Learning Sim-Grounded Policies for Bimanual Rope Manipulation from Human Teleoperation Data

Fig. 1: Open-loop rollout of the particle-based ACT policy on an overhand-knotted rope excluded from the training set where t indexes simulation frames at 30 Hz. The policy predicts a single macro action (grasp and pull) from the initial particle state at t=0; for overhand knot configurations like the one shown, a single well-placed pull is sufficient to fully resolve the knot. The policy successfully transfers this strategy to some unseen rope instances.

***Abstract*—Deformable Linear Objects (DLOs) such as ropes and cables are widely encountered in both household and industrial applications, yet remain challenging to manipulate due to their infinite-dimensional configuration space and frequent self-occlusion. Imitation learning from teleoperation offers a practical path to bimanual DLO manipulation, but its scalability is limited by human effort, making the choice of observation space critical for generalization from small datasets. In this study, we investigate whether the lack of generalization in egocentric visual policies for the knot-untangling task stems from the observation space itself, rather than from the policy architecture or data scale. We compare two Action Chunking with Transformers policies trained on the same bimanual teleoperation data: a vision-based policy conditioned on two egocentric RGB streams from wrist-mounted cameras, and a state-based policy conditioned on the DLO's 3D particle state, extracted from an initial observation via multi-view fusion and evolved in a particle-based eXtended Position-Based Dynamics simulation. Evaluated open-loop on an unseen rope configuration, the state-based policy outperforms its visual counterpart with a 30.8% reduction in L1 error when predicting the initial grasp-and-pull action, quantifying the observability gap between pixels and physics-consistent state, and pointing toward more data-efficient robot learning for the DLO manipulation task from limited human demonstrations.**

I. Introduction

Tasks involving the manipulation of Deformable Linear Objects (DLOs), such as untangling ropes, remain one of the most challenging tasks in robotics due to high-dimensional state spaces and frequent self-occlusions [1], [2]. Traditional imitation learning approaches often attempt to map continuous RGB video streams from wrist-mounted cameras directly to robot actions [3], [4]. However, this "end-to-end" visual paradigm suffers from two major bottlenecks: computational inefficiency due to high-bandwidth data processing [5], and perceptual vulnerability when the robot's own end-effectors (EEFs) occlude the object of interest [6].

In this work, we propose a shift from continuous visual tracking to a *Snapshot-to-Action* paradigm grounded in physics-based simulation. Instead of requiring a persistent video stream, our method extracts the rope's topological state once, from a multi-view RGB-D observation fused across wrist-mounted cameras. This state is then mapped into a particle-based eXtended Position-Based Dynamics (XPBD) simulation using a Cosserat rod model [7]–[10]. This approach offers a dual advantage: it replaces high-dimensional visual inputs with a compact state representation, reducing compute and memory overhead and, more importantly, provides a physically consistent state representation that persists even when raw visual tracking fails.

We argue that for bimanual manipulation using end-effectors with wrist-mounted cameras, simulation is not merely an optional data source but a necessary state estimator. During manipulation, the robot arms inevitably occlude large segments of the rope, causing standard tracking algorithms to lose the trace of the rope [11]. By grounding the initial observation in an XPBD-based "Digital Twin", we create a representation anchor that maintains physical integrity where pixels provide no signal.

We utilize Action Chunking with Transformers (ACT) to learn from human teleoperation demonstrations, collected without predefined action scripts or motion templates [3]. By focusing on the initial manipulation action, we benchmark the policy's ability to interpret complex knots from a single observation state.

To isolate the role of the observation space itself from confounders such as policy architecture or training data volume, we frame our evaluation as a controlled ablation: two policies with identical ACT backbones and identical training trajectories, differing only in whether they condition on raw egocentric pixels or on a physics-consistent particle state. We perform zero-shot evaluation on a different rope that does not appear in the training dataset, examining whether the observation space by itself dictates how effectively a policy generalizes across rope instances. Recent work on 3D representations has answered this question positively for rigid and articulated

objects, but it remains unresolved for DLOs [12]–[14].

Our main contributions and arguments are:

- **Occlusion-Robust State Grounding:** We demonstrate that mapping a multi-view observation into a particle-based simulation provides a more robust representation for DLOs than raw pixels, specifically addressing the failure modes of visual tracking during bimanual occlusion.
- **Observational Efficiency:** We show that a single multi-view state extraction, evolved in simulation, is sufficient for predicting complex human-like intent without continuous visual feedback, leading to reduced dependency on high-frequency visual feedback. We further show that inference time is reduced by 86.2% and peak VRAM usage is reduced by 46.4% using our data-efficient representation, enabling our approach to run on resource-constrained edge systems.

Through this controlled comparison, we quantify the observability gap between pixel-level observations and physics-consistent states in bimanual knot untangling.

II. Related Work

Deformable Linear Object Manipulation. DLO manipulation is challenging due to infinite-dimensional configuration spaces, visual homogeneity, and frequent self-occlusion [15]–[17]. Recent work approaches rope untangling through topological state estimation and discrete action primitives. Grannen et al. [1] predict keypoints for pull-and-pin primitives directly from RGB observations without reconstructing the full cable state. Sundaresan et al. [2] refine this pipeline and extend it to non-planar knots by adding learned components for grasp refinement and recovery from failure modes, while other work generalizes the approach to dense multi-cable knots by planning over predicted topological graphs [18]. Recent assistive teleoperation work [19] instead keeps the human in the loop as the planner. These methods share a common structure, where either perception produces a discrete state or a keypoint set that a planner executes using fixed motion primitives, or the human supplies topological reasoning at runtime. We instead learn continuous bimanual trajectories end-to-end from diverse limited human demonstrations, using a particle-based representation that captures finer geometric nuances than purely topological models and sidesteps the need to enumerate primitives at planning time.

3D vs. 2D Observation Spaces for Imitation Learning. Recent work shows that structured 3D representations outperform raw 2D images when the training data is small. DP3 shows that diffusion policies conditioned on sparse point clouds generalize across space, viewpoint, appearance, and instance variations where equivalent RGB policies fail, requiring only 10 to 40 demonstrations per task [12]. iDP3 extends this to egocentric 3D representations for humanoid manipulation and empirically demonstrates that image-based counterparts overfit to the training scenario [13]. GROOT makes a related argument with object-centric 3D priors, showing that end-to-end visual policies are brittle to background, viewpoint, and instance changes that structured representations absorb [14]. These results are established primarily for rigid and articulated objects. We extend this approach to DLOs, where the relevant 3D structure is not a set of discrete entities but an ordered one-dimensional manifold, and where the gripper is the dominant source of occlusion.

Imitation Learning with ACT. Imitation learning for bimanual tasks has been significantly advanced by ACT [3], which uses a Conditional Variational Autoencoder (CVAE) Transformer architecture to predict synchronized action chunks from multi-view RGB observations. ACT and related methods usually rely on raw visual streams throughout the trajectory, making the policy's effectiveness dependent on the quality and comprehensiveness of the visual observations during execution. We extend this framework to a *Snapshot-to-Action* paradigm by grounding the ACT policy in an XPBD-based particle state initialized from a multi-view RGB-D observation, ensuring spatial equivariance and robustness to the occlusions inherent in bimanual manipulation.

III. Methodology

Our objective is to learn a policy π that predicts a human-like manipulation action a given an initial state s of a tangled rope, comparing raw visual input and sim-grounded particle coordinates generated through physical re-execution.

A. Data Acquisition and Bimanual Setup

We utilize a dataset of human teleoperation demonstrations for the task of bimanual overhand-knot untangling, performed across four distinct ropes varying in length, color, and stiffness. The dataset comprises 96 demonstrations in which the human provides diverse, unconstrained manipulation strategies. Data was collected via a bimanual teleoperation interface, recording gripper poses $(x, y, z, q_x, q_y, q_z, q_w)$ and continuous gripper openness at 100 Hz, resampled to 30 Hz to match the simulation rate. Rather than filtering for "perfect" trajectories, we explicitly include sub-optimal demonstrations, treating the data as a diverse distribution of human topological reasoning that captures failed attempts and exploratory "probing" actions. Thus, we do not aim to train policies that fully untangle ropes. Instead, we focus on capturing human intent by predicting only the first macro action.

B. Ground-Truth Generation via Simulation Playback

To collect our dataset of particle states from teleoperation data, we employ a Simulation Playback pipeline:

Snapshot Initialization: Because the rope is heavily occluded by the robot's EEFs during manipulation, we temporarily extract multi-view RGB-D frames before the gripper initiates its grasp. Using the Segment Anything Model 2 (SAM2) [20], we generate binary masks of the rope and combine them with aligned depth maps, aggregating the views into a unified, voxel-downsampled 3D point cloud. To extract the rope's one-dimensional topology, we project this point cloud onto the 2D table plane and apply a Voronoi-based skeletonization to estimate its centerline [19], [20]. We then lift these keypoints

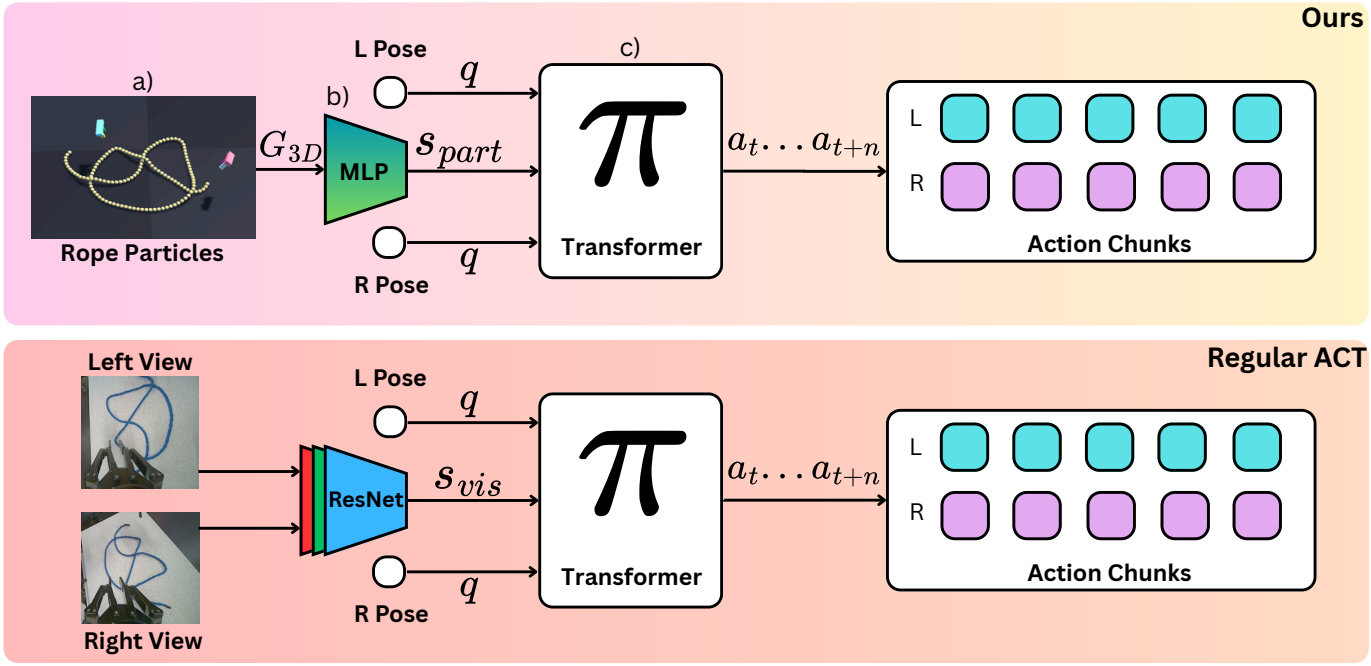

Fig. 2: Comparative overview of our architecture and the regular ACT policy we are comparing against. a) Rope particle positions are extracted from an initial observation of the rope. b) The particle positions are encoded into a state vector through an MLP. c) A transformer model acts as the policy and predicts the next chunk of EEF movements.

back into a 3D ordered structure, resolving ambiguous self-intersections by evaluating the relative depth gradient ∇Z at crossing nodes. To standardize input dimensionality, we uniformly resample this centerline to a fixed sequence of $N{=}100$ particles via arc-length interpolation, producing the particle state $G_{3D} \in \mathbb{R}^{100\times 3}$. This canonicalizes raw particle counts that ranged from 70 to 150 across rope instances, so each index corresponds to a consistent fractional position along the arc length.

Action Re-execution and Labeling: We spawn a simulated gripper in the XPBD environment, execute the human's recorded teleoperation actions $a_{\text{human}}(t)$, and record the resulting particle trajectory $\{G_{3D}(t)\}_{t=0}^{T}$ synchronized with those actions. This yields a continuous, fully observed state signal throughout the demonstration, including phases in which the gripper heavily occludes the rope and visual tracking would fail, producing the $(\text{state}, \text{action})$ pairs used to train the particle-based policy.

C. Policy Learning with Action Chunking

We build upon the ACT framework [3]. ACT models the policy π as a CVAE. During training, a transformer encoder compresses the ground-truth action sequence into a latent style variable z, while a transformer decoder conditions on z, the current observation, and learned action queries to predict a chunk of k future bimanual actions. At inference time, z is sampled from the prior, and overlapping chunks are smoothed via temporal ensembling.

We define two input variants for the observation s, adapting the encoder accordingly while sharing the decoder architecture, action head, and CVAE latent structure. Both variants additionally receive the current bimanual proprioception $q \in \mathbb{R}^{16}$ as an embedded input token. For each arm, this consists of an 8-dimensional state: the end-effector pose (3D position and 4D quaternion) and a 1D scalar for gripper openness.

Vision-based State (s_{vis}): Multi-view RGB images from both wrist-mounted cameras serve as input. A shared, pre-trained ResNet-18 backbone independently extracts spatial feature maps from each view; these are projected to the transformer's hidden dimension ($d{=}512$) via a $1{\times}1$ convolution, concatenated along the spatial axis, and augmented with sinusoidal 2D position encodings.

Particle-based State (s_{part}): To replace high-dimensional image processing with a structurally compact input, we introduce a custom ParticleEncoder that bypasses the ResNet backbone. The $N{=}100$ resampled 3D particle coordinates from the simulation playback are projected through a two-layer MLP with GELU activation and LayerNorm, producing one token per particle. Sinusoidal 1D positional encodings are added to preserve the sequential ordering along the rope's arc length.

Figure 2 shows both architectures side by side. Apart from the encoder, both variants share the same backbone and are trained with the standard ACT loss combining L1 action reconstruction and KL regularization.

IV. Experiments

The experimental setup consists of two 7-DoF Franka Panda arms with Robotiq 2F-140 grippers and wrist-mounted ZEDX

stereo cameras. Demonstrations are collected via VR teleoperation (HTC Vive Pro) and provide 6-DoF EEF commands using VR controllers. The 96 demonstrations are split into 64 training, 15 validation, and 17 test samples. The test set consists exclusively of demonstrations on the held-out rope, the stiffest of the four ropes and thus the most dynamically distinct from the training distribution. This evaluates zero-shot generalization to an unseen rope instance with unseen physical properties.

We quantitatively evaluate both policies through an offline comparison, using human teleoperation trajectories from the held-out test set as ground truth. For our particle-based policy, we extract an initial multi-view RGB-D snapshot to initialize the rope's geometry in simulation. To continuously provide the policy with updated particle state information, we propagate the simulation forward using Newton Physics, which employs an XPBD solver based on the Cosserat rod model [8], [21]. Our implementation builds upon the rope simulation by Cai et al. and incorporates additional physical constraints from [10], [22]. In contrast, the baseline vision policy processes continuous multi-view RGB observations directly at each evaluation timestep. Based on their respective state representations, both policies predict open-loop action chunks. We evaluate their ability to replicate human manipulation intent on unseen rope configurations using the L1 error, defined as the mean absolute distance between predicted and ground-truth bimanual EEF poses at the end of a k=20-step trajectory. Table I reports these results alongside computational metrics evaluated on an NVIDIA RTX 4060 Ti GPU.

TABLE I: **Action prediction performance metrics.** Inference time is measured per single step, and Peak VRAM is the maximum PyTorch memory allocated during evaluation.

Policy	**L1** $(10^{-3})\downarrow$	**Inf. Time (ms)** $\downarrow$	**Peak VRAM (MB)** $\downarrow$
Particles (S)	$\mathbf{9.70 \pm 7.95}$	$\mathbf{2.95 \pm 0.05}$	**334.40**
Vision (V)	14.02 ± 8.44	21.34 ± 0.21	623.66

To complement the quantitative evaluation, Figure 1 shows an open-loop rollout of the particle-based policy on the held-out rope, which differs from the training instances in both appearance and stiffness. From the initial particle state alone, the policy predicts a grasp-and-pull action that resolves the overhand knot in a single macro step, transferring zero-shot to a rope instance with physical properties unseen during training. While this level of success is not achieved uniformly across all test samples, the rollout qualitatively demonstrates that the particle representation can encode sufficient information about the initial knot topology to plan a geometrically valid untangling strategy from a single snapshot.

V. Discussion and Limitations

While the particle-based representation offers significant advantages in terms of inference speed, memory footprint, and occlusion robustness, it introduces a reliance on the underlying physics model. Our current approach utilizes a fixed set of XPBD parameters for the Cosserat rod model and does not yet perform online system identification, so the simulation does not account for variations in rope stiffness, friction, or linear density. Any divergence in these parameters will eventually lead to drift between the simulated particle state and the physical rope. We envision two pathways to mitigate this: integrating an active System Identification module that estimates material properties from the initial snapshot, or a periodic re-grounding strategy that re-initializes the particle state from a fresh RGB-D snapshot after every grasp-and-pull action, effectively resetting tracking error between macro-actions.

Second, our vision baseline uses standard RGB ACT, while our particle pipeline implicitly exploits depth via multi-view reconstruction. The measured gap thus reflects the combined contribution of depth and explicit topological structure. Future work will evaluate an RGB-D ACT baseline to isolate the specific gains of depth versus structural representation.

Finally, raw RGB data remains inherently flexible, and vision-based policies are more intuitive and require no manual modeling of the object. However, their generalization scales with the size and diversity of the training data, so covering multiple rope types or changing environments demands extensive data collection. Our sim-grounded approach instead encodes physical priors directly into the state representation, offering a more sample-efficient path toward generalizable manipulation at the cost of requiring a reliable initial state estimator.

VI. Conclusion

In this work, we investigated the impact of state representation on learning dexterous bimanual manipulation from limited and diverse human demonstrations. By benchmarking a vision-based ACT policy against a sim-grounded particle-based policy, we quantified the significant observability gap that exists when mapping raw pixels to complex deformable dynamics. Grounding a single visual snapshot into a particle representation yielded a 30.8% reduction in L1 error and a 7.2x speedup in inference on a zero-shot rope instance.

These gains suggest that for high-degree-of-freedom tasks like rope untangling, the bottleneck is not necessarily the amount of human data or the capacity of the transformer, but the structural clarity of the observation space. By shifting the burden of representation from end-to-end visual learning to physics-based grounding, we provide a path toward more data-efficient and computationally lean robot learning.

References

[1] J. Grannen, P. Sundaresan, B. Thananjeyan, J. Ichnowski, A. Balakrishna, V. Viswanath, M. Laskey, J. Gonzalez, and K. Goldberg, "Untangling Dense Knots by Learning Task-Relevant Keypoints," in *Proceedings of the 2020 Conference on Robot Learning*. PMLR, Oct. 2021, pp. 782–800. [Online]. Available: https://proceedings.mlr.press/v155/grannen21a.html

[2] P. Sundaresan, J. Grannen, B. Thananjeyan, A. Balakrishna, J. Ichnowski, E. Novoseller, M. Hwang, M. Laskey, J. Gonzalez, and K. Goldberg, "Untangling Dense Non-Planar Knots by Learning Manipulation Features and Recovery Policies," in *Robotics: Science and Systems XVII*. Robotics: Science and Systems Foundation, Jul. 2021. [Online]. Available: http://www.roboticsproceedings.org/rss17/p013.pdf

[3] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware," Apr. 2023, arXiv:2304.13705 [cs]. [Online]. Available: http://arxiv.org/abs/2304.13705

[4] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion Policy: Visuomotor Policy Learning via Action Diffusion," Mar. 2024, arXiv:2303.04137 [cs]. [Online]. Available: http://arxiv.org/abs/2303.04137

[5] K. Chen, Y. Long, and M. Shang, "PIPHEN: Physical Interaction Prediction with Hamiltonian Energy Networks," Nov. 2025, arXiv:2511.16200 [cs]. [Online]. Available: http://arxiv.org/abs/2511.16200

[6] Y. Zhu, A. Joshi, P. Stone, and Y. Zhu, "VIOLA: Imitation Learning for Vision-Based Manipulation with Object Proposal Priors," Mar. 2023, arXiv:2210.11339 [cs]. [Online]. Available: http://arxiv.org/abs/2210.11339

[7] M. Macklin, M. Müller, and N. Chentanez, "XPBD: position-based simulation of compliant constrained dynamics," in *Proceedings of the 9th International Conference on Motion in Games*, ser. MIG '16. New York, NY, USA: Association for Computing Machinery, Oct. 2016, pp. 49–54. [Online]. Available: https://dl.acm.org/doi/10.1145/2994258.2994272

[8] E. Cosserat and F. Cosserat, "Theorie des Corps deformables," *Nature*, vol. 81, no. 2072, pp. 67–67, Jul. 1909. [Online]. Available: https://www.nature.com/articles/081067a0

[9] T. Kugelstadt and E. Schömer, "Position and Orientation Based Cosserat Rods," 2016.

[10] J. Hsu, T. Wang, K. Wu, and C. Yuksel, "Stable Cosserat Rods," in *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, ser. SIGGRAPH Conference Papers '25. New York, NY, USA: Association for Computing Machinery, Jul. 2025, pp. 1–10. [Online]. Available: https://dl.acm.org/doi/10.1145/3721238.3730618

[11] J. Xiang, H. Dinkel, H. Zhao, N. Gao, B. Coltin, T. Smith, and T. Bretl, "TrackDLO: Tracking Deformable Linear Objects Under Occlusion With Motion Coherence," *IEEE Robotics and Automation Letters*, vol. 8, no. 10, pp. 6179–6186, Oct. 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10214157

[12] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3D Diffusion Policy: Generalizable Visuomotor Policy Learning via Simple 3D Representations," Sep. 2024, arXiv:2403.03954 [cs]. [Online]. Available: http://arxiv.org/abs/2403.03954

[13] Y. Ze, Z. Chen, W. Wang, T. Chen, X. He, Y. Yuan, X. B. Peng, and J. Wu, "Generalizable Humanoid Manipulation with 3D Diffusion Policies," Sep. 2025, arXiv:2410.10803 [cs]. [Online]. Available: http://arxiv.org/abs/2410.10803

[14] Y. Zhu, Z. Jiang, P. Stone, and Y. Zhu, "Learning Generalizable Manipulation Policies with Object-Centric 3D Representations," Oct. 2023, arXiv:2310.14386 [cs]. [Online]. Available: http://arxiv.org/abs/2310.14386

[15] J. Sanchez, J. A. Corrales Ramon, B.-C. Bouzgarrou, and Y. Mezouar, "Robotic Manipulation and Sensing of Deformable Objects in Domestic and Industrial Applications: A Survey," *The International Journal of Robotics Research*, vol. 37, no. 7, pp. 688 – 716, Jun. 2018. [Online]. Available: https://uca.hal.science/hal-01816189

[16] H. Yin, A. Varava, and D. Kragic, "Modeling, learning, perception, and control methods for deformable object manipulation," *Science Robotics*, vol. 6, no. 54, p. eabd8803, May 2021. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.abd8803

[17] J. Zhu, A. Cherubini, C. Dune, D. Navarro-Alarcon, F. Alambeigi, D. Berenson, F. Ficuciello, K. Harada, J. Kober, X. Li, J. Pan, W. Yuan, and M. Gienger, "Challenges and Outlook in Robotic Manipulation of Deformable Objects," Dec. 2021, arXiv:2105.01767 [cs]. [Online]. Available: http://arxiv.org/abs/2105.01767

[18] V. Viswanath, K. Shivakumar, J. Kerr, B. Thananjeyan, E. Novoseller, J. Ichnowski, A. Escontrela, M. Laskey, J. E. Gonzalez, and K. Goldberg, "Autonomously Untangling Long Cables," Jul. 2022, arXiv:2207.07813 [cs]. [Online]. Available: http://arxiv.org/abs/2207.07813

[19] B. Güler, K. Pompetzki, S. Manschitz, and J. Peters, "Towards Assistive Teleoperation for Knot Untangling," in *German Robotics Conference (GRC)*, Mar. 2025.

[20] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "SAM 2: Segment Anything in Images and Videos," Oct. 2024, arXiv:2408.00714 [cs]. [Online]. Available: http://arxiv.org/abs/2408.00714

[21] The Newton Contributors, "Newton: GPU-accelerated physics simulation for robotics and simulation research," Apr. 2025. [Online]. Available: https://github.com/newton-physics/newton

[22] Y. Cai, P. Jansonnie, C. d. Farias, O. Arenz, and J. Peters, "GaussTwin: Unified Simulation and Correction with Gaussian Splatting for Robotic Digital Twins," Mar. 2026, arXiv:2603.05108 [cs]. [Online]. Available: http://arxiv.org/abs/2603.05108