSE*, *IK-Cycle*) and ablated diffusion models that predict $\hat{\mathbf{q}}$ only (DDPM-Tip, DDPM-3Pts, DDPM-Shape), isolating the contribution of shape reconstruction.

E. Evaluation Metrics

In joint space we report the aggregate angular error $e_\alpha = \big(\sum_{i=1}^3 e_{\alpha,i}^2\big)^{1/2}$, where $e_{\alpha,i}$ is the wrapped per-tube angle difference recovered via atan2 from the sine–cosine representation, and the translation error $e_\beta = \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2$ in millimetres. In task space we report the tip position error $e_{\text{tip}} = \|\hat{\mathbf{p}}_3 - \mathbf{p}_3\|_2$, the sparse shape error $e_{\text{shape}} = \frac{1}{3}\sum_{i=1}^3 \|\hat{\mathbf{p}}_i - \mathbf{p}_i\|_2$, and the geodesic orientation error $e_R = \|\log(\mathbf{R}^\top \hat{\mathbf{R}})\|_2$ on $SO(3)$ for the tip and sparse shape frames.

III. EXPERIMENTS

A. Dataset and Setup

We evaluate on *CRL-Dataset-CTR-Pose* [8], the first open benchmark from a physical three-tube CTR, containing 100,000 samples of joint values $(\alpha_i, \beta_i)_{i=1}^3$ paired with four $SE(3)$ sensor frames per configuration. Following [8], we split into 80,000 / 10,000 / 10,000 train/val/test samples. Sensor noise is 0.7 mm RMS (position) and 0.20° RMS (orientation).

All DDPM variants are trained for 1,000 epochs (AdamW, lr $= 2 \times 10^{-4}$, batch size 1,024) with a squared cosine noise schedule over T=1,000 steps and 100 inference steps. The $SE(3)$-DDPM uses hidden size 384, depth 5; ablation baselines use hidden size 256, depth 4.

B. Kinematic Redundancy and Multimodality

A three-tube CTR with 6-DoF nominally provides a unique IK solution under full pose constraints. Empirically, however, we observe residual redundancy in the benchmark: at a 0.5 mm / 1.0° tolerance, 50.4% of samples have at least two distinct configurations mapping to the same pose, with mean cluster size 1.80. A concrete example is shown in Fig 2 (a.(3)). We attribute this to inter-tube coupling induced by pre-curvature, which effectively reduces the independent DoF of the robot. Additionally, encoding rotations as $(\cos\alpha_i, \sin\alpha_i)$ induces a bimodal angular distribution seen in Fig 2 (a.(2)) that MSE-based regressors cannot represent. This motivates the generative formulation.

C. Results

FFN mode collapse. A feedforward network trained with MSE loss (IK-MSE) collapses to a near-constant prediction approximating the dataset mean, yielding tip errors exceeding 16 mm (Table I, Fig. 3). This is a direct consequence of averaging over a multimodal, redundant IK distribution. Replacing the joint-space loss with a cycle-consistency term through a **pretrained** FK network (IK-Cycle) substantially recovers distributional coverage, reducing tip error to 10.1 mm, yet the FFN remains unable to represent the full multimodal structure and still collapses bimodal angular marginals to unimodal peaks.

Diffusion models recover the full distribution. DDPMs accurately approximate the joint distribution regardless of conditioning modality (Fig. 3), with all variants closely tracking the ground-truth marginals including bimodal rotations and plateau-shaped translations. The strongest purely positional baseline, DDPM-3Pts (Best-K), achieves 0.224 mm tip position and 1.029° tip orientation, reductions of over 97% relative to IK-MSE.

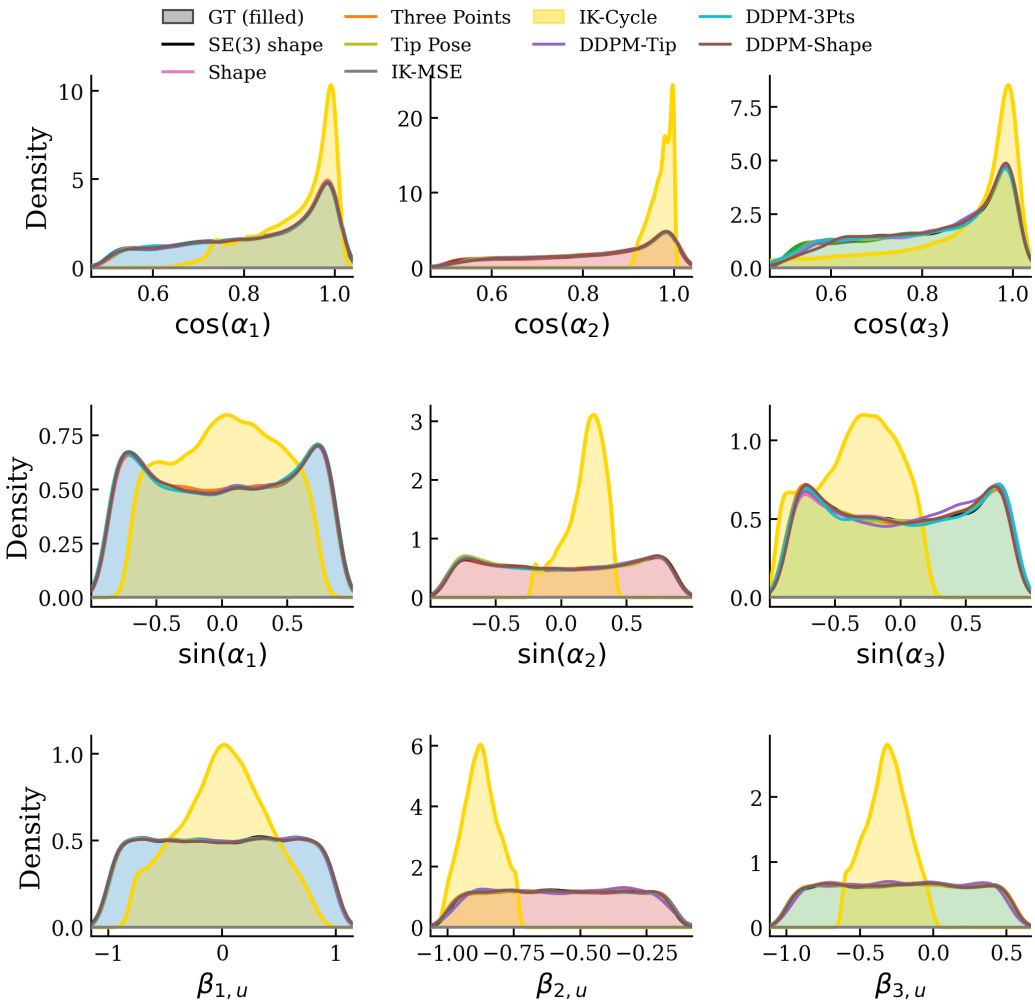

Fig. 3. Marginal joint-space distributions of FFN-based and DDPMs on the test split (Best-K=32)

$SE(3)$ **shape augmentation.** Augmenting the diffusion target with Ξ consistently improves every task-space metric, with gains scaling with the geometric richness of the conditioning input. Under three-point and shape conditioning, tip-position error is approximately halved and sparse shape position error reduced around $3\times$ relative to the corresponding DDPM baselines. The full-frame variant (Best-K, K=32) achieves $\mathbf{0.083 \pm 0.141}$ mm tip position and $\mathbf{0.218^\circ \pm 0.285^\circ}$ tip orientation—reductions of 69% and 75% over the strongest DDPM baseline and over two orders of magnitude over IK-MSE. Physical constraint satisfaction remains $\geq$98.1% across all diffusion variants without any explicit feasibility terms in training.

Inference speed. One-shot sampling with 100 denoising steps runs at around 719 Hz (1.39 ms latency), exceeding EM tracking (40 Hz), FBG sensing (100 Hz), and real-time shape estimation (333 Hz) update rates reported in CTR literature [12]–[14]. While this falls below 1 kHz servo loops, the diffusion model targets high-level IK generation rather than low-level actuation, fitting naturally into a hierarchical control architecture.

IV. Discussion and Conclusion

CTR IK is inherently multimodal: deterministic regression averages over valid modes, producing configurations that are neither physically meaningful nor task-accurate. Framing IK as a conditional generative problem lets the method approximate $p_\theta(\mathbf{y}|\mathbf{c})$ and return multiple feasible solutions per query. Sparse $SE(3)$ target features embed backbone geometry into the diffusion target, serving as a shape-aware regulariser whose benefit grows with the geometric richness of the conditioning input. Consequently, single-sample MSE against a recorded $\mathbf{q}$ is not an adequate metric for redundant IK; distribution coverage, FK consistency, constraint satisfaction, and best-of-K accuracy are required.

For surgical autonomy, the method produces accurate, physically feasible IK candidates at about 719 Hz, suitable as a high-level planner in hierarchical control. Limitations include evaluation on a single free-space prototype and the added sampling cost of best-of-K inference. Future work will explore few-step diffusion samplers, closed-loop FK feedback, uncertainty-aware mode selection, and safety-constrained planning for minimally invasive surgery deployment.

REFERENCES

[1] H. Alfalahi, F. Renda, and C. Stefanini, "Concentric tube robots for minimally invasive surgery: Current applications and future opportunities," *IEEE Transactions on Medical Robotics and Bionics*, vol. 2, no. 3, pp. 410–424, 2020.

[2] K. Leibrandt, C. Bergeles, and G.-Z. Yang, "Concentric tube robots: Rapid, stable path-planning and guidance for surgical use," *IEEE Robotics & Automation Magazine*, vol. 24, no. 2, pp. 42–53, 2017.

[3] P. E. Dupont, J. Lock, B. Itkowitz, and E. Butler, "Design and control of concentric-tube robots," *IEEE Transactions on Robotics*, vol. 26, no. 2, pp. 209–225, 2009.

[4] D. C. Rucker, B. A. Jones, and R. J. Webster III, "A geometrically exact model for externally loaded concentric-tube continuum robots," *IEEE transactions on robotics*, vol. 26, no. 5, pp. 769–780, 2010.

[5] R. Grassmann, V. Modes, and J. Burgner-Kahrs, "Learning the forward and inverse kinematics of a 6-dof concentric tube continuum robot in se (3)," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 5125–5132.

[6] N. Liang, R. M. Grassmann, S. Lilge, and J. Burgner-Kahrs, "Learning-based inverse kinematics from shape as input for concentric tube continuum robots," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 1387–1393.

[7] P. H. Kang, R. Gondokaryono, M. Roshanfar, R. H. Nguyen, T. Looi, J. M. Drake, and D. Podolsky, "Learning inverse kinematics multiplicity of concentric tube robots using invertible neural networks," in *2025 International Symposium on Medical Robotics (ISMR)*. IEEE, 2025, pp. 157–163.

[8] R. M. Grassmann, R. Z. Chen, N. Liang, and J. Burgner-Kahrs, "A dataset and benchmark for learning the kinematics of concentric tube continuum robots," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 9550–9557.

[9] R. Grassmann and J. Burgner-Kahrs, "On the merits of joint space and orientation representations in learning the forward kinematics in se (3)." in *Robotics: science and systems*, 2019.

[10] S. S. Antman, *Nonlinear problems of elasticity*. Springer, 2005.

[11] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[12] Q. Boyer, S. Voros, P. Roux, F. Marionnet, K. Rabenorosoa, and M. T. Chikhaoui, "On high performance control of concentric tube continuum robots through parsimonious calibration," *IEEE Robotics and Automation Letters*, vol. 9, no. 11, pp. 9621–9628, 2024.

[13] R. Xu, A. Yurkewich, and R. V. Patel, "Curvature, torsion, and force sensing in continuum robots using helically wrapped fbg sensors," *IEEE Robotics and Automation Letters*, vol. 1, no. 2, pp. 1052–1059, 2016.

[14] H. Donat, J. Gu, and J. J. Steil, "Real-time shape estimation for concentric tube continuum robots with a single force/torque sensor," *Frontiers in Robotics and AI*, vol. 8, p. 734033, 2021.