CRAFT: Video Diffusion for Bimanual Robot Data Generation

Anonymous Authors

***Abstract*—Bimanual robot learning from demonstrations is fundamentally limited by the cost and narrow visual diversity of real-world data, which constrains policy robustness across viewpoints, object configurations, and embodiments. We present Canny-guided Robot Data Generation using Video Diffusion Transformers (CRAFT), a video diffusion-based framework for scalable bimanual demonstration generation that synthesizes temporally coherent manipulation videos while producing action labels. By conditioning video diffusion on edge-based structural cues extracted from simulator-generated trajectories, CRAFT produces physically plausible trajectory variations and supports a unified augmentation pipeline spanning object pose changes, camera viewpoints, lighting and background variations, cross-embodiment transfer, and multi-view synthesis. We leverage a pre-trained video diffusion model to convert simulated videos, along with action labels from the simulation trajectories, into action-consistent demonstrations. Starting from only a few real-world demonstrations, CRAFT generates a large, visually diverse set of photorealistic training data, bypassing the need to replay demonstrations on the real robot (Sim2Real). Across simulated and real-world bimanual tasks, CRAFT improves success rates over existing augmentation strategies and straightforward data scaling, demonstrating that diffusion-based video generation can substantially expand demonstration diversity and improve generalization for dual-arm manipulation tasks. Our project website is available at: https://craftaug.github.io/.**

I. Introduction

Imitation learning with teleoperated datasets has enabled capable bimanual manipulation [1]–[3], but scaling to diverse embodiments, viewpoints, and task variations remains data-intensive. While data augmentation is a promising strategy [4]–[6], existing works address only subsets of augmentations, e.g., single camera views [4, 7] or cross-embodiment transfer [8, 9] without a unified pipeline.

We present **CRAFT**, a unified data augmentation framework that constructs a digital twin to generate simulation trajectories, extracts Canny-edge control videos, and conditions a video diffusion model [10] on these edges with a real-world reference image and language instruction. Canny-edge conditioning preserves structural contours while abstracting simulation details, enabling augmentations spanning object pose, color, background, lighting, camera viewpoints, multi-view generation, and cross-embodiment transfer in a single pipeline. We demonstrate that policies trained on CRAFT-generated data significantly outperform baselines across simulation and real-world experiments.

II. Related Work

A. Video Generation For Robotics

Recent diffusion-based video generative models produce high-fidelity frames from conditioning inputs such as text or images [11]–[13]. We use Wan 2.1 [10], though any model supporting Canny-edge conditioning [14] is applicable. Prior works use video diffusion for trajectory prediction [15, 16] or world model learning [17]; we instead synthesize action-labeled demonstrations for imitation learning directly. Most related, AnchorDream [18] conditions on rendered robot motion traces without a simulator, limiting augmentation diversity. CRAFT leverages a simulator and digital twin to produce physically plausible trajectories across diverse scene configurations, yielding higher-fidelity demonstrations across bimanual and multi-view settings, at the cost of requiring simulator and object access.

B. Data Augmentation for Imitation Learning

Data augmentation has emerged as a practical tool for scaling imitation learning without additional demonstrations. Prior work uses generative models to alter visual context such as backgrounds or objects while keeping actions fixed [5, 6, 19], unlike state-based approaches [20, 21]. CRAFT similarly adjusts visual context but additionally expands the action distribution without requiring high-fidelity scene reconstruction [22]. Viewpoint augmentation has been studied for third-person [4, 23] and wrist-camera [7, 24] settings separately; CRAFT unifies both. Finally, Real2Sim2Real methods [25, 26] augment both state and action data via simulation rollouts, but require a final Sim2Real step. CRAFT instead uses a video diffusion model to synthesize photorealistic images directly from simulator trajectories, preserving coordination constraints and contact dynamics without real-world rollout collection.

III. Problem Statement

Our focus is on scalable data generation for vision-based imitation learning in bimanual manipulation, where a policy π_θ with parameters θ is trained from expert demonstrations using third-person RGB, wrist-camera, or combined RGB image observations. We denote a camera image at time t as I_t, simulation-generated images as I_t^{s}, video-diffusion-synthesized images as I_t^{d}, and ground truth deployment images as I_t^{g}. At deployment, the policy receives I_t^{g} and produces actions $a_t = \pi_\theta(I_t^{\mathrm{g}})$, where $a_t = (a_t^{\mathrm{l}}, a_t^{\mathrm{r}})$ specifies target joint positions and gripper actuation for the left and right arms, respectively.

We assume access to a small set of M real-world teleoperation demonstrations $\mathscr{D}^{\mathrm{real}} = \{\tau_1^{\mathrm{real}}, \ldots, \tau_M^{\mathrm{real}}\}$ and a simulation environment generating source videos $\mathbf{V}^{\mathrm{s}}$ via a digital twin pipeline. Each demonstration is a sequence of ground truth image observations and corresponding actions:

$$\tau_i^{\mathrm{real}} = (I_1^{\mathrm{g}}, a_1^{\mathrm{l}}, a_1^{\mathrm{r}}, \ldots, I_T^{\mathrm{g}}, a_T^{\mathrm{l}}, a_T^{\mathrm{r}}), \tag{1}$$

for a demonstration of T timesteps. Our goal is to synthesize a large, visually diverse set of generated demonstrations $\mathscr{D}^{\mathrm{gen}}$, with $|\mathscr{D}^{\mathrm{gen}}| \gg |\mathscr{D}^{\mathrm{real}}|$, where each synthesized demonstration contains diffusion-synthesized observations I_t^{d} resembling real-world images, to train a policy on $\mathscr{D}^{\mathrm{real}} \cup \mathscr{D}^{\mathrm{gen}}$.

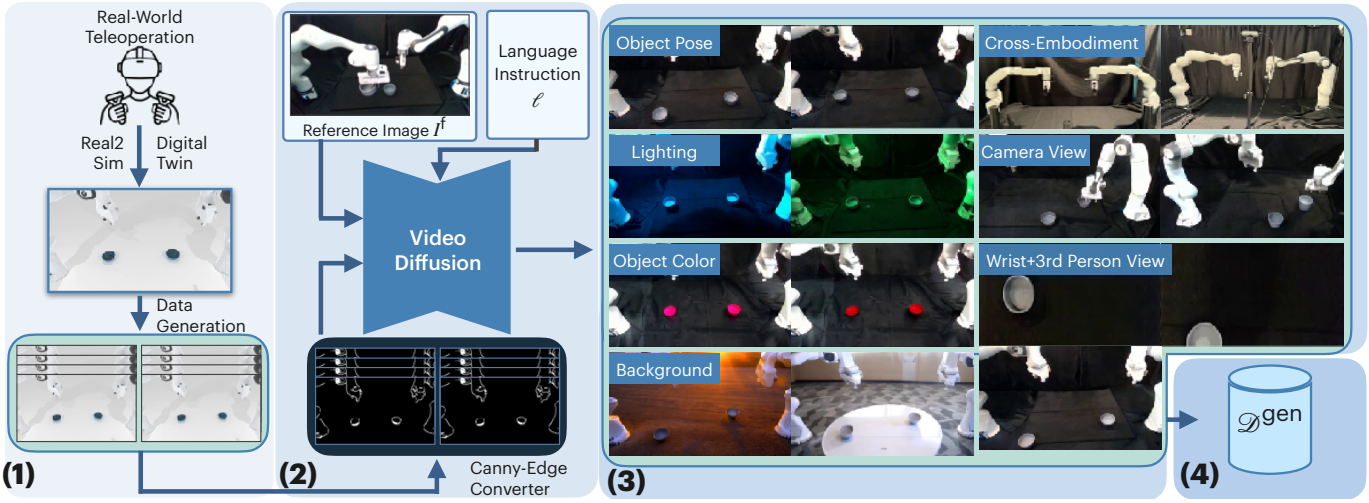

Fig. 1: **Method Overview. (1) Trajectory Expansion:** Real-world teleoperation data is first collected, and a digital twin pipeline transfers the objects and robot into simulation (Real2Sim). This simulation environment is then used for large-scale data generation. **(2) Video Generation:** The simulation trajectories are rendered into source videos and passed through a Canny-Edge Converter to extract structural edge representations, which are then combined with a real-world reference image and language instructions to condition a video diffusion model that synthesizes photorealistic video outputs. **(3) Augmented Dataset Construction:** The resulting generated videos support a wide range of visual variations, including object pose, lighting conditions, object color, background, cross-embodiment transfer, camera viewpoint, and combined wrist and third-person camera perspectives. **(4) Generated Dataset:** The synthesized videos are paired with action labels from the simulation trajectories, producing action-consistent demonstrations $\mathscr{D}^{\text{gen}}$ for downstream policy training.

IV. Method: CRAFT

CRAFT leverages a video diffusion model to synthesize photorealistic and visually diverse training videos for bimanual manipulation. Given a simulation-generated source video $\mathbf{V}^{\text{s}}$ produced from a digital twin pipeline, a real-world reference image I^{f}, and a language instruction ℓ, the model outputs a photorealistic target video $\mathbf{V}^{\text{d}} = \{I_1^{\text{d}}, \ldots, I_T^{\text{d}}\}$ that preserves robot motion structure while matching diverse real-world visual appearance. This is achieved through three stages: trajectory expansion (Section IV-A), video generation (Section IV-B), and augmented dataset construction for policy training (Section IV-C). CRAFT repeatedly applies this procedure to obtain $\mathscr{D}^{\text{gen}}$. Figure 1 provides an overview.

A. Trajectory Expansion

We construct a simulation counterpart $\mathscr{D}^{\text{sim}}$ from $\mathscr{D}^{\text{real}}$ using a digital twin pipeline, leveraging AprilTags [27] for object localization and known object meshes from RoboTwin [28], though any pipeline reconstructing object meshes and robot models in simulation is applicable [29]. Each real trajectory τ_i^{real} is replayed in simulation to generate a source video $\mathbf{V}^{\text{s}}$ and corresponding simulation trajectory τ_i^{sim}, resulting in a new simulation data of equal size as the original: $|\mathscr{D}^{\text{sim}}| = |\mathscr{D}^{\text{real}}|$.

To scale up data collection, we expand $\mathscr{D}^{\text{sim}}$ inspired by DexMimicGen [26]: each trajectory τ_i^{real} is decomposed into object-centric subtasks by annotating per-arm timestep boundaries, and a transformation operator $\mathscr{T}$ is applied to produce a new candidate trajectory $\mathscr{T}(\tau_i^{\text{real}})$ consistent with a novel sampled scene configuration. Each candidate is executed in simulation and validated for task success, retaining only successful trajectories to expand $\mathscr{D}^{\text{sim}}$. The validated trajectories are rendered into source videos $\mathbf{V}^{\text{s}}$, from which Canny-edge control videos $\mathbf{V}^{\text{c}}$ are extracted by filtering for salient structural edges, and fed into the video generation stage (Section IV-B) to synthesize $\mathscr{D}^{\text{gen}}$.

B. Video Generation

We model the conditional distribution:

$$p_\phi(\mathbf{V}^{\text{d}} \mid I^{\text{ref}}, \mathbf{V}^{\text{c}}, \ell), \tag{2}$$

where $\mathbf{V}^{\text{c}}$ is the Canny-edge control video, I^{ref} is a real-world reference image, ℓ is the language instruction, and ϕ denotes model parameters. We use Wan2.1-Fun-Control [10], which supports Canny-edge, depth, and skeleton pose conditioning. We choose Canny edges as they balance structural preservation of robot arms and objects while discarding fine-grained details, skeleton pose ignores object information entirely, while depth retains too much scene detail. By varying I^{ref}, ℓ, or both while keeping $\mathbf{V}^{\text{c}}$ fixed, the model synthesizes diverse videos without altering robot motion structure, with an LLM generating semantic variants of ℓ (Section IV-C).

C. Augmented Dataset Construction

We leverage Canny edges to guide demonstration augmentation across seven dimensions: object pose, lighting, object color, background, cross-embodiment, camera viewpoint, and wrist with third-person view generation. *CRAFT is flexible and modular where users can apply any subset of augmentation techniques and control the number of generated demonstrations.* Several augmentation techniques leverage LLMs to automatically generate diverse prompts and complete prompt lists are provided in the supplementary material.

1) Object Pose: To augment object poses, we introduce variations during trajectory expansion (Section IV-A). For each source trajectory τ_i^{real}, the simulator applies random translations and rotations to the target object's pose, sampled from

a uniform distribution with ranges set based on the physically feasible workspace. We also find that using a reference image capturing gripper-object contact yields higher fidelity contact synthesis in generated videos.

2) Lighting: To generate diverse lighting conditions, we augment the reference image I^{f} by prompting an image generation model, Veo3 [30], to synthesize variants under different ambient illumination, such as blue or green lighting. Unlike simple color jitter or RGB channel manipulation, this approach preserves scene properties such as shadows and surface reflections. The augmented reference images are then used to condition the video diffusion model, producing target videos $\mathbf{V}^{\mathrm{d}}$ with photorealistic lighting variations while preserving the underlying robot motion structure.

3) Object Color: To generate diverse object colors, we use a reference image I^{f} of the empty table scene without any objects. Conditioning on a reference image that contains objects would anchor the generated scene to the object color present in the reference, limiting color diversity. Since the reference image contains no objects, the Canny-edge control video $\mathbf{V}^{\mathrm{c}}$ provides the object contours to inform the diffusion model of their location, while the language instruction ℓ specifies the desired color to guide the appearance of the synthesized objects. By modifying the language instruction to specify the desired object color, the video diffusion model synthesizes target videos $\mathbf{V}^{\mathrm{d}}$ with the specified object appearance while preserving the scene layout and robot motion structure. To avoid manual prompt editing, we prompt an LLM to generate a list of object colors, from which we sample randomly during dataset construction.

4) Background: To generate diverse backgrounds, we omit the reference image I^{f} from the video diffusion model, as conditioning on it anchors the generated scene to the original environment. Instead, we modify the instruction ℓ to describe the desired background. To scale background diversity without manual prompting, we leverage an LLM to automatically generate a large set of varied background descriptions, which are then used to condition the video diffusion model to produce target videos $\mathbf{V}^{\mathrm{d}}$ with diverse scene appearances.

5) Cross-Embodiment: To enable cross-embodiment transfer, we map demonstrations from a source robot to a target robot using forward and inverse kinematics, and replace images of the source robot with photorealistic images of the target robot. This allows us to directly use the transferred demonstrations as training data for the target robot, without requiring any additional real-world data collection.

6) Camera Viewpoint: To generate diverse camera viewpoints, we place additional cameras inside the simulator and tile up to four simultaneous views into a single image. Formally, given $1 \leq N \leq 4$ camera views $\{I_t^{\mathrm{s},1}, \ldots, I_t^{\mathrm{s},N}\}$, we construct a tiled source image $I_t^{\mathrm{s,tile}} = \{I_t^{\mathrm{s},1}, \ldots, I_t^{\mathrm{s},N}\}$ from which the Canny-edge control video $\mathbf{V}^{\mathrm{c}}$ is extracted. The reference image I^{f} is similarly tiled to match, and both are fed into the video diffusion model to synthesize target videos $\mathbf{V}^{\mathrm{d}}$ across multiple camera perspectives simultaneously. The video diffusion model automatically preserves the tiled structure in the generated output and each viewpoint remains spatially contained within its corresponding tile without going into adjacent tiles.

7) Wrist and Third-Person View: Here, we follow the same tiling approach as camera viewpoint generation. Instead of tiling multiple third-person views, we tile the left wrist camera $I_t^{\mathrm{s,l}}$, right wrist camera $I_t^{\mathrm{s,r}}$, and a third-person (external) camera $I_t^{\mathrm{s,ext}}$ into a single image $I_t^{\mathrm{s,tile}} = \{I_t^{\mathrm{s,l}}, I_t^{\mathrm{s,r}}, I_t^{\mathrm{s,ext}}, \emptyset\}$, leaving the fourth tile empty, from which the Canny-edge control video $\mathbf{V}^{\mathrm{c}}$ is extracted. Tiling ensures spatial consistency across all viewpoints, and the reference image I^{f} is tiled accordingly before being fed into the video diffusion model to synthesize target videos $\mathbf{V}^{\mathrm{d}}$.

V. Real-World Experiments

A. Real-World Experiment Setup

We use a bimanual Franka Research 3 with GELLO [32], one or three Intel RealSense D435i cameras depending on whether wrist-camera observations are needed, an NVIDIA RTX 5090 for ACT training and inference, and **zero-shot** video generation via Wan2.1-Fun-Control 1.3B.

We evaluate each policy's success rate over 20 trials per task across three tasks spanning a range of bimanual coordination strategies.

- **`Lift Roller (LR)`:** A coordinated task where both arms simultaneously grasp and lift a dough roller.
- **`Place Cans in Plasticbox (PC)`:** A parallel task where both arms independently pick up cans and place them into a container.
- **`Stack Bowls (SB)`:** A sequential task where two bowls must be stacked on top of each other in order.

B. Real-World Results

We evaluate CRAFT across seven augmentation techniques (see Section IV-C): object pose, lighting, object color, background, cross-embodiment, camera viewpoint, and wrist with third-person view generation.

For each augmentation type, we compare the same baselines and evaluate under test conditions that vary only along that specific dimension. For example, unseen lighting conditions for lighting and unseen backgrounds for background, while all other visual factors remain fixed. Due to task simplicity, `Lift Roller` uses fewer demonstrations than `Place Cans in Plasticbox` and `Stack Bowls` across all methods:

- **ACT w/o Aug:** 50 (`LR`) / 100 (`PC`, `SB`) real-world demonstrations collected under standard conditions trained on ACT [33].
- **CRAFT Pose-Only:** 100 (`LR`) / 200 (`PC`, `SB`) demonstrations with object pose augmentation only. Inspired by RoboSplat [22], we include this baseline to assess the standalone impact of varying object poses.
- **ACT with Baseline Aug (augmentation-specific):** 50 (`LR`) / 100 (`PC`, `SB`) demonstrations with an augmentation-specific method. The specific baseline used for each augmentation type is noted in the corresponding subsection.

Method	Lighting			Background			Camera View			Object Color			Wrist + 3rd Person			Cross-Embodiment		
	LR	PC	SB	LR	PC	SB	LR	PC	SB	LR	PC	SB	LR	PC	SB	LR	PC	SB
ACT w/o Aug	3 / 20	1 / 20	0 / 20	4 / 20	0 / 20	0 / 20	6 / 20	3 / 20	2 / 20	2 / 20	0 / 20	1 / 20	15 / 20	11 / 20	13 / 20	5 / 20	2 / 20	3 / 20
CRAFT Pose-Only	5 / 20	3 / 20	2 / 20	7 / 20	2 / 20	3 / 20	13 / 20	5 / 20	7 / 20	5 / 20	2 / 20	3 / 20	13 / 20	8 / 20	10 / 20	4 / 20	1 / 20	2 / 20
ACT w/ Baseline Aug	13 / 20	9 / 20	8 / 20	4 / 20	5 / 20	6 / 20	14 / 20	8 / 20	6 / 20	15 / 20	9 / 20	11 / 20	N/A†	N/A†	N/A†	2 / 20	1 / 20	1 / 20
CRAFT (Ours)	**17 / 20**	**14 / 20**	**12 / 20**	**18 / 20**	**15 / 20**	**10 / 20**	**19 / 20**	**18 / 20**	**18 / 20**	**18 / 20**	**18 / 20**	**17 / 20**	**20 / 20**	**19 / 20**	**20 / 20**	**17 / 20**	**15 / 20**	**16 / 20**

† No suitable baseline augmentation method exists for this augmentation type.

TABLE I: **Real-World Results.** Success rates out of 20 for LR, PC, and SB across five augmentation techniques and cross-embodiment transfer. For augmentation techniques, all methods are evaluated under test conditions that vary only along that specific dimension, while all other visual factors remain fixed. Cross-Embodiment evaluates transfer from a bimanual xArm7 to a bimanual Franka Panda on LR, PC, and SB (see Section V-B), where CRAFT (Ours) uses 1000 generated demos without collecting any target robot demos, in contrast to Collected Target (ACT w/o Aug) which requires 100 demos on the target robot. All CRAFT (Ours) augmentation columns use 1000 generated demonstrations combined with the real-world collected demonstrations (100 for LR, 200 for PC, and 150 for SB). The "ACT w/ Baseline Aug" row refers to a different baseline for each augmentation type: Lighting (Color Jitter), Background (RoboEngine [6]), Camera View (VISTA [4]), Object Color (SAM3 [31]), and Cross-Embodiment (Shadow [8]). All methods are trained and evaluated using an ACT policy on the bimanual Franka.

- **CRAFT (Ours):** 1000 (LR, PC, SB) generated demonstrations and the original real-world demonstrations using our full augmentation pipeline.

VI. Conclusion

We present CRAFT, a scalable data generation pipeline for bimanual imitation learning that synthesizes photorealistic demonstrations across seven augmentation techniques via video diffusion conditioned on Canny-edge control videos, reference images, and language instructions. CRAFT consistently outperforms augmentation-specific baselines in the real world, demonstrating that scalable data generation can substitute for costly real-world data collection. We hope CRAFT inspires further work in video generation for robot learning.

References

[1] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, and et al., "π_0: A Vision-Language-Action Flow Model for General Robot Control," in Robotics: Science and Systems (RSS), 2024.

[2] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, and et al., "$\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization," in Conference on Robot Learning (CoRL), 2025.

[3] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu, "RDT-1B: a Diffusion Foundation Model for Bimanual Manipulation," in International Conference on Learning Representations (ICLR), 2025.

[4] S. Tian, B. Wulfe, K. Sargent, K. Liu, S. Zakharov, V. Guizilini, and J. Wu, "View-Invariant Policy Learning via Zero-Shot Novel View Synthesis," in Conference on Robot Learning (CoRL), 2024.

[5] X. Wang, K. Wu, Z. Zhao, H. Cao, Y. Zhao, Z. Xu, M. Li, S. Fan, D. Wu, Y. Zhang, N. Liu, Z. Che, and J. Tang, "RoboAug: One Annotation to Hundreds of Scenes via Region-Contrastive Data Augmentation for Robotic Manipulation," arXiv preprint arXiv:2602.14032, 2026.

[6] C. Yuan, S. Joshi, S. Zhu, H. Su, H. Zhao, and Y. Gao, "RoboEngine: Plug-and-Play Robot Data Augmentation with Semantic Robot Segmentation and Background Generation," in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2025.

[7] I.-C. A. Liu, J. Chen, G. Sukhatme, and D. Seita, "D-CODA: Diffusion for Coordinated Dual-Arm Data Augmentation," in Conference on Robot Learning (CoRL), 2025.

[8] M. Lepert, R. Doshi, and J. Bohg, "Shadow: Leveraging Segmentation Masks for Cross-Embodiment Policy Transfer," in Conference on Robot Learning (CoRL), 2024.

[9] L. Y. Chen, K. Hari, K. Dharmarajan, C. Xu, Q. Vuong, and K. Goldberg, "Mirage: Cross-Embodiment Zero-Shot Policy Transfer with Cross-Painting," in Robotics: Science and Systems (RSS), 2024.

[10] T. Wan, A. Wang, B. Ai, B. Wen, C. Mao, C. Xie, D. Chen, and et al., "Wan: Open and Advanced Large-Scale Video Generative Models," arXiv preprint arXiv:2503.20314, 2025.

[11] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in Neural Information Processing Systems (NeurIPS), 2020.

[12] NVIDIA, "Cosmos World Foundation Model Platform for Physical AI," arXiv preprint arXiv:2501.03575, 2025.

[13] Z. Yang, J. Teng, W. Zheng, B. Xu, X. Gu, Y. Dong, J. Tang, and et al., "CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer," in International Conference on Learning Representations (ICLR), 2026.

[14] J. Canny, "A Computational Approach to Edge Detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-8, no. 6, pp. 679–698, 1986.

[15] J. Jang, S. Ye, Z. Lin, J. Xiang, D. Fox, J. Kautz, S. Reed, Y. Zhu, L. Fan, , and et al., "DreamGen: Unlocking Generalization in Robot Learning through Video World Models," arXiv preprint arXiv:2505.12705, 2025.

[16] Y. Guo, L. X. Shi, J. Chen, and C. Finn, "Ctrl-World: A Controllable Generative World Model for Robot Manipulation," in International Conference on Learning Representations (ICLR), 2026.

[17] J. Mao, S. He, H.-N. Wu, Y. You, S. Sun, Z. Wang, Y. Bao, H. Chen, L. Guibas, V. Guizilini, H. Zhou, and Y. Wang, "Robot Learning from a Physical World Model," in IEEE International Conference on Robotics and Automation (ICRA), 2026.

[18] J. Ye, R. Xue, B. V. Hoorick, P. Tokmakov, M. Z. Irshad, Y. Wang, and V. Guizilini, "AnchorDream: Repurposing Video Diffusion for Embodiment-Aware Robot Data Synthesis," in IEEE International Conference on Robotics and Automation (ICRA), 2026.

[19] Z. Chen, Z. Mandi, H. Bharadhwaj, M. Sharma, S. Song, A. Gupta, and V. Kumar, "Semantically Controllable Augmentations for Generalizable Robot Learning," in International Journal of Robotics Research (IJRR), 2024.

[20] L. Ke, Y. Zhang, A. Deshpande, S. Srinivasa, and A. Gupta, "CCIL: Continuity-based Data Augmentation for Corrective Imitation Learning," in International Conference on Learning Representations (ICLR), 2024.

[21] M. Laskey, J. Lee, R. Fox, A. D. Dragan, and K. Goldberg, "DART: Noise Injection for Robust Imitation Learning," in Conference on Robot Learning (CoRL), 2017.

[22] S. Yang, W. Yu, J. Zeng, J. Lv, K. Ren, C. Lu, D. Lin, and J. Pang, "Novel demonstration generation with gaussian splatting enables robust one-shot manipulation," in Robotics: Science and Systems (RSS), 2025.

[23] J. Chen, I.-C. A. Liu, G. Sukhatme, and D. Seita, "ROPA: Synthetic Robot Pose Generation for RGB-D Bimanual Data Augmentation," in IEEE International Conference on Robotics and Automation (ICRA), 2026.

[24] X. Zhang, M. Chang, P. Kumar, and S. Gupta, "Diffusion Meets DAgger: Supercharging Eye-in-hand Imitation Learning," in Robotics: Science and Systems (RSS), 2024.

[25] A. Mandlekar, S. Nasiriany, B. Wen, I. Akinola, Y. Narang, L. Fan, Y. Zhu, and D. Fox, "MimicGen: A Data Generation System for Scalable Robot Learning using Human Demonstrations," in Conference on Robot Learning (CoRL), 2023.

[26] Z. Jiang, Y. Xie, K. Lin, Z. Xu, W. Wan, A. Mandlekar, L. Fan, and Y. Zhu, "DexMimicGen: Automated Data Generation for Bimanual Dexterous Manipulation via Imitation Learning," in IEEE International Conference on Robotics and Automation (ICRA), 2025.

[27] J. Wang and E. Olson, "Apriltag 2: Efficient and robust fiducial detection," in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2016.

[28] T. Chen, Z. Chen, B. Chen, Z. Cai, Y. Liu, Z. Li, Q. Liang, X. Lin, Y. Ge, Z. Gu et al., "Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation," arXiv preprint arXiv:2506.18088, 2025.

[29] Z. Chen, A. Walsman, M. Memmel, K. Mo, A. Fang, K. Vemuri, A. Wu, D. Fox, and A. Gupta, "Urdformer: A pipeline for constructing articulated simulation environments from real-world images," in Robotics: Science and Systems (RSS), 2024.

[30] Google, "Veo 3," https://deepmind.google/models/veo/, 2025.

[31] N. Carion, L. Gustafson, Y.-T. Hu, N. Ravi, K. Saenko, P. Zhang, C. Feichtenhofer, and et al., "Sam 3: Segment anything with concepts," arXiv:2511.16719, 2025.

[32] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel, "GELLO: A General, Low-Cost, and Intuitive Teleoperation Framework for Robot Manipulators," in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2024.

[33] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware," in Robotics: Science and Systems (RSS), 2023.