Robotics: Science and Systems 2026
Sydney, Australia, July 13–17, 2026

One-Shot Real-World Demonstration Synthesis for Scalable Bimanual Manipulation

Huayi Zhou* and Kui Jia*†‡
*The Chinese University of Hong Kong, Shenzhen. †DexForce, Shenzhen. ‡The Corresponding Author.
https://hnuzhy.github.io/projects/BiDemoSyn

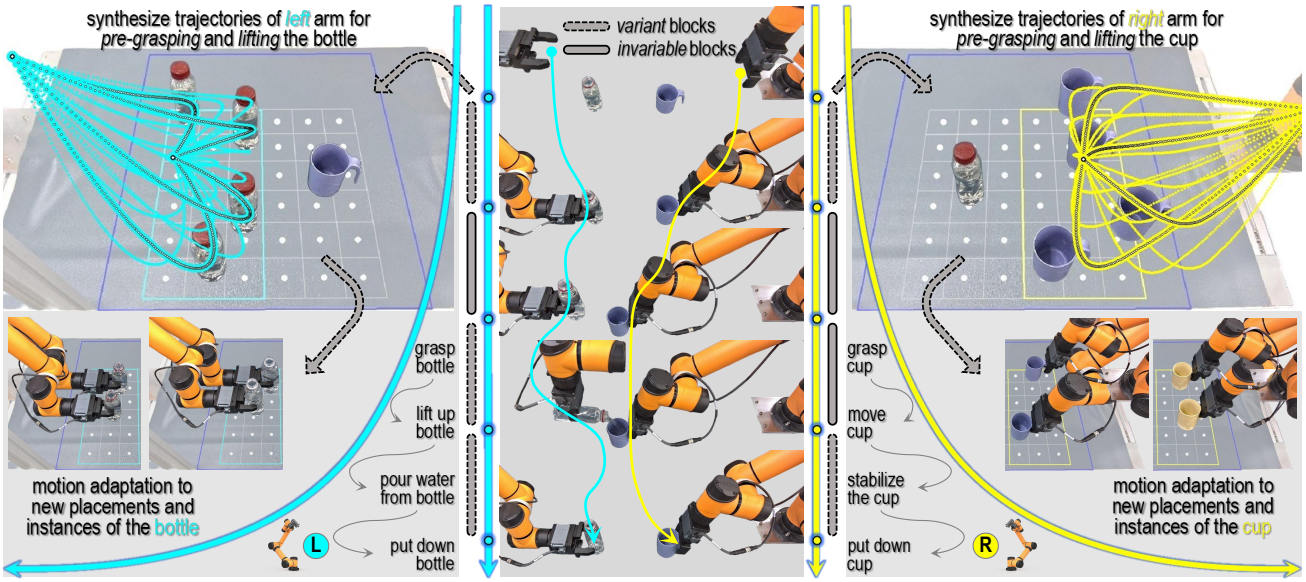

Fig. 1: **From One to Many** ①→Ⓝ. Taking the dual-arm coordinated `pouring` water task as an example, we illustrate how to synthesize corresponding pre-grasping and lifting trajectories for different new placements and novel instances of manipulated objects (*e.g.*, bottle and cup) during the initial frame alignment phase for pre-grasping.

***Abstract*—Learning dexterous bimanual manipulation policies critically depends on large-scale, high-quality demonstrations, yet current paradigms face inherent trade-offs: teleoperation provides physically grounded data but is prohibitively labor-intensive, while simulation-based synthesis scales efficiently but suffers from sim-to-real gaps. We present BiDemoSyn, a framework that synthesizes contact-rich, physically feasible bimanual demonstrations from a single real-world example. The key idea is to decompose tasks into invariant coordination blocks and variable, object-dependent adjustments, then adapt them through vision-guided alignment and lightweight trajectory optimization. This enables the generation of thousands of diverse and feasible demonstrations within several hours, without repeated teleoperation or reliance on imperfect simulation. Across six dual-arm tasks, we show that policies trained on BiDemoSyn data generalize robustly to novel object poses and shapes, significantly outperforming recent strong baselines. Beyond the one-shot setting, BiDemoSyn naturally extends to few-shot-based synthesis, improving object-level diversity and out-of-distribution generalization while maintaining strong data efficiency. Moreover, policies trained on BiDemoSyn data exhibit zero-shot cross-embodiment transfer to new robotic platforms, enabled by object-centric observations and a simplified 6-DoF end-effector action representation that decouples policies from embodiment-specific dynamics. By bridging the gap between efficiency and real-world fidelity, BiDemoSyn provides a scalable path toward practical imitation learning for complex bimanual manipulation without compromising physical grounding.**

I. Introduction

Recent advances in robot manipulation have been propelled by data-driven imitation learning, where visuomotor policies trained on large-scale demonstration datasets enable dexterous single-arm and bimanual tasks previously deemed intractable. Methods like ACT [103], DP [16] and π_0 [5] exemplify this trend, achieving remarkable performance in contact-rich scenarios such as continuously pushing, battery assembly, and coordinated garment folding. These successes, however, hinge on a critical yet underappreciated premise: the availability of diverse, high-quality demonstrations, which are typically collected via expert teleoperation. As the community shifts toward more complex long-horizon bimanual tasks, the data scalability bottleneck becomes increasingly apparent. Even those largest manipulation datasets like RH20T [24], DROID [44], Open X-Embodiment [68] and AgiBot World [6] struggle to provide sufficient geometric and kinematic diversity for robust generalization, underscoring a pivotal question: *how to scale real-world bimanual demonstration collection without compromising physical fidelity?*

The answer, unfortunately, lies in a fundamental trade-off. Existing data acquisition methods fall into two paradigms, each with critical shortcomings. On one hand, human-operated

systems (*e.g.*, the ALOHA series [103, 26, 104]) yield physically-accurate trajectories but demand prohibitive expertise, rendering them impractical for scaling to long-horizon tasks. On the other hand, simulation-based methods (*e.g.*, MimicGen [62, 40], RoboGen [85], and RoboCasa [66]) bypass human labor through programmatic data generation utilizing massive 3D assets, yet discrepancies in contact dynamics and visual rendering inevitably plague real-world deployment [2, 64]. Thus, the tension between data quality (grounded in reality) and quantity (enabled by automation) persists as a key barrier to scalable imitation learning, particularly for contact-rich bimanual tasks requiring precise dual-arm coordination.

To bridge this gap, we present **BiDemoSyn** that synthesizes real-world bimanual demonstrations from a single exemplar, by algorithmically amplifying task semantics through a hierarchical decomposition of invariance and adaptability (refer Fig. 1 and Fig. 2). Our motivation is inherited from the well-established single-arm one-shot imitation learning [63, 60, 4, 100, 21, 106]. The key insight lies in recognizing that complex manipulation tasks exhibit a duality: while certain motion primitives (*e.g.*, stable grasp sequences, dual-arm synchronization) remain invariant across instances, others (*e.g.*, trajectory segments adapting to object shifts) require dynamic adjustments to contextual variations. Specifically, through a three-stage process, **BiDemoSyn** first deconstructs the initial demonstration into invariant primitives and geometry-sensitive components. Then, it leverages visual perception to spatially align these elements in novel scenes, enabling generalization across object geometries and workspace layouts with minor effort (primarily involving reinitializing objects). Finally, a physics-aware optimizer jointly modulates dual-arm trajectories, injecting diversity into adaptable components while enforcing real-world constraints on collision avoidance and kinematic coordination. This integrated approach achieves *One-to-Many* demonstration synthesis: a single exemplar spawns hundreds of trajectories that preserve task intent while adapting to unseen configurations, all grounded in real-world dynamics rather than simulated proxies.

Unlike prior simulation-based methods [62, 85, 66, 40], **BiDemoSyn** operates entirely in the physical domain, ensuring synthesized data inherits the fidelity of human demonstrations. By unifying algorithmic scalability with physical realism, our framework empowers policies to transcend the limitations of narrow manually-collected data distributions. Experiments on real dual-arm platforms validate the effectiveness of **BiDemoSyn**. Visuomotor policies trained with synthesized demonstrations can achieve superior success rates and generalization in tasks requiring precise contact coordination (*e.g.*, small hole insertion, rotation angle control, purposeful handover and articulated object manipulation).

Our contributions are threefold: (1) *One-Shot Synthesis Framework*: A systematic pipeline combining task decomposition, vision-guided adaptation, and contact-aware trajectory optimization to generate scalable real-world bimanual demonstrations. (2) *Reality-Grounded Data Generation*: A completely simulator-free method for synthesizing bimanual demonstrations, ensuring physical fidelity by construction. (3) *Empirical Validation in Complex Tasks*: Comprehensive real-robot experiments demonstrating significant improvements in policy robustness and cross-configuration generalization on various bimanual manipulation tasks.

II. Related Works

Bimanual Robotic Manipulation. Prior literatures primarily target task-specific challenges (such as cloth folding [20, 88, 7], untangling [29, 70], untwisting [54, 1], and handover [37, 51]) through handcrafted controllers or narrow skill libraries, limiting generalization due to rigid motion primitives and excessive assumptions. General methods [90, 18, 10, 11, 33, 105, 106] either rely on predefined hierarchical arm roles (*e.g.*, leader-follower [48, 57, 32] and stabilizer-actor [30, 56]) or hinge generalization entirely on dataset diversity, both of which are ultimately bottlenecked by the impractical costs of data acquisition. Recently, ALOHA series [103, 26, 104] have revolutionized bimanual manipulation by dexterous low-cost teleoperation. Meanwhile, the visuomotor policy learning paradigm [16, 98, 94] based on diffusion models [35, 79] has largely expanded the action modeling space and data utilization efficiency. Their followers [80, 45, 58, 5, 71, 53] expect to train universal policies using large-scale teleoperation data but face scalability barriers from human effort. To improve reachability and dexterity, some studies use specialized hardware (*e.g.*, multi-finger hands [82, 76, 25, 15] or tactile sensors [55, 9]), which complicate real-world adoption. In contrast, we utilize a fixed-base dual-arm platforrm with parallel grippers, synthesizing diverse demonstrations from a single example. By optimizing coordination as dynamic constraints rather than predefined hierarchies, we bypass hardware specificity and data scalability limitations, enabling task-agnostic policies grounded in physical feasibility.

Bimanual Demonstration Collection. The dominant path for acquiring bimanual demonstrations is *human teleoperation* [44, 6], which delivers high-quality data but suffers from prohibitive scalability costs. To alleviate it, two alternative routes have emerged: *simulation-based synthesis* [28, 36, 52, 96, 83] and *learning from human videos* [72, 97, 102, 42, 3]. The former, exemplified by MimicGen [62] and DexMimicGen [40], augments a few demonstrations by generating variations in simulation using geometric transformations and kinematic constraints. Analogously, RoboGen [85] and RoboCasa [66] leverage 3D assets to procedurally generate full-scene manipulation data, yet such methods inevitably inherit *sim-to-real* gaps, from unrealistic textures to unphysical dynamics. The latter extracts bimanual hand-object manipulation from egocentric videos [99, 59, 31] and maps them to dual arms via motion retargeting [50, 43, 12, 13] or non-privileged representations (*e.g.*, keypoints [69, 27, 87], affordances [41, 65] and correspondences [70, 46, 101]). While human videos are abundant and low-cost, large morphological mismatches between humanity and robotic embodiments often need heuristic translation rules and hinder direct applicability. Our work navigates these trade-offs by synthesizing demonstrations directly

in real-world. Given a single exemplar, we diversify trajectories through vision-based adaptation and physics-compliant optimization. This balances the fidelity and scalability, while avoiding retargeting hurdles, thereby enabling practical and scalable data acquisition for contact-rich bimanual tasks.

Several concurrent works exhibit notable limitations. DemoGen [93] and YOTO [107, 108] diversify demos via 3D editing of the initial single-view point cloud, but perspective ambiguities induce visual artifacts in synthesized data. ODIL [84] avoids editing by segmenting objects for visual servoing and iteratively replanning trajectories, which is cumbersome for scalability. MoMaGen [49] is a long-horizon mobile bimanual demonstrations generator yet mainly focusing on the simulation. Our **BiDemoSyn** sidesteps these issues: synthesizing trajectories directly in real-world ensures visual-physical consistency without manual edits or robot replaying.

III. Preliminaries

Problem Formulation. Bimanual imitation learning aims to train visuomotor policies $\pi_\theta(\mathbf{o}_t) \rightarrow \mathbf{a}_t$, where $\mathbf{o}_t$ denotes multimodal observations (*e.g.*, RGB images, joint states or point clouds) and $\mathbf{a}_t$ represents dual-arm actions. While architectures like ACT [103] and DP [16] enhance policy expressivity, their generalization critically depends on expert demonstrations $\mathcal{D} = \{\tau_i\}_{i=0}^{N}$ that densely span the feasible state-action manifold. Existing approaches either collect $\mathcal{D}$ via labor-intensive teleoperation or generate data in simulation, both failing to balance real-world fidelity and scalability. We address this gap by proposing the following problem:

Consider a bimanual task defined by an initial state s_0 and a specific goal g. Given a single real-world demonstration $\tau = \{\mathbf{o}_t, \mathbf{a}_t\}_{t=0}^{T}$, our objective is to synthesize a dataset $\mathcal{D}_{syn} = \{\tau_i\}_{i=0}^{M} (M \gg 1)$ such that: (1) *Task Consistency*: Every $\tau_i \in \mathcal{D}_{syn}$ achieves the task goal g under perturbed initial states (*e.g.*, new object poses, scene layouts). (2) *Physical Admissibility*: Trajectories adhere to kinematic and dynamic constraints of the real-world system (*e.g.*, collision avoidance, dual-arm coordination). (3) *Diversity*: $\mathcal{D}_{syn}$ covers task-relevant variations to enable robust policy training.

Visuomotor Policy Learning. Our π_θ maps observations $\mathbf{o}_t$ to dual-arm actions $\mathbf{a}_t = \left[\mathbf{a}_t^L, \mathbf{a}_t^R\right] \in \mathbb{R}^{2d}$, where $\mathbf{a}_t^L$, $\mathbf{a}_t^R$ denotes left/right end-effector poses. We adapt a diffusion policy model [16, 98, 94] to generate action sequences $\mathbf{A} = \left[\mathbf{a}_{t:t+H}^L, \mathbf{a}_{t:t+H}^R\right] \in \mathbb{R}^{2H\times d}$ over a horizon H. The diffusion process iteratively denoises a noisy action sequence $\mathbf{A}_k$ toward kinematic feasibility over K steps:

$$\mathbf{A}_k = \sqrt{\alpha_k}\mathbf{A}_0 + \sqrt{1-\alpha_k}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{\Sigma}), \tag{1}$$

where α_k follows a cosine schedule [67]. And $\mathbf{\Sigma} \in \mathbb{R}^{2H\times 2H}$ is a block-diagonal covariance matrix $\begin{bmatrix} \mathbf{\Sigma}_L & \rho\mathbf{\Sigma}_{LR} \\ \rho\mathbf{\Sigma}_{LR}^\top & \mathbf{\Sigma}_R \end{bmatrix}$ encoding arm coordination priors, where $\mathbf{\Sigma}_L$, $\mathbf{\Sigma}_R$ governing temporal smoothness per arm, $\mathbf{\Sigma}_{LR}$ modeling inter-arm dependencies, and $\rho \in [0,1]$ controlling coordination strength. Then, a denoiser μ_θ processes the noised sequence conditioned on visual observations $\mathbf{o}_t$ via:

$$\mu_\theta(\mathbf{A}_k, \mathbf{o}_t, k) = \texttt{UNet}_\theta(\texttt{Concat}\left[\mathbf{A}_k, \mathsf{E}(\mathbf{o}_t)\right], k), \tag{2}$$

where $\mathsf{E}(\mathbf{o}_t)$ is a vision encoder (ResNet [34] or PointNet [73]) extracting latent features. The $\texttt{UNet}_\theta$ employs asymmetric layers with early stages processing each arm independently, while deeper layers fusing bilateral coordinated features. Training minimizes:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{A}_0, \mathbf{o}_t, k}\left[\left\|\mu_\theta(\mathbf{A}_k, \mathbf{o}_t, k) - \mathbf{A}_0\right\|_{\mathbf{W}}^2\right], \tag{3}$$

where $\|\cdot\|_{\mathbf{W}}$ is a Mahalanobis norm with $\mathbf{W} = \mathbf{\Sigma}^{-1}$ to prioritize arm interactions kinematically. By training on $\mathcal{D}_{syn}$ with feasible trajectories synthesized from a single exemplar, the policy bypasses sim-to-real gaps while handling diverse bimanual coordination patterns.

IV. Methodology

Here, we detail **BiDemoSyn** for synthesizing bimanual demonstrations $\mathcal{D}_{syn}$ from a single exemplar τ with three compact stages (see Fig. 2): Sec. IV-A *Deconstruction of One-Shot Teaching*, which extracts invariant patterns and adaptable primitives; Sec. IV-B *Vision-based Initial Frame Alignment*, enabling generalization across geometric variations via efficient scene perception; and Sec. IV-C *Trajectory Modulation and Optimization*, ensuring physical feasibility via hierarchical kinematic constraints. Each stage sequentially addresses scalability, adaptability, and real-world fidelity to bridge the gap between data efficiency and policy robustness.

A. Deconstruction of One-Shot Teaching

Given a demonstration $\tau = \{\mathbf{o}_t, \mathbf{a}_t\}_{t=0}^{T}$, we decompose it into a *bimanual execution blocks* set $\{\mathcal{B}_i\}_{i=1}^{n}$, where each block $\mathcal{B}_i = (\mathbf{s}_i^L, \mathbf{s}_i^R)$ represents a discrete phase of dual-arm interaction. Here, $\mathbf{s}_i^L, \mathbf{s}_i^R \in \mathbb{R}^d$ denote the left/right arm state sequences (*e.g.*, end-effector 6-DoF poses and gripper status) within block $\mathcal{B}_i$. Blocks are categorized based on motion coordination. For **single-arm motion**, one arm exhibits significant transitions while the other remains static. Formally, for threshold δ (minimal motion saliency) and ζ (static tolerance):

$$\begin{aligned}&(\left\|\mathbf{s}_{i,e}^L - \mathbf{s}_{i,s}^L\right\| \geq \delta \wedge \left\|\mathbf{s}_{i,e}^R - \mathbf{s}_{i,s}^R\right\| \leq \zeta) \vee \\ &(\left\|\mathbf{s}_{i,e}^R - \mathbf{s}_{i,s}^R\right\| \geq \delta \wedge \left\|\mathbf{s}_{i,e}^L - \mathbf{s}_{i,s}^L\right\| \leq \zeta).\end{aligned} \tag{4}$$

While, for **dual-arm coordination**, both arms synchronously or asynchronously adjust states:

$$\left\|\mathbf{s}_{i,e}^L - \mathbf{s}_{i,s}^L\right\| \geq \delta \wedge \left\|\mathbf{s}_{i,e}^R - \mathbf{s}_{i,s}^R\right\| \geq \delta. \tag{5}$$

1) From Blocks to Atomic Execution Primitives (AEPs): To enable modular adaptation, each block $\mathcal{B}_i$ is further refined into a minimal motion unit AEP, where an arm undergoes a salient state transition (*e.g.*, gripper closure, moving or pose shifts). For arm $A \in \{L, R\}$, an AEP is defined as:

$$\begin{aligned}\texttt{AEP}_A = \{&\mathbf{s}_t^A \rightarrow \mathbf{s}_{t+\Delta t}^A \mid \left\|\mathbf{s}_{t+\Delta t}^A - \mathbf{s}_t^A\right\|_2 \geq \gamma, \\ &\Delta t \leq T_{\max}, \texttt{CollisionFree}(\mathbf{s}_{[t,\Delta t]}^A)\},\end{aligned} \tag{6}$$

where γ is a distance threshold, $T_{\max}$ limits execution duration, and $\texttt{CollisionFree}(\cdot)$ enforces no collisions between the arm and objects/scene (optimized via forward kinematics). Unlike keyframe-based segments [39, 78, 61, 107], AEPs ignore acceleration profiles (assuming quasi-static motions) and prioritize task-oriented transitions over temporal granularity.

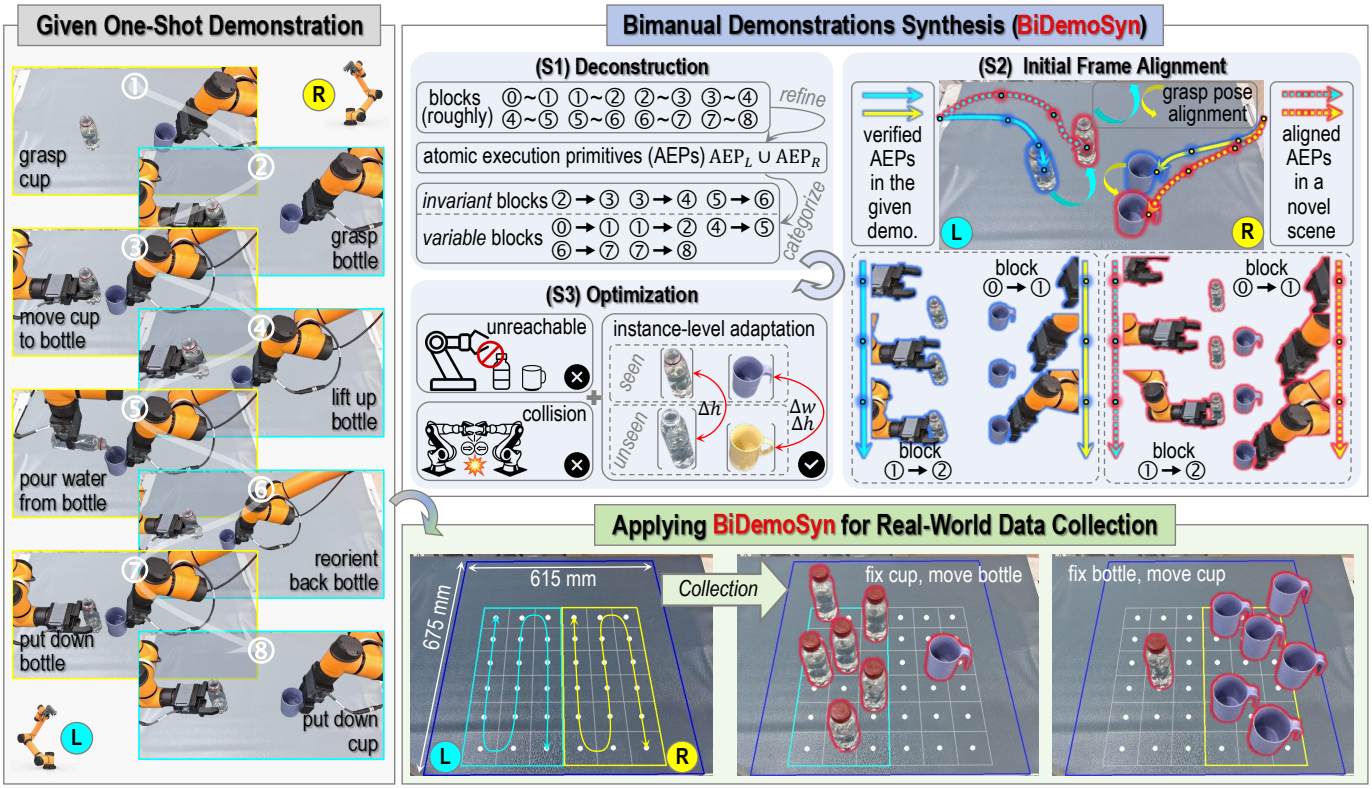

Fig. 2: The overview of **BiDemoSyn**. It consists of three stages (*e.g.*, deconstruction, alignment, and optimization) based on a given demonstration. Then, we can apply our method to complete data collection efficiently and conveniently in real-world.

2) Semantic Categorization for Adaptive Synthesis: The final step categorizes refined blocks into *invariant* and *variable* types. The invariant blocks encode task-semantic primitives (*e.g.*, screwing, pressing) or general motions (*e.g.*, lifting, transferring), which remain structurally consistent across variations. The variable blocks, such as object-centric goal-conditioned grasping, adapt to instance-level geometric variations including object poses and shapes through:

$$\mathcal{B}_i' = \mathcal{B}_i \circ \Phi(g, \mathbf{o}_{\text{novel}}), \tag{7}$$

where $\Phi(\cdot)$ is a visual adapter, g is the task goal, and $\mathbf{o}_{\text{novel}}$ is the novel observation. This structured decomposition enables selective diversification, allowing variable blocks to adapt to new scenes while preserving task fidelity in invariant blocks.

B. Vision-based Initial Frame Alignment

The variable blocks identified in Sec. IV-A require adapting to novel object poses and geometries. To achieve this, we propose a vision-driven adapter $\Phi(\cdot)$ that generalizes object-centric interactions in the given exemplar to new scenes. Given a novel observation $\mathbf{o}_{\text{novel}}$, our method executes three steps: *object perception*, *state estimation*, and *pose alignment*, ensuring the initial robotic grasp aligns with the task goal g. Illustrations are shown in Fig. 3.

1) Object Perception: We first detect and segment the target object in $\mathbf{o}_{\text{novel}}$ using an open-vocabulary detector (*e.g.*, YOLO-World [14] or YOLOE [81]) or vision foundation models (*e.g.*, Florence2 [89] with SAM2 [74] for robustness to rare categories). In practice, we prioritize foundation models to avoid detection failures. Now let $M \subseteq \mathbf{o}_{\text{novel}}$ denote the obtained object's binary mask.

2) State Estimation: Instead of uitlizing category-level CAD-based 6D pose estimation models (*e.g.*, FoundationPose [86]), we estimate the instance-level 6D object pose $\mathbf{P}_{\text{obj}} \in SE(3)$ using geometry-aware processing. Sepcifically, we adopt classic image moments [8, 47] to calculate the object mask centroid $\mathbf{c}$ and extract principal axes $\mathbf{R}$:

$$\begin{cases} \mathbf{c} &= \frac{1}{|M|}\sum_{(u,v)\in M}(u, v, \mathrm{d}(u,v)) \\ \mathbf{R} &= \mathtt{PCA}(\{(u, v, \mathrm{d}(u,v)) \mid (u,v) \in M\}) \end{cases} \tag{8}$$

where $\mathrm{d}(u,v)$ is the depth value. $\mathtt{PCA}$ is used to fit 3D points of M to determine orientation $\mathbf{R} \in SO(3)$. This yields $\mathbf{P}_{\text{obj}} = (\mathbf{R}, \mathbf{c})$, robust to arbitrary object states (e.g., fallen, inverted).

3) Pose Alignment: Let $\mathbf{P}_{\text{demo}}$ denote the object pose in the given demonstration. We compute the rigid transformation $\mathbf{T} \in SE(3)$ that maps $\mathbf{P}_{\text{demo}}$ to $\mathbf{P}_{\text{obj}}$. This transformation is applied to the initial grasp pose $\mathbf{P}_{\text{grasp,demo}}$ in the variable block $\mathcal{B}_i$, yielding the adapted grasp pose:

$$\mathbf{P}_{\text{grasp,novel}} = \mathbf{T} \cdot \mathbf{P}_{\text{grasp,demo}} = (\mathbf{P}_{\text{obj}} \cdot \mathbf{P}_{\text{demo}}^{-1}) \cdot \mathbf{P}_{\text{grasp,demo}}. \tag{9}$$

The robot arm then executes a collision-free motion trajectory to reach $\mathbf{P}_{\text{grasp,novel}}$, ensuring goal-conditioned grasping without relying on task-irrelevant dense grasp proposals produced via 6-DoF grasp pose detectors [22, 23]. By integrating with the proposed decomposition stage, the adapted grasp pose directly modifies the variable block $\mathcal{B}_i$ into $\mathcal{B}_i'$, enabling synthesis of new trajectories.

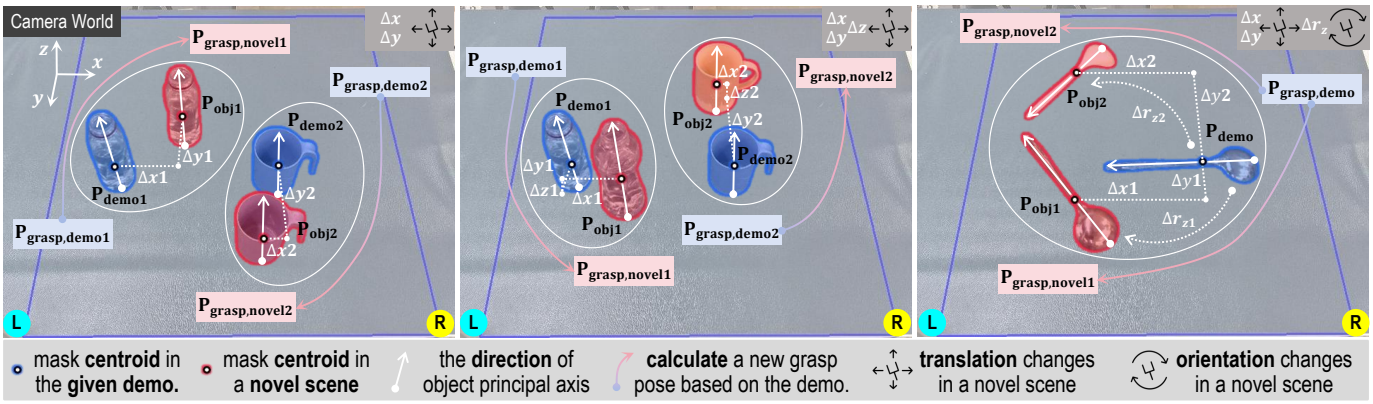

Fig. 3: Illustrations of the initial frame alignment applied to tasks `pouring` (left and middle) and `reorient` (right). It shows that we can automatically adjust the grasp pose after the position, orientation and shape of the manipulated object changes.

C. Trajectory Modulation and Optimization

After deconstructing the provided demonstration into blocks (Sec. IV-A) and adapting variable blocks via vision-based alignment (Sec. IV-B), the synthesized trajectory still requires two critical refinements to ensure executability and robustness as expalined below:

1) Collision-Aware Validation: Each adapted block undergoes kinematic feasibility checks. First, we validate the reachability of target poses $\mathbf{P}_{\text{grasp,novel}}$ using Inverse Kinematics (IK), ensuring the robot joint limits are satisfied. If the result returned is singular or unreachable (rarely happens), we will delete this sample. Second, a motion planner [19, 75] verifies potential collision between arms and scene objects by solving for collision-free paths between the start and end states in each block. This simplified two-point constraint reduces computational overhead while preserving safety, assuming the trajectory of original given demonstration is collision-free.

2) Instance-Level Motion Adaptation: To handle geometric variations across object instances with differing shapes, we adjust motion primitives in variable blocks based on the object 3D bounding box dimensions. Let Δl, Δw and Δh denote the length, width and height differences between the novel object bounding box $\mathbf{b}_{\text{novel}}$ and the original $\mathbf{b}_{\text{demo}}$. Motion endpoints (*e.g.*, grasp or release positions) are offset by:

$$\mathbf{s}_{i,e}^{A'} = \mathbf{s}_{i,e}^{A} + \lambda(\Delta l, \Delta w, \Delta h), \tag{10}$$

where λ scales the adjustment (empirically set to $0.8 \sim 1.0$), and $A \in \{L, R\}$. For irregular shapes, depth-aware masks M extracted in Sec. IV-B can approximate volumetric differences, or minimal human input provides precise measurements.

This hierarchical optimization stage operates at the block level rather than full trajectories, and enables efficient scaling while preserving the task semantics encoded in invariant blocks. Finally, validated and adapted blocks are recombined into synthetic trajectories $\tau' = \{\tau_{\text{inv}}, \{\mathcal{B}_i'\}\}$, populating diverse and physically feasible demonstrations $\mathcal{D}_{syn}$.

V. Experiments

Our experiments address five core questions. Q1: Is BiDemoSyn truly efficient and user-friendly? Q2: Does BiDemoSyn enable scalable visuomotor imitation learning? Q3: Does synthetic demonstrations generalize to spatial and object variations? Q4: Can BiDemoSyn seamlessly integrate with few-shot demonstrations to further improve robustness and data efficiency? Q5: Do policies trained on BiDemoSyn data transfer zero-shot across different robot embodiments while maintaining high success rates?

A. Experimental Setup and Protocol

Tasks. We evaluate on six bimanual manipulation tasks requiring contact-rich coordination: `plugpen`, `inserting`, `unscrew`, `pouring`, `pressing` and `reorient`. These tasks cover diverse primitive skills (*e.g.*, grasping, rotating and handover), and involve both rigid and articulated objects. The one-shot demonstration for each task is collected via kinesthetic teaching. For the platform, two fixed-base arms equipped with parallel grippers are used as the main workspace. An auxiliary humanoid dual-arm platform was also used to verify the cross-platform deployment capability of the learned policy. Scene perception is provided by a stereo camera capturing binocular images. More details of tasks, harwares and data collection are in ***Supplementary Materials***.

Policies and Baselines. Compared baselines contain two categories. One category is *purely for data collection* including point cloud editing (DemoGen [93]), real robot auto-rollout (YOTO [107]), and human drag teaching (close to Teleoperation). Generally, the quality of collected demonstrations by these baselines is better in turn, but the cost is more time-consuming. After preparing the training data, we adapt three advanced visuomotor policies (DP [16], DP3 [98] and EquiBot [94]) to bimanual settings by modifying their modeling spaces to dual-arm actions (refer Sec. III). Observation inputs are RGB-only images or segmented 3D point clouds of task-relevant objects. Final trained policies operate in the open-loop discrete keyposes prediction to align with our synthesized demonstration format. The another category is *directly for bimanual manipulation without retraining* including the zero-shot ReKep [38], an advanced ReKep+ with oracle-level grasp labels at the beginning, and two one-shot imitation learning methods ODIL [84] and MAGIC [60].

Metrics. We evaluate (1) *synthesis efficiency*: average time per demo over 50 tests, (2) *data quality*: visual authenticity

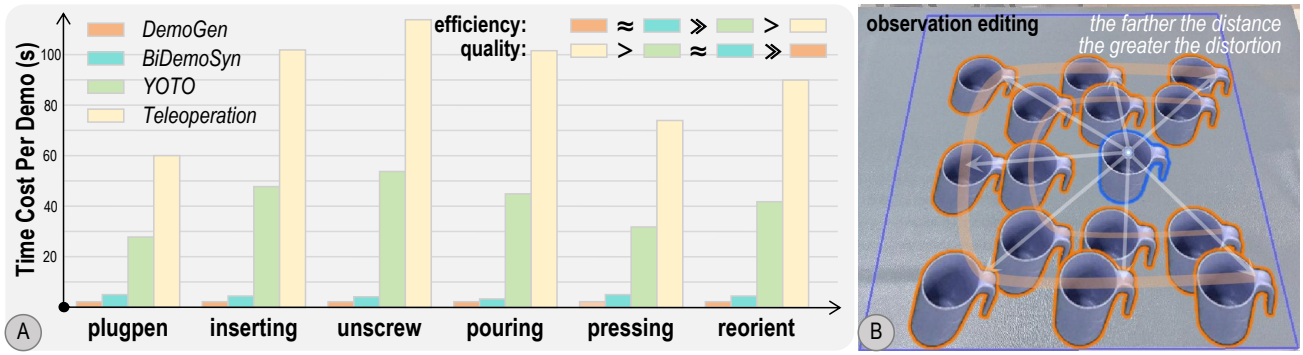

Fig. 4: (A) The data collection efficiency comparison of different baselines and our BiDemoSyn, and (B) the generated data quality illustration of DemoGen. Although DemoGen has the highest synthesis efficiency, it cannot avoid visual artifacts caused by perspective transformation, so its data quality is the lowest. Our method can achieve a good speed-quality balance.

TABLE I: Quantitative results of different methods under *in-distribution* (ID) and *out-of-distribution* (OOD) evaluations. The OOD means removing one or a pair of objects from the training set for testing purposes only. The # means the training size.

Policy (Input Mode)	Method	*in-distribution* (ID) evaluations							*out-of-distribution* (OOD) evaluations						
		plugpen #3888	inserting #7776	unscrew #1152	pouring #2592	pressing #1296	reorient #1008	Average Success Rate	plugpen #2916	inserting #3888	unscrew #1008	pouring #972	pressing #324	reorient #756	Average Success Rate
Training -Free (RGB)	ReKep	15/30	13/30	12/30	15/30	12/30	09/30	42.2%	12/30	10/30	10/30	09/30	09/30	06/30	31.1%
	ReKep+	17/30	17/30	14/30	19/30	20/30	11/30	54.4%	15/30	14/30	12/30	12/30	11/30	09/30	40.6%
	ODIL	18/30	19/30	13/30	18/30	18/30	12/30	54.4%	12/30	14/30	08/30	11/30	13/30	08/30	36.7%
	MAGIC	19/30	18/30	14/30	18/30	19/30	12/30	55.6%	13/30	14/30	10/30	12/30	13/30	09/30	39.4%
DP (RGB)	DemoGen	18/30	20/30	16/30	16/30	14/30	15/30	55.0%	08/30	03/30	10/30	04/30	03/30	07/30	19.4%
	YOTO	19/30	20/30	16/30	17/30	14/30	16/30	56.7%	08/30	04/30	10/30	05/30	04/30	07/30	21.1%
	BiDemoSyn	22/30	24/30	22/30	21/30	17/30	19/30	**67.8%**	13/30	10/30	18/30	15/30	08/30	12/30	**42.2%**
DP3 (PCD)	DemoGen	21/30	24/30	19/30	20/30	18/30	17/30	66.1%	11/30	06/30	14/30	07/30	05/30	11/30	30.0%
	YOTO	22/30	23/30	20/30	20/30	18/30	19/30	67.8%	12/30	06/30	15/30	08/30	06/30	14/30	33.9%
	BiDemoSyn	26/30	28/30	25/30	24/30	21/30	22/30	**81.1%**	16/30	14/30	21/30	18/30	11/30	18/30	**54.4%**
EquiBot (PCD)	DemoGen	22/30	24/30	20/30	20/30	19/30	19/30	68.9%	13/30	10/30	16/30	11/30	10/30	13/30	40.6%
	YOTO	23/30	24/30	20/30	21/30	20/30	19/30	71.1%	15/30	11/30	18/30	14/30	12/30	16/30	47.8%
	BiDemoSyn	28/30	28/30	26/30	25/30	24/30	25/30	**86.7%**	18/30	20/30	24/30	21/30	17/30	20/30	**66.7%**

of newly acquired observations, (3) *success rates*: real robot deployment with 30 trials per task, and (4) *generalization*: performance on unseen object placements and geometries.

B. Comparison and Presentation of Results

We here answer three questions raised earlier, demonstrating that BiDemoSyn can efficiently synthesize high-quality real-world demonstrations, enabling scalable and generalizable visuomotor policy training with minimal human input. Then there are some snapshots of real robot execution effects.

(A1): **The efficiency and usability of BiDemoSyn have obvious advantages over baselines**. Without bells and whistles, our BiDemoSyn demonstrates superior efficiency and usability as illustrated in Fig. 4. It requires *about 5 seconds per demonstration* to synthesize new trajectories, where primarily involving manual object repositioning and visual verification of initial frame alignment, followed by the trajectory optimization. In contrast, DemoGen generates trajectories in *less then 1 second* via scripted edits but suffers from perspective distortion artifacts, degrading data quality. The specific qualitative effect can be compared with Fig. 4A and our results in Fig. 2. By relying on manual alignment and auto-replay, YOTO takes *42 seconds per demo* which is equivalent to real robot execution time. While, teleoperation demands *91 seconds per demo* for near-perfect but labor-intensive collection. These results highlight BiDemoSyn's unique balance of speed (automated optimization) and reliability (artifact-free synthesis), making it the only method scalable to thousands of demonstrations without compromising real-world fidelity.

(A2): **Demonstrations obtained via BiDemoSyn can support scalable imitation learning**. We applied BiDemoSyn for efficient collection of thousands demonstrations per task, densely covering the workspace with instance-level object diversity, thereby providing abundant training data for imitation learning. For two reproduced baselines, DemoGen [93] matches our data scale but suffers from visual artifacts, while YOTO [107] collects only 1/10 the data per task for being limited by its time-consuming replay mechanism. After obtaining adequate data, we trained three advanced visuomotor policies DP [16], DP3 [98] and EquiBot [94]. As shown in Tab. I left, BiDemoSyn achieves superior *in-distribution* (ID) success rates across all six tasks, outperforming both baselines regardless of the policy architecture. This confirms that BiDemoSyn's artifact-free synthesis and efficient data scaling reliably support the scaling laws of imitation learning. Besides, it is unsurprising that policies trained with demonstrations consistently outperform three training-free baselines ReKep[38], ODIL [84], and MAGIC [60]. To further quantify scalability,

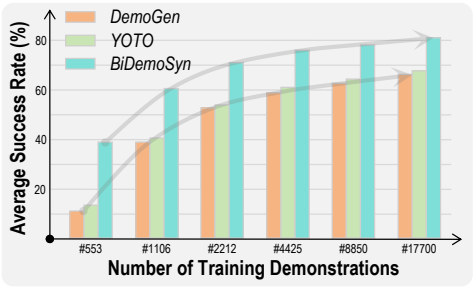

Fig. 5: Comparison between training scale and success rate. Less sized data is randomly sampled out of the total dataset at task-level. DP3 is chosen as the visuomotor policy.

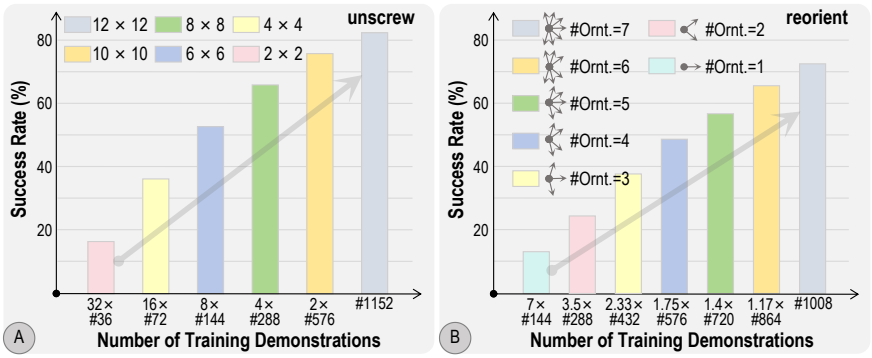

Fig. 6: Analysis of spatial generalization for variations in (A) position sampling density and (B) orientation sampling diversity. DP3 is chosen as the policy.

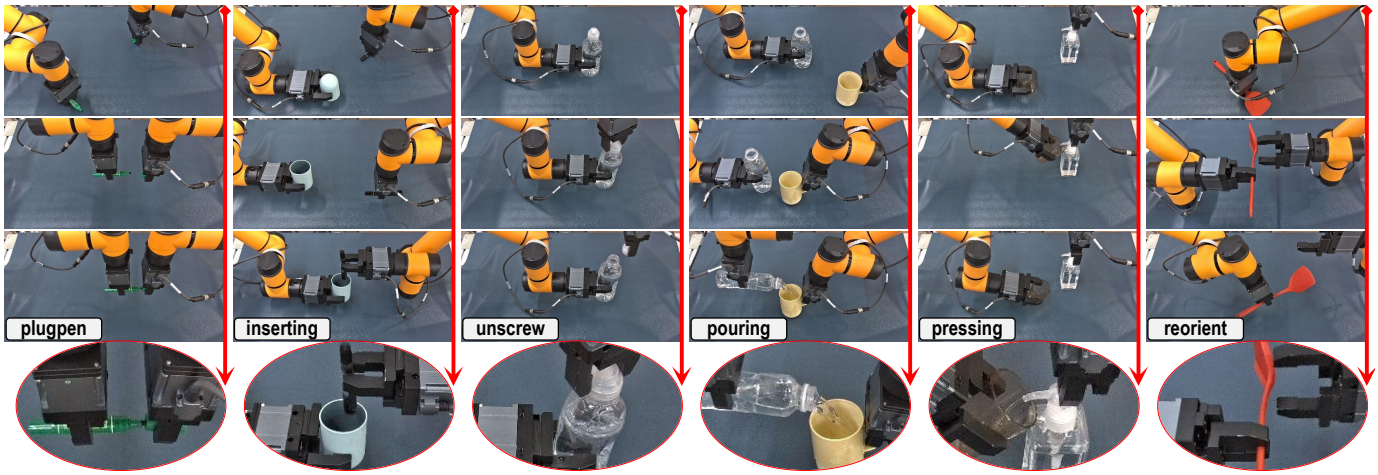

Fig. 7: Visualization of all six bimanual tasks performed on real robots. All models are trained and tested under the ID evaluations. EquiBot is chosen as the visuomotor policy. Key highlighted dual-arm coordination movements associated with each task are partially enlarged at the bottom row for quick review.

we evaluated training size-performance relationships on six tasks of all methods. As in Fig. 5, BiDemoSyn-trained policies exhibit stronger scaling trends than DemoGen (*e.g.*, 60% vs. 39% at 1106 demos), aligning with community observations that data quality and quantity jointly drive visuomotor policy performance [58, 5, 71, 53].

(A3): **Policies trained on BiDemoSyn data can achieve better generalization to unseen variations**. To evaluate the generalization to novel instances, we conduct controlled tests across all six tasks. For each task, we exclude demonstrations associated with one or a pair of randomly selected objects (*e.g.*, an arbitrary bottle in `unscrew` or a bottle-cup pair in `pouring`) from the training set, resulting in training sizes of 2916/3888/1008/972/324/756 demos for six tasks, respectively. These excluded objects are reintroduced exclusively during real-world testing, with identical data splits applied to DemoGen and YOTO baselines for fair comparison. As shown in Tab. I right, policies trained on BiDemoSyn data consistently achieve higher success rates compared to all baselines in *out-of-distribution* (OOD) settings. While all methods exhibit performance drops relative to ID evaluations, BiDemoSyn's superior OOD robustness highlights its ability to capture task-invariant features through vision-aligned synthesis. To further analyze spatial generalization, we vary the density of positional and orientation coverage in training data: for `unscrew`, we synthesize trajectories with sparse-to-dense workspace coverage (see Fig. 6A), and for `reorient`, we incrementally increase the diversity of object orientations (see Fig. 6B). To exclude the effect of data size differences, the total number of train-set is always augmented to a standard size (*e.g.*, `#1152` and `#1008`) via the DemoGen to edit some observations at small distances. Finally, policies trained on BiDemoSyn data with dense spatial coverage (*e.g.*, 10×10 workspace coverage in `unscrew`) achieve 76.7% success on unseen positions, compared to 53.3% for sparse coverage (6×6 grid cells). Similarly, orientation diversity in `reorient` improves generalization from 37.3% (lower diversity with `#Ornt.=3`) to 73.3% (higher diversity with `#Ornt.=7`). These results confirm that BiDemoSyn's ability to cheaply synthesize positionally and orientationally diverse demos directly translates to enhanced policy generalization, aligning with empirical insights that broad data coverage mitigates spatial distribution shifts.

Qualitative Results: Fig. 7 shows effectiveness across six tasks. For example, the `plugpen` policy precisely aligns and inserts the pen into the narrow cap even with $\pm$5mm positional noise. In `inserting`, dual-arm coordination achieves rough but proximate peg-in-hole precision. The `unscrew` demon-

TABLE II: Quantitative results for few-shot BiDemoSyn. Although the number of training demonstrations is the same for each setting, the number of seed demonstrations for trajectory synthesis is different. The # means the training size.

Trained Policy	Few-Shot BiDemoSyn	*in-distribution* (ID) evaluations							*out-of-distribution* (OOD) evaluations						
		plugpen	inserting	unscrew	pouring	pressing	reorient	Average Success Rate	plugpen	inserting	unscrew	pouring	pressing	reorient	Average Success Rate
		#3888	#7776	#1152	#2592	#1296	#1008		#2916	#3888	#1008	#972	#324	#756	
EquiBot	1-shot	28/30	28/30	26/30	25/30	24/30	25/30	86.7%	18/30	20/30	24/30	21/30	17/30	20/30	66.7%
	5-shot	28/30	28/30	26/30	26/30	25/30	25/30	87.8%	20/30	21/30	24/30	23/30	18/30	21/30	70.0%
	10-shot	28/30	28/30	26/30	27/30	25/30	26/30	88.9%	21/30	21/30	25/30	23/30	18/30	22/30	71.7%
	15-shot	28/30	28/30	27/30	27/30	26/30	26/30	90.0%	21/30	22/30	25/30	23/30	18/30	22/30	72.2%
	20-shot	28/30	28/30	27/30	27/30	26/30	26/30	90.0%	21/30	22/30	25/30	23/30	18/30	22/30	72.2%

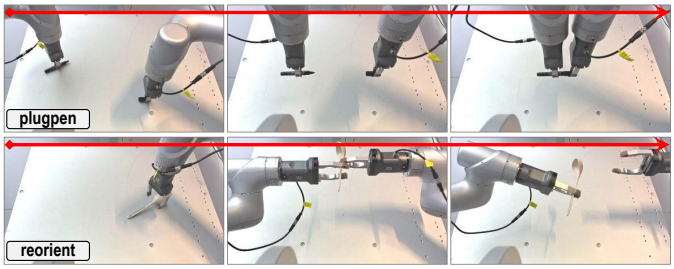

Fig. 8: Visualization for the cross-embodiment deployment of BiDemoSyn-trained policies on an unseen humanoid-style dual-arm robot for the `plugpen` and `reorient` tasks.

strates robust rotational synchronization, successfully loosening lids with variable thread tightness. The `pouring` policy adapts liquid flow control to container geometries, minimizing spills despite tilted orientations. For `pressing`, predicted actions reliably activate nozzles without over-pressuring, while `reorient` handles object handovers and flips with smooth dual-arm transitions. These results corroborate our quantitative findings, emphasizing BiDemoSyn's ability to handle real-world complexity in contact-rich cases.

(A4): **BiDemoSyn can seamlessly integrate with few-shot demonstrations to improve robustness**. Although BiDemoSyn is formulated in a one-shot paradigm, it can seamlessly incorporate additional real demonstrations (5–20 samples) as complementary seeds without altering the synthesis pipeline. We summarized this ablation results in Tab. II. Empirically, we can find that few-shot extensions provide limited gains under ID settings, where one-shot–generated datasets already yield near-saturated performance on seen settings. In contrast, moderate but consistent improvements are observed under OOD conditions, primarily due to expanded coverage of object geometries and intermediate contact states introduced by new demonstrations. These results suggest that BiDemoSyn's one-shot synthesis already produces high-quality, physically grounded demonstrations, while few-shot seeds mainly contribute by enriching representational diversity rather than correcting systematic failure modes. Overall, this study confirms that BiDemoSyn offers a favorable trade-off: it remains highly effective in the one-shot regime, benefits selectively from few-shot data in OOD scenarios, and supports hybrid one-shot + few-shot usage without increasing complexity.

(A5): **Policies trained on BiDemoSyn data transfer effectively across robot embodiments in a zero-shot manner**. BiDemoSyn facilitates zero-shot cross-embodiment policy deployment by design, owing to its object-centric observation formulation and the use of end-effector–centric 6-DoF action representation. These choices decouple policy learning from robot-specific kinematics and enable direct transfer to new hardware platforms. To evaluate this capability, we deploy visuomotor policies trained on the primary dual-arm platform to a previously unseen humanoid-style bimanual robot, without additional training or fine-tuning. We consider two representative tasks, `plugpen` and `reorient`, using EquiBot policies trained under ID settings. As shown in Fig. 8, the transferred policies successfully reproduce precise bimanual alignment for insertion and reorientation. Quantitative evaluations indicate that success rates on the new embodiment are comparable to those reported in Tab. I left. These results demonstrate that BiDemoSyn-trained policies exhibit strong embodiment-agnostic behavior, highlighting the potential for scalable deployment across heterogeneous robotic platforms.

VI. Conclusion and Limitation

We introduce BiDemoSyn, a real-world demonstration synthesis framework that enables scalable bimanual manipulation learning from a single kinesthetic example. By decomposing demonstrations into invariant coordination blocks and object-dependent adaptations, and instantiating them via vision-based object anchoring and trajectory optimization, BiDemoSyn generates large-scale, physically feasible demonstrations without simulation or repeated teleoperation. Across diverse bimanual tasks, policies trained on BiDemoSyn data achieve strong generalization to novel objects, long-horizon compositions, and unseen robot embodiments. We further show that the framework naturally extends to few-shot synthesis, improving robustness with minimal additional supervision. This work establishes a new paradigm for scalable imitation learning, bridging the divide between data quantity and physical fidelity in bimanual robotic manipulation.

Limitations: While BiDemoSyn excels in static, geometrically varied scenes, it faces challenges in fully dynamic environments and extreme shape variations (*e.g.*, highly deformable or articulated objects). The current hierarchical optimization assumes quasi-static motions, limiting its applicability to highly dynamic tasks like catching. Additionally, synthesizing trajectories for multi-stage tasks with interdependent contacts (*e.g.*, lacing ropes) requires further extension. Future work will integrate dynamic perception and adaptive contact modeling to address these constraints.

ACKNOWLEDGMENTS

This work was supported by the Key-Area Research and Development Program of Guangdong Province, China under Grant 2024B0101040004, the Shenzhen Science and Technology Program under Grant KJZD20240903104008012, and the Shenzhen Science and Technology Program under Grant ZDCY20250901113000001.

REFERENCES

[1] Arpit Bahety, Priyanka Mandikal, Ben Abbatematteo, and Roberto Martín-Martín. Screwmimic: Bimanual imitation from human videos with screw space projection. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[2] Homanga Bharadhwaj. Position: scaling simulation is neither necessary nor sufficient for in-the-wild robot manipulation. In *International Conference on Machine Learning*, 2024.

[3] Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation. In *European Conference on Computer Vision*, pages 306–324. Springer, 2024.

[4] Ondrej Biza, Skye Thompson, Kishore Reddy Pagidi, Abhinav Kumar, Elise van der Pol, Robin Walters, Thomas Kipf, Jan-Willem van de Meent, Lawson LS Wong, and Robert Platt. One-shot imitation learning via interaction warping. In *Conference on Robot Learning*, pages 2519–2536. PMLR, 2023.

[5] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[6] Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xu Huang, Shu Jiang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

[7] Alper Canberk, Cheng Chi, Huy Ha, Benjamin Burchfiel, Eric Cousineau, Siyuan Feng, and Shuran Song. Cloth funnels: Canonicalized-alignment for multi-purpose garment manipulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5872–5879. IEEE, 2023.

[8] François Chaumette. Image moments: a general and useful set of features for visual servoing. *IEEE Transactions on Robotics*, 20(4):713–723, 2004.

[9] Sirui Chen, Chen Wang, Kaden Nguyen, Li Fei-Fei, and C Karen Liu. Arcap: Collecting high-quality human demonstrations for robot learning with augmented reality feedback. *arXiv preprint arXiv:2410.08464*, 2024.

[10] Yuanpei Chen, Tianhao Wu, Shengjie Wang, Xidong Feng, Jiechuan Jiang, Zongqing Lu, Stephen McAleer, Hao Dong, Song-Chun Zhu, and Yaodong Yang. Towards human-level bimanual dexterous manipulation with reinforcement learning. *Advances in Neural Information Processing Systems*, 35:5150–5163, 2022.

[11] Yuanpei Chen, Yiran Geng, Fangwei Zhong, Jiaming Ji, Jiechuang Jiang, Zongqing Lu, Hao Dong, and Yaodong Yang. Bi-dexhands: Towards human-level bimanual dexterous manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):2804–2818, 2023.

[12] Yuanpei Chen, Chen Wang, Yaodong Yang, and Karen Liu. Object-centric dexterous manipulation from human motion data. In *8th Annual Conference on Robot Learning*, 2024.

[13] Zerui Chen, Shizhe Chen, Etienne Arlaud, Ivan Laptev, and Cordelia Schmid. Vividex: Learning vision-based dexterous manipulation from human videos. *arXiv preprint arXiv:2404.15709*, 2024.

[14] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024.

[15] Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-television: Teleoperation with immersive active visual feedback. In *8th Annual Conference on Robot Learning*, 2024.

[16] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.

[17] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*, 2024.

[18] Rohan Chitnis, Shubham Tulsiani, Saurabh Gupta, and Abhinav Gupta. Efficient bimanual manipulation using learned task schemas. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1149–1155. IEEE, 2020.

[19] Sachin Chitta, Ioan Sucan, and Steve Cousins. Moveit![ros topics]. *IEEE Robotics & Automation Magazine*, 19(1):18–19, 2012.

[20] Adria Colomé and Carme Torras. Dimensionality reduction for dynamic movement primitives and application to bimanual manipulation of clothes. *IEEE Transactions on Robotics*, 34(3):602–615, 2018.

[21] Kamil Dreczkowski, Pietro Vitiello, Vitalis Vosylius, and Edward Johns. Learning a thousand tasks in a day. *Science Robotics*, 10(108):eadv7594, 2025.

[22] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*

Recognition, pages 11444–11453, 2020.

[23] Hao-Shu Fang, Chenxi Wang, Hongjie Fang, Minghao Gou, Jirong Liu, Hengxu Yan, Wenhai Liu, Yichen Xie, and Cewu Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics*, 39(5):3929–3945, 2023.

[24] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 653–660. IEEE, 2024.

[25] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans. In *8th Annual Conference on Robot Learning*, 2024.

[26] Zipeng Fu, Tony Z Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation using low-cost whole-body teleoperation. In *8th Annual Conference on Robot Learning*, 2024.

[27] Jianfeng Gao, Xiaoshu Jin, Franziska Krebs, Noémie Jaquier, and Tamim Asfour. Bi-kvil: Keypoints-based visual imitation learning of bimanual manipulation tasks. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16850–16857. IEEE, 2024.

[28] Caelan Reed Garrett, Ajay Mandlekar, Bowen Wen, and Dieter Fox. Skillmimicgen: Automated demonstration generation for efficient skill learning and deployment. In *8th Annual Conference on Robot Learning*, 2024.

[29] Jennifer Grannen, Priya Sundaresan, Brijen Thananjeyan, Jeffrey Ichnowski, Ashwin Balakrishna, Vainavi Viswanath, Michael Laskey, Joseph Gonzalez, and Ken Goldberg. Untangling dense knots by learning task-relevant keypoints. In *Conference on Robot Learning*, pages 782–800. PMLR, 2021.

[30] Jennifer Grannen, Yilin Wu, Brandon Vu, and Dorsa Sadigh. Stabilize to act: Learning to coordinate for bimanual manipulation. In *Conference on Robot Learning*, pages 563–576. PMLR, 2023.

[31] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024.

[32] Markus Grotz, Mohit Shridhar, Tamim Asfour, and Dieter Fox. Peract2: Benchmarking and learning for robotic bimanual manipulation tasks. *arXiv preprint arXiv:2407.00278*, 2024.

[33] Valentin N Hartmann, Andreas Orthey, Danny Driess, Ozgur S Oguz, and Marc Toussaint. Long-horizon multi-robot rearrangement planning for construction assembly. *IEEE Transactions on Robotics*, 39(1):239–252, 2022.

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[35] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[36] Pu Hua, Minghuan Liu, Annabella Macaluso, Yunfeng Lin, Weinan Zhang, Huazhe Xu, and Lirui Wang. Gensim2: Scaling robot data generation with multi-modal and reasoning llms. In *8th Annual Conference on Robot Learning*, 2024.

[37] Binghao Huang, Yuanpei Chen, Tianyu Wang, Yuzhe Qin, Yaodong Yang, Nikolay Atanasov, and Xiaolong Wang. Dynamic handover: Throw and catch with bimanual hands. In *Conference on Robot Learning*, pages 1887–1902. PMLR, 2023.

[38] Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. In *8th Annual Conference on Robot Learning*, 2024.

[39] Stephen James, Kentaro Wada, Tristan Laidlow, and Andrew J Davison. Coarse-to-fine q-attention: Efficient learning for visual robotic manipulation via discretisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13739–13748, 2022.

[40] Zhenyu Jiang, Yuqi Xie, Kevin Lin, Zhenjia Xu, Weikang Wan, Ajay Mandlekar, Linxi Fan, and Yuke Zhu. Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning. *arXiv preprint arXiv:2410.24185*, 2024.

[41] Yuanchen Ju, Kaizhe Hu, Guowei Zhang, Gu Zhang, Mingrun Jiang, and Huazhe Xu. Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation. In *European Conference on Computer Vision*, pages 222–239. Springer, 2024.

[42] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. *arXiv preprint arXiv:2410.24221*, 2024.

[43] Justin Kerr, Chung Min Kim, Mingxuan Wu, Brent Yi, Qianqian Wang, Ken Goldberg, and Angjoo Kanazawa. Robot see robot do: Imitating articulated object manipulation with monocular 4d reconstruction. In *8th Annual Conference on Robot Learning*, 2024.

[44] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. In *Proceedings of Robotics: Science and Systems (RSS)*,

2024.

[45] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. Openvla: An open-source vision-language-action model. In *8th Annual Conference on Robot Learning*, 2024.

[46] Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B Tenenbaum. Learning to act from actionless videos through dense correspondences. In *The Twelfth International Conference on Learning Representations*, 2024.

[47] Leonidas Kotoulas and Ioannis Andreadis. Accurate calculation of image moments. *IEEE Transactions on Image Processing*, 16(8):2028–2037, 2007.

[48] Franziska Krebs and Tamim Asfour. A bimanual manipulation taxonomy. *IEEE Robotics and Automation Letters*, 7(4):11031–11038, 2022.

[49] Chengshu Li, Mengdi Xu, Arpit Bahety, Hang Yin, Yunfan Jiang, Huang Huang, Josiah Wong, Sujay Garlanka, Cem Gokmen, Ruohan Zhang, et al. Momagen: Generating demonstrations under soft and hard constraints for multi-step bimanual mobile manipulation. *arXiv preprint arXiv:2510.18316*, 2025.

[50] Jinhan Li, Yifeng Zhu, Yuqi Xie, Zhenyu Jiang, Mingyo Seo, Georgios Pavlakos, and Yuke Zhu. Okami: Teaching humanoid robots manipulation skills through single video imitation. In *8th Annual Conference on Robot Learning*, 2024.

[51] Yunfei Li, Chaoyi Pan, Huazhe Xu, Xiaolong Wang, and Yi Wu. Efficient bimanual handover and rearrangement via symmetry-aware actor-critic learning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3867–3874. IEEE, 2023.

[52] Yongyuan Liang, Tingqiang Xu, Kaizhe Hu, Guangqi Jiang, Furong Huang, and Huazhe Xu. Make-an-agent: A generalizable policy network generator with behavior-prompted diffusion. *Advances in Neural Information Processing Systems*, 2024.

[53] Fanqi Lin, Yingdong Hu, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. Data scaling laws in imitation learning for robotic manipulation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[54] Toru Lin, Zhao-Heng Yin, Haozhi Qi, Pieter Abbeel, and Jitendra Malik. Twisting lids off with two hands. In *8th Annual Conference on Robot Learning*, 2024.

[55] Toru Lin, Yu Zhang, Qiyang Li, Haozhi Qi, Brent Yi, Sergey Levine, and Jitendra Malik. Learning visuotactile skills with two multifingered hands. *arXiv preprint arXiv:2404.16823*, 2024.

[56] I-Chun Arthur Liu, Sicheng He, Daniel Seita, and Gaurav S Sukhatme. Voxact-b: Voxel-based acting and stabilizing policy for bimanual manipulation. In *8th Annual Conference on Robot Learning*, 2024.

[57] Junjia Liu, Yiting Chen, Zhipeng Dong, Shixiong Wang, Sylvain Calinon, Miao Li, and Fei Chen. Robot cooking with stir-fry: Bimanual non-prehensile manipulation of semi-fluid objects. *IEEE Robotics and Automation Letters*, 7(2):5159–5166, 2022.

[58] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[59] Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21740–21751, 2024.

[60] Yuyao Liu, Jiayuan Mao, Joshua B Tenenbaum, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. One-shot manipulation strategy learning by making contact analogies. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 15387–15393. IEEE, 2025.

[61] Xiao Ma, Sumit Patidar, Iain Haughton, and Stephen James. Hierarchical diffusion policy for kinematics-aware multi-task robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18081–18090, 2024.

[62] Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations. In *Conference on Robot Learning*, pages 1820–1864. PMLR, 2023.

[63] Jiayuan Mao, Tomás Lozano-Pérez, Joshua B Tenenbaum, and Leslie Pack Kaelbling. Learning reusable manipulation strategies. In *Conference on Robot Learning*, pages 1467–1483. PMLR, 2023.

[64] Robert McCarthy, Daniel CH Tan, Dominik Schmidt, Fernando Acero, Nathan Herr, Yilun Du, Thomas G Thuruthel, and Zhibin Li. Towards generalist robot learning from internet video: A survey. *arXiv preprint arXiv:2404.19664*, 2024.

[65] Soroush Nasiriany, Sean Kirmani, Tianli Ding, Laura Smith, Yuke Zhu, Danny Driess, Dorsa Sadigh, and Ted Xiao. Rt-affordance: Affordances are versatile intermediate representations for robot manipulation. *arXiv preprint arXiv:2411.02704*, 2024.

[66] Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[67] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[68] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri,

Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.

[69] Georgios Papagiannis, Norman Di Palo, Pietro Vitiello, and Edward Johns. R+x: Retrieval and execution from everyday human videos. *arXiv preprint arXiv:2407.12957*, 2024.

[70] Weikun Peng, Jun Lv, Yuwei Zeng, Haonan Chen, Siheng Zhao, Jichen Sun, Cewu Lu, and Lin Shao. Tiebot: Learning to knot a tie from visual demonstration through a real-to-sim-to-real approach. In *8th Annual Conference on Robot Learning*, 2024.

[71] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025.

[72] Georgy Ponimatkin, Martin Cífka, Tomáš Souček, Médéric Fourmy, Yann Labbé, Vladimir Petrik, and Josef Sivic. 6d object pose tracking in internet videos for robotic manipulation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[73] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems*, 30, 2017.

[74] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[75] John Schulman, Yan Duan, Jonathan Ho, Alex Lee, Ibrahim Awwal, Henry Bradlow, Jia Pan, Sachin Patil, Ken Goldberg, and Pieter Abbeel. Motion planning with sequential convex optimization and convex collision checking. *The International Journal of Robotics Research*, 33(9):1251–1270, 2014.

[76] Kenneth Shaw, Yulong Li, Jiahui Yang, Mohan Kumar Srirama, Ray Liu, Haoyu Xiong, Russell Mendonca, and Deepak Pathak. Bimanual dexterity for complex tasks. In *8th Annual Conference on Robot Learning*, 2024.

[77] Modi Shi, Li Chen, Jin Chen, Yuxiang Lu, Chiming Liu, Guanghui Ren, Ping Luo, Di Huang, Maoqing Yao, and Hongyang Li. Is diversity all you need for scalable robotic manipulation? *arXiv preprint arXiv:2507.06219*, 2025.

[78] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR, 2023.

[79] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[80] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[81] Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yoloe: Real-time seeing anything. *arXiv preprint arXiv:2503.07465*, 2025.

[82] Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C Karen Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[83] Lirui Wang, Yiyang Ling, Zhecheng Yuan, Mohit Shridhar, Chen Bao, Yuzhe Qin, Bailin Wang, Huazhe Xu, and Xiaolong Wang. Gensim: Generating robotic simulation tasks via large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[84] Yilong Wang and Edward Johns. One-shot dual-arm imitation learning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025.

[85] Yufei Wang, Zhou Xian, Feng Chen, Tsun-Hsuan Wang, Yian Wang, Katerina Fragkiadaki, Zackory Erickson, David Held, and Chuang Gan. Robogen: Towards unleashing infinite data for automated robot learning via generative simulation. In *International Conference on Machine Learning*, pages 51936–51983. PMLR, 2024.

[86] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17868–17879, 2024.

[87] Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[88] Thomas Weng, Sujay Man Bajracharya, Yufei Wang, Khush Agrawal, and David Held. Fabricflownet: Bimanual cloth manipulation with a flow-based policy. In *Conference on Robot Learning*, pages 192–202. PMLR, 2022.

[89] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024.

[90] Fan Xie, Alexander Chowdhury, M De Paolis Kaluza, Linfeng Zhao, Lawson Wong, and Rose Yu. Deep imitation learning for bimanual robotic manipulation. *Advances in Neural Information Processing Systems*, 33:2327–2337, 2020.

[91] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin

Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21919–21928, 2023.

[92] Mengda Xu, Han Zhang, Yifan Hou, Zhenjia Xu, Linxi Fan, Manuela Veloso, and Shuran Song. Dexumi: Using human hand as the universal manipulation interface for dexterous manipulation. *arXiv preprint arXiv:2505.21864*, 2025.

[93] Zhengrong Xue, Shuying Deng, Zhenyang Chen, Yixuan Wang, Zhecheng Yuan, and Huazhe Xu. Demogen: Synthetic demonstration generation for data-efficient visuomotor policy learning. In *Proceedings of Robotics: Science and Systems (RSS)*, 2025.

[94] Jingyun Yang, Ziang Cao, Congyue Deng, Rika Antonova, Shuran Song, and Jeannette Bohg. Equibot: Sim (3)-equivariant diffusion policy for generalizable and data efficient learning. In *8th Annual Conference on Robot Learning*, 2024.

[95] Jingyun Yang, Congyue Deng, Jimmy Wu, Rika Antonova, Leonidas Guibas, and Jeannette Bohg. Equivact: Sim (3)-equivariant visuomotor policies beyond rigid object manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9249–9255. IEEE, 2024.

[96] Yandan Yang, Baoxiong Jia, Peiyuan Zhi, and Siyuan Huang. Physcene: Physically interactable 3d scene synthesis for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16262–16272, 2024.

[97] Weirui Ye, Fangchen Liu, Zheng Ding, Yang Gao, Oleh Rybkin, and Pieter Abbeel. Video2policy: Scaling up manipulation tasks in simulation through internet videos. *arXiv preprint arXiv:2502.09886*, 2025.

[98] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[99] Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 445–456, 2024.

[100] Tong Zhang, Yingdong Hu, Hanchen Cui, Hang Zhao, and Yang Gao. A universal semantic-geometric representation for robotic manipulation. In *Conference on Robot Learning*, pages 3342–3363. PMLR, 2023.

[101] Xinyu Zhang and Abdeslam Boularias. One-shot imitation learning with invariance matching for robotic manipulation. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[102] Hongxiang Zhao, Xingchen Liu, Mutian Xu, Yiming Hao, Weikai Chen, and Xiaoguang Han. Taserob: Advancing video generation of task-oriented hand-object interaction for generalizable robotic manipulation. *arXiv preprint arXiv:2503.11423*, 2025.

[103] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[104] Tony Z Zhao, Jonathan Tompson, Danny Driess, Pete Florence, Seyed Kamyar Seyed Ghasemipour, Chelsea Finn, and Ayzaan Wahid. Aloha unleashed: A simple recipe for robot dexterity. In *8th Annual Conference on Robot Learning*, 2024.

[105] Yan Zhao, Ruihai Wu, Zhehuan Chen, Yourong Zhang, Qingnan Fan, Kaichun Mo, and Hao Dong. Dualafford: Learning collaborative visual affordance for dual-gripper manipulation. In *The Eleventh International Conference on Learning Representations*, 2023.

[106] Huayi Zhou and Kui Jia. Vlbiman: Vision-language anchored one-shot demonstration enables generalizable bimanual robotic manipulation. In *The Fourteenth International Conference on Learning Representations*, 2026.

[107] Huayi Zhou, Ruixiang Wang, Yunxin Tai, Yueci Deng, Guiliang Liu, and Kui Jia. You only teach once: Learn one-shot bimanual robotic manipulation from video demonstrations. In *Proceedings of Robotics: Science and Systems (RSS)*, 2025.

[108] Huayi Zhou, Ruixiang Wang, Yunxin Tai, Yueci Deng, Guiliang Liu, and Kui Jia. Yoto++: Learning long-horizon closed-loop bimanual manipulation from one-shot human video demonstrations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

APPENDIX

This supplementary provides detailed explanations and extensions to the main paper. Sec. VII outlines the design rationale of six bimanual manipulation tasks, including platform specifications, task-specific objectives, and one-shot demonstration acquisition protocols. Sec. VIII details the data collection standards, UI-assisted collection tools, and post-processing workflows. These constitute the trajectory synthesis pipelines for generating real-world demonstrations with BiDemoSyn. We have also added more details about BiDemoSyn to help to understand it. Sec. IX elaborates on policy training, covering implementation specifics for baselines (DemoGen [93], YOTO [107]) and diffusion policies (DP3 [98], EquiBot [94]), alongside real-robot deployment. Sec. X provides additional visualizations of real-robot executions and failure case analyses, offering insights into current limitations. Sec. XI concludes with reflections and future directions, aiming to extend BiDemoSyn to challenge more complex tasks and enhance cross-domain generalization.

VII. DESIGN OF BIMANUAL MANIPULATION TASKS

A. Hardware and Platform

Our main experimental platform comprises a rectangular workspace (110cm×70cm) with two 6-DoF AUBO-i5 collaborative arms[1] (880mm reach) mounted on opposite short edges of the table (refer Fig. 9 top). This opposing-arm configuration maximizes shared workspace while minimizing self-collision risks, albeit differing from anthropomorphic designs. Each arm is equipped with a DH-Robotics parallel gripper[2] (80mm max opening, 50mm effective length), controlled in binary states (open/closed). Tool length compensation accounts for 160mm absolute length of the gripper. The scene perception is provided by a binocular stereo Kingfisher R-6000 camera (960×540 RGB resolution), mounted 100cm above the table long edge to capture a third-person view of the workspace. The calibrated stereo setup reconstructs high-fidelity 3D point clouds [91], eliminating the need for wrist-mounted cameras while ensuring full task visibility. To further demonstrate the cross-embodiment transferability of trained policies based on BiDemoSyn, as shown in Fig. 9 bottom, we have prepared another new dual-arm robotic platform configured in a popular humanoid style. This new platform consists of two Rokae xMate CR7[3] 6-DoF collaborative arms (reach: 988 mm), each equipped with a parallel gripper (Jodell Robotics RG75-300[4], max opening: 75 mm). A binocular camera Kingfisher R-6000 is mounted centrally at the head position. We will present how to utilize this dual-arm platform in Sec. X-C to deploy the learned models and perform real robot testing.

[1]https://www.aubo-cobot.com/public/i5product3
[2]https://en.dh-robotics.com/product/pg
[3]https://www.rokae.com/en/product/show/545/xMateCR.html
[4]https://www.jodell-robotics.com/product-detail?id=5

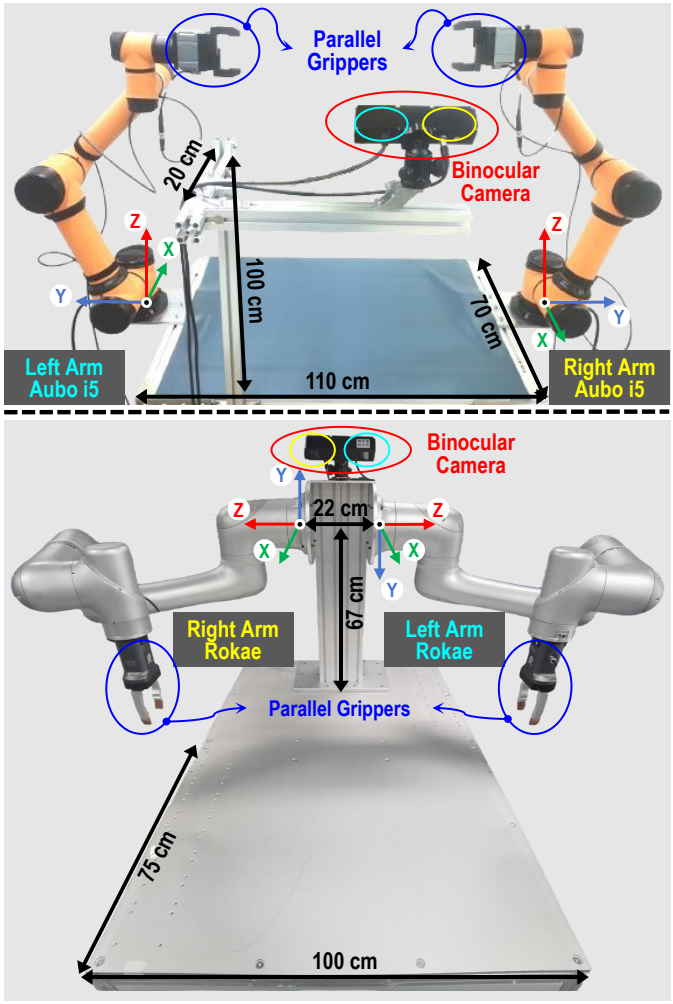

Fig. 9: *(Top) Primary*. The fixed-base dual-arm platform having a table with two robot arms, two grippers and the binocular camera. *(Bottom) Auxiliary*. The humanoid dual-arm platform for validating cross-embodiment capabilities.

B. Bimanual Task Formulation

We design six bimanual tasks (`plugpen`, `inserting`, `unscrew`, `pouring`, `pressing` and `reorient`) to comprehensively evaluate dual-arm coordination and generalization. Each task involves at least two category-level object instances (Fig. 10) and requires diverse primitive skills, such as single-arm actions (grasping, placing, precision rotation) and dual-arm coordination (fixed-twisting, plug-in, handover). All tasks are inherently **long-horizon**, starting from goal-conditioned initial grasping (objects fully separated from robots), contrasting prior works already grasping/holding the manipulated object and focusing on short-horizon atomic skills such as untwisting [54, 1] or handover [37]. Fig. 11 shows some examples. Below we detail each task:

- `plugpen`: *plug the pen with the pen cap*. **States**: A marker pen body and its cap are placed separately on the table. The pen body lies horizontally with its tip roughly pointing to the right arm, while the cap is also positioned horizontally with its socket facing the left arm. **Steps**:

Fig. 10: Object assets involved in six bimanual manipulation tasks. All objects have been scaled down proportionally.

Left arm grasps pen body; right arm grasps cap. Arms lift and spatially align the pen tip with the cap's socket before insertion. **Challenges**: Directional grasping alignment, sub-millimeter positional tolerance for insertion.

- `inserting`: *insert the marker pen into the cup*. **States**: An empty handle-free cup is placed upside-down on the table, and a closed marker pen lies horizontally nearby. **Steps**: Left arm flips the cup (about 180° rotation); right arm reorients the marker vertically. Arms coordinate to align and insert the marker into the upright cup. **Challenges**: Dual-arm rotational synchronization, tight insertion tolerance ($\pm$2mm).
- `unscrew`: *open the bottle by twisting cap counterclockwise*. **States**: A translucent plastic bottle (filled with colorless water) stands upright on the table, with its cap tightly screwed on. **Steps**: Left arm stabilizes the bottle; right arm grasps the cap and rotates it vertically (with multiple degree-fixed rotations). **Challenges**: Force-agnostic cap grasping (no torque sensing), rotational precision ($\pm 5^\circ$ per turn).
- `pouring`: *pour water from the bottle into the mug cup*. **States**: An uncapped water bottle (about $3/4$ full) and an empty handled mug are placed on opposite sides of the table. **Steps**: Left arm tilts the bottle (about 90°); right arm positions the mug to catch water. **Challenges**: Fluid dynamics approximation, nozzle-cup alignment under gripper deflection.
- `pressing`: *press bottle and catch water using the cup*. **States**: A shampoo bottle (with a pressable nozzle) containing water and an upright empty handle-free cup are placed apart. **Steps**: Right arm vertically presses the nozzle; left arm tilts the cup to catch water. **Challenges**: Nozzle positioning accuracy ($\pm$5mm), open-loop fine-grained force control.
- `reorient`: *flip the spoon so that the bottom is facing up*. **States**: A long spoon or shovel (metal or plastic) lies concave-down on the table, with randomized orientation and position. **Steps**: Right arm grasps the spoon center, reorients it mid-air, and hands it to the left arm, which then flips and places it. **Challenges**: Unstable grasp during handover, rotational precision for flipping ($\pm 10^\circ$).

These tasks collectively stress *spatial reasoning*, *contact-aware coordination*, and *generalization* to *instance-level variations*, which are cornerstones of real-world bimanual manipulation. Besdies, it is important to note that, we have **shared hardware constraints**: (1) No force/torque sensing limits contact-rich adjustments (*e.g.*, pen plug-in force, cap twisting force, or pressing force). (2) Binary gripper states (open/closed) restrict fine-grained manipulation (*e.g.*, plastic bottles are always deformed by excessive clamping). (3) Third-person view occlusions occasionally hinder precise alignment (*e.g.*, cannot discern the groove side of the pen cap).

C. Obtaining and Decomposing the One-Shot Demonstration

As stated in the main content, we collect one-shot demonstrations via kinesthetic teaching: an operator manually guides both arms through task-critical waypoints, recording the 6DoF end-effector poses (relative to each robot base frame) and gripper binary states at each pause. Objects are placed in fixed initial configurations (allowing small positional tolerance) to ensure consistency. The recorded waypoints are then executed autonomously by the robot control API, which solves inverse kinematics between consecutive given poses and synchronizes gripper actions (*e.g.*, closing after reaching a pre-grasp pose). During auto-execution, stereo camera observations (10Hz) and dual-arm joint/end-effector states are logged. For the real rollout effect of the one-shot demonstration related to each task, please refer to our **Supplementary Videos**. Finally, demonstrations are deconstructed into task-aware blocks to support subsequent trajectory synthesis.

To enable reliable downstream synthesis, the one-shot kinesthetic demonstration must be collected in a way that naturally exposes task-relevant structure. During kinesthetic teaching, the demonstrator physically guides the two manipulators through the task while pausing at intuitive semantic boundaries, such as after a stable grasp, after completing an alignment adjustment, or before switching from transportation to interaction. These pauses, which arise organically from human motion patterns, produce clear temporal discontinuities that align well with the block-based representation introduced in the main paper. Importantly, this process *does not require any manual annotation or labeling*. The demonstration is captured in a continuous stream exactly as in standard kinesthetic teleoperation, and the boundaries between coarse blocks emerge directly from the dynamics of human-guided motion. This makes the procedure no more labor-intensive than collecting a single ordinary demonstration.

We explored alternative strategies for fully automatic segmentation, such as keyframe/keypose detection and curvature-based change-point identification [39, 78, 61]. However, reliably identifying semantically meaningful waypoints from a single demonstration proved brittle, especially when waypoints must correspond to subtle but crucial phases of contact establishment, re-grasping, or dual-arm synchronization. Hardware-assisted approaches (e.g., instrumented gloves as in DexCap [82] or end-effector teaching interfaces as in UMI [17] and DexUMI [92]) can offer additional cues for segmentation but

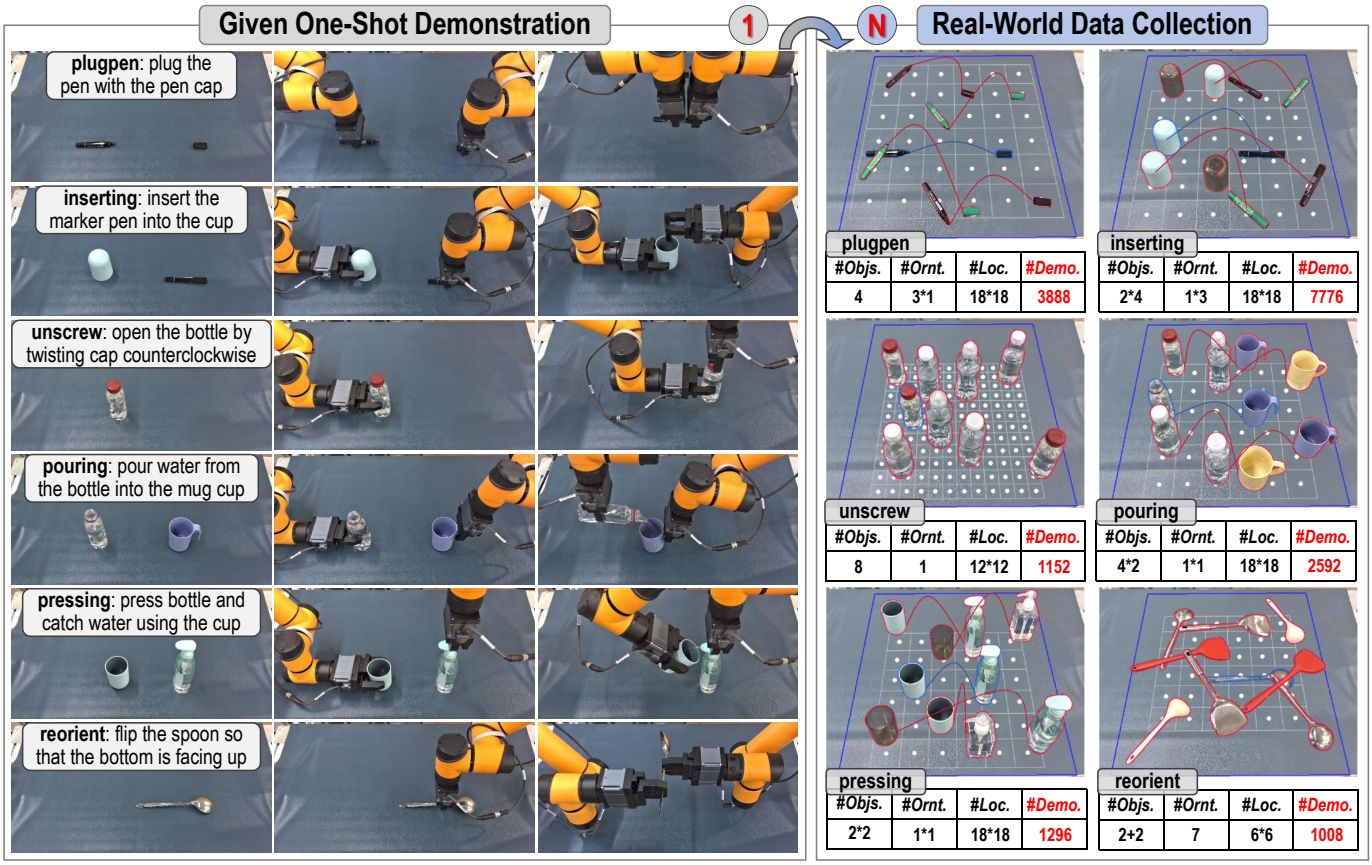

Fig. 11: **From One to Many** ①→Ⓝ. *Left*: Six representative bimanual manipulation tasks with their one-shot demonstrations and task-specific descriptors. *Right*: Real-world data collection diagrams, showing object instances with varied geometries and spatial arrangements used to synthesize diverse demonstrations (*e.g.*, thousands physically consistent trajectories per task).

introduce alignment overhead, cross-morphology calibration challenges, and potential loss of semantic correspondence between human motion and robot embodiment. Given these trade-offs, the kinesthetic-teaching-driven segmentation remains a pragmatic and robust choice: it imposes no additional annotation burden, preserves semantic alignment across blocks, and provides boundaries well suited for constructing invariant and variant components for real-world synthesis.

VIII. Details of Data Collection and Process

A. Task-Oriented Collection Standards

To systematically study factors influencing policies, we establish task-specific data collection protocols that ensure spatial uniformity and instance balance. Each task defines a common workspace (blue quadrilateral in Fig. 12, mapped to a 615mm×675mm rectangular area on the table) where objects are placed. For single-object tasks (*e.g.*, `unscrew`), objects are positioned across grid cells (white dots/boxes in Fig. 12) covering the entire workspace. For dual-object tasks, the workspace is split into left-right regions (*e.g.*, bottles occupy the left half, mugs the right half in `pouring`), with objects distributed uniformly within their zones. Object orientations are randomized for `plugpen`, `inserting`, and `reorient` but fixed for `unscrew`, `pouring`, and `pressing` (see Fig. 1 in the main paper for orientation counts). A grid cell is marked as collected if its centroid hosts a demonstration, allowing local variations with positional tolerance. This protocol balances coverage and practicality, enabling controlled studies on spatial and instance-level generalization.

B. Collection UI Interface Development

To streamline data collection, we develop a real-time UI interface (Fig. 13) that visualizes task-specific grid cells, dynamically tracks object positions/poses, and validates alignment with the one-shot demonstration. The interface overlays grid layouts onto live camera feeds, highlights detected objects (using YOLOE [81] for fast detection and segmentation), and marks completed cells using red crosses. The position of objects is represented by red dots (some objects also use yellow arrows to indicate their orientation). In addition, for cups and bottles that do not need to pay attention to changes in orientation, we did not use the mask centroid as their representive point during collection, but instead used the lowest endpoint in contact with the desktop. A sidebar displays positional/orientational deviations from the reference demonstration, enabling quick verification. If object perception fails, the system falls back to Florence2 [89] and SAM2 [74] for robust detection and segmentation. Operators trigger a *capture-align-validate-save* cycle every 2 seconds. And successful saves prompt visual feedback (UI flash of corre-

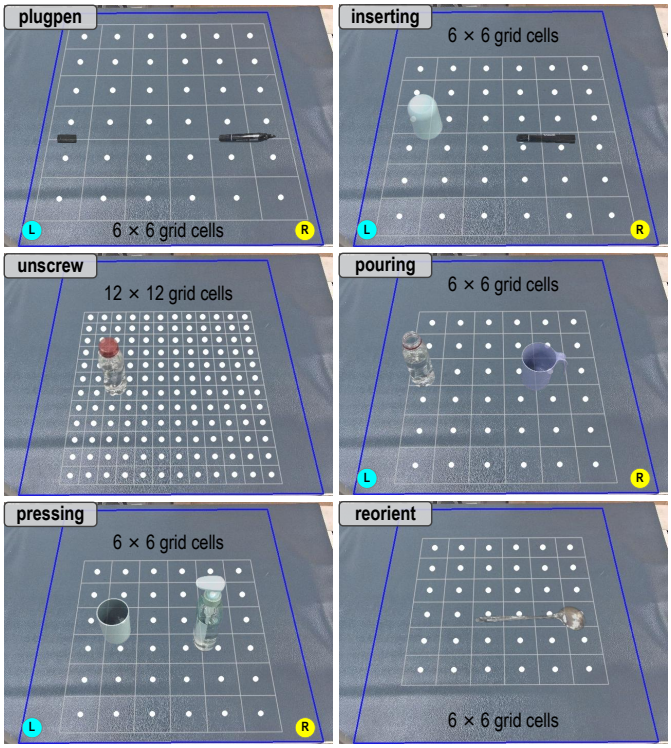

Fig. 12: The grid cell division for each task. For tasks involving two manipulated objects, the total number of grid cells will be divided equally between the left and right side.

sponding grid cells), signaling readiness for the next sample. This setup eliminates manual grid marking and minimizes human idle time, achieving a high speedup while ensuring spatial consistency. Please refer our **Supplemental Videos** that demonstrate the vivid workflow, illustrating how the UI balances efficiency with precision.

C. Post-Processing of Collected Data

This part includes the last two stages of BiDemoSyn: initial frame alignment and trajectory optimization. First, pre-recorded positional/orientational deviations (validated via the UI interface during collection) align novel observations with the one-shot demonstration reference frame using the rigid transformation. Next, trajectory optimization ensures kinematic feasibility. We prune singular configurations and validate collision-free paths using the robot IK solvers and motion planners. Over 96% of trajectories pass validation, with failures (*e.g.*, unreachable corner grid cells) either re-collected or discarded. Finally, validated trajectories are stored in $\mathcal{D}_{syn}$ to complete the synthesis pipeline.

D. Hyperparameters and Predefined Symbols

In the method section, several thresholds and parameters are introduced for clarity and reproducibility. Here we further explain their meaning and selection strategy. For instance, threshold symbols δ, ζ and γ in Eqn. 4, Eqn. 5 and Eqn. 10 are predefined only for one-shot deconstruction. The symbol δ is a minimal motion saliency threshold, used to identify boundaries between coarse blocks. The symbol ζ is a static tolerance threshold to segment motion boundaries under quasi-static assumptions. The symbol γ is a refinement threshold used to create fine-grained AEPs after bounding block duration ($T_{\max}$). And the scaling factor λ in Eqn. 10 controls the adaptation from the canonical one-shot demonstration to new object instances with varying sizes. Theoretically, $\lambda \in \mathbb{R}^{+}$, yet empirical tuning shows that values within $[0.8, 1.0]$ are sufficient for stable grasp transfer without destabilizing contact. Larger or smaller values tend to shift grasp points away from functional regions (e.g., bottle mid-section in `unscrew`), while $\lambda = 1.0$ already suffices for most tasks. Similarly, thresholds for collision-checking or pose-alignment rejection (e.g., rejecting infeasible IK solutions or near-singular samples) were determined through small-scale grid search and validated in pilot runs. Importantly, such hyperparameters are not per-trial tunable knobs, but rather fixed system constants, ensuring consistent data synthesis across tasks.

E. Evaluation of Intermediate Processes

We provide further evaluation to justify intermediate design choices from three aspects: Segmentation Quality, Pose Estimation Robustness and Impact on Data Reliability. (1) Firstly, object masks were obtained from Florence2 [89] and SAM2 [74], two state-of-the-art vision foundation models (VFMs). On held-out captures without foreground occlusion, the segmentation accuracy exceeded 97%, ensuring reliable geometric inputs for subsequent processing. (2) Secondly, the axis alignment for objects with a distinct orientation was based on classical image-moment theory: first-order moments define centroids, second-order moments define orientation. This method, given reliable masks, yields highly stable results. Quantitative inspection of $\sim$200 sampled trials revealed $<$2% pose estimation errors, and only under severe occlusion. (3) Thirdly, in Sec. X-D, we showed that perception and orientation errors together account for $\sim$47% of failures, highlighting their importance. Nevertheless, these intermediate modules are already close to optimal under current open-loop constraints, which explains the high success rate ($\sim$98%) of the data collection system.

F. Positioning with Respect to Prior Works

To avoid potential confusion with prior related works such as ReKep [38], ODIL [84] and MAGIC [60], we emphasize both the similarities and the fundamental differences. Like these methods, BiDemoSyn exploits one-shot demonstrations and compositional reasoning; however, *its core objective is not direct policy execution*, but rather *rapid and scalable synthesis of real-world training demonstrations*. This distinction is crucial: while prior methods target zero-/few-shot policy generalization (often trading reliability for flexibility), BiDemoSyn deliberately focuses on maximizing data quality and efficiency, enabling downstream training of robust visuomotor policies (e.g., DP3 [98] and EquiBot [94]). Our evaluation against DemoGen [93] and YOTO [107], which are two strong baselines addressing the same bottleneck, demonstrates that BiDemoSyn achieves a superior balance between efficiency

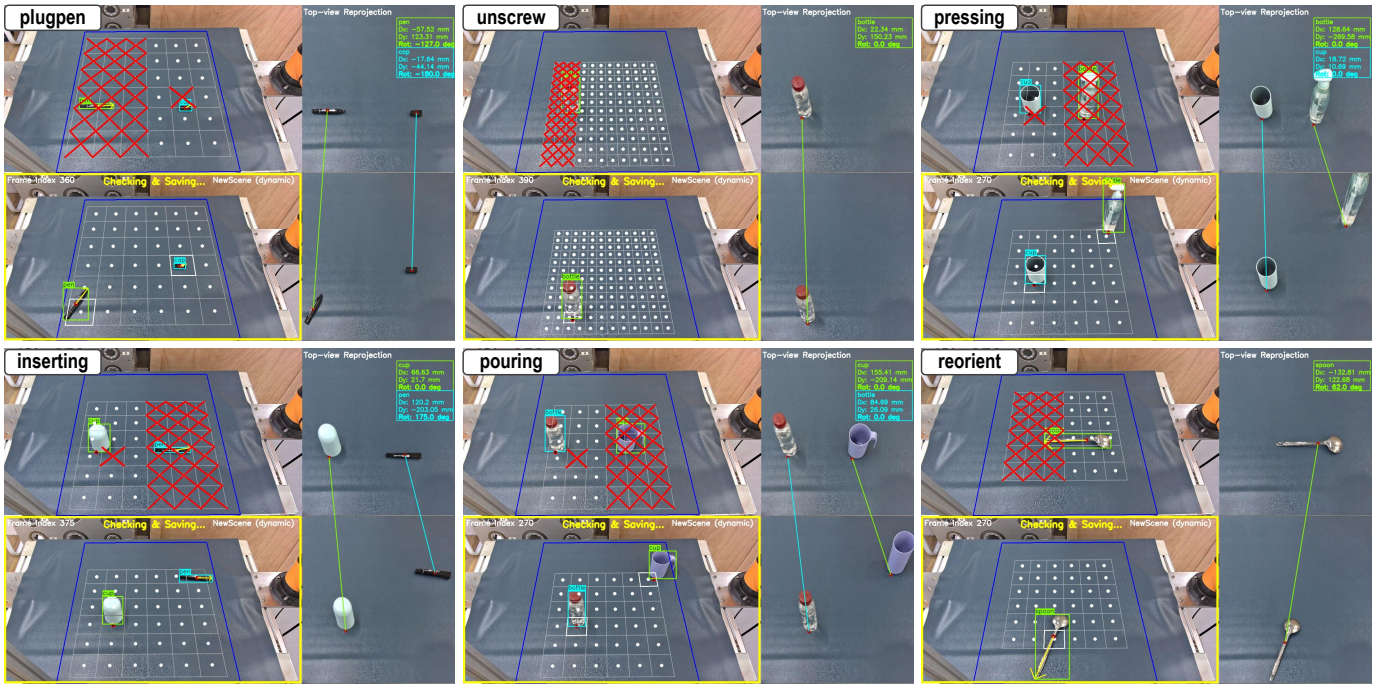

Fig. 13: UI interfaces. When collecting data, six tasks need to display different grid cells in real time, as well as perceive and track task-related objects. Note that these points, crosses and lines are drawn digitally, which are not marks in the real world.

and data reliability, which is the practical need for scaling imitation learning.

A concurrent work MoMaGen [49] tackles scalable bimanual mobile manipulation by generating diverse trajectories in *simulation*, focusing on base reachability and camera visibility through constrained optimization. Although sharing a high-level goal of reducing human effort, MoMaGen and BiDemoSyn operate under fundamentally different assumptions: MoMaGen relies on noise-free simulation assets and virtual scene perturbations, whereas BiDemoSyn performs *real-world demonstration synthesis*, directly addressing physical-domain challenges such as noisy perception, contact feasibility, and one-shot decomposition under real sensor observations. As a result, MoMaGen expands diversity through simulated scene variation, while BiDemoSyn enables *physically grounded generalization* across spatial and category-level object variations without sim-to-real transfer. Thus, BiDemoSyn should be viewed not as a direct competitor to simulation-based frameworks like MoMaGen, but as a *complementary real-world solution* specifically designed for scalable, high-fidelity data generation where simulation pipelines cannot fully capture real-world physical constraints.

G. Evaluation of Demonstration Quality

Evaluating the quality of synthesized demonstrations is essential because the downstream visuomotor policies depend on both the fidelity of observations and the physical feasibility of actions. We assess demonstration quality along two orthogonal dimensions that align with the structure of visuomotor imitation learning:

Visual Fidelity. This dimension evaluates whether the observations included in a synthesized demonstration faithfully reflect the real physical scene without introducing artifacts. Teleoperation and BiDemoSyn preserve raw RGB-D observations, ensuring perfect consistency with the physical environment. YOTO [107] and DemoGen [93] modify the seed point cloud to reposition objects, which introduces geometric distortions and depth inconsistencies, particularly as the displacement grows (see Fig. 4B in the main paper). These artifacts degrade the policy's perception module by coupling unrealistic geometry with real sensor noise.

Trajectory Feasibility & Physical Grounding. A demonstration is considered high-quality if the associated action sequence: (1) yields a collision-free, dynamically stable trajectory, (2) respects the contact geometry implied by the task semantics, and (3) can be executed on the real robot without additional correction or replanning. Teleoperation and YOTO [107] provide ground-truth trajectories. BiDemoSyn produces partially reused (invariant blocks) and partially adapted (variant blocks) trajectories aligned with real-world object pose and geometry. DemoGen [93] synthesizes trajectories directly from edited scene representations, which can deviate from physically stable contact geometries.

Finally, Fig. 4A in the main paper reflects the interplay of these two components. Demonstrations with high-quality observations and physically feasible actions rank highest. This ordering also predicts downstream policy performance: policies trained with cleaner, physically grounded demonstrations demonstrate higher success rates and better robustness under OOD evaluation. These expanded analyses clarify the qualitative and quantitative factors driving the reported comparisons and support the role of demonstration quality as a key determinant of real-world imitation learning performance.

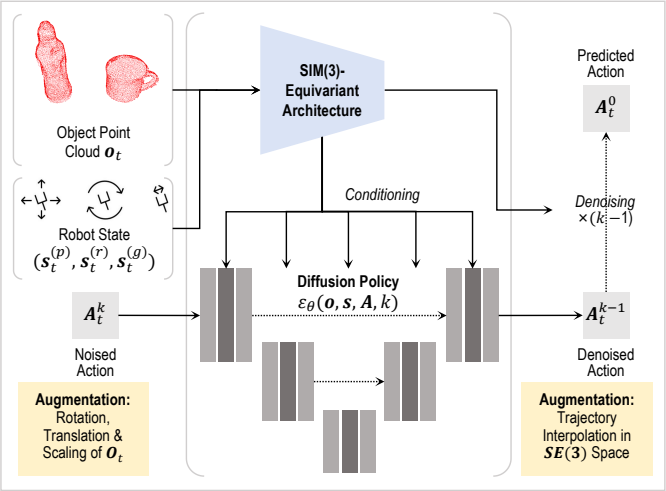

Fig. 14: The adapted visuomotor network architecture for imitation learning of bimanual tasks. The input visual observation is simplified to a point cloud of manipulated objects. Note that the vision encoder is slightly different in DP3 [98] and EquiBot [94]. The illustrated one is for EquiBot.

H. Diversity of Synthetic Demonstrations

A natural question is whether additional diversity (such as reordering or randomizing block sequences) would improve the effectiveness of the synthesized dataset. In BiDemoSyn, we intentionally do not introduce such sequence-level variations. Our empirical findings align with recent analyses in large-scale teleoperation datasets focusing on the diversity exploration and ablation for scalable robotic manipulation [77], which show that low-level behavioral variability (e.g., minor differences in grasp location, end-effector speed profiles, or redundant micro-motions) does not necessarily benefit policy learning and may even hinder convergence. Instead, we adopt the principle that *task success is primarily determined by achieving correct interactions, not by how those interactions are executed.*

Accordingly, in BiDemoSyn's design, we maintain the original causal structure of the task by preserving the block ordering extracted from the one-shot demonstration, while introducing diversity along the *dimensions that matter for generalization*, namely: (1) spatial variation of object pose (translation + rotation), and (2) category-level instance variation (shape and size differences within the same object class). These forms of variation are directly tied to visuomotor generalization and produce clear benefits in downstream policy performance. In contrast, randomizing the internal block structure introduces behavioral diversity that is orthogonal to task semantics and does not improve success rates in our preliminary tests. We therefore avoid unnecessary variation and instead focus on *environment-induced diversity*, which is both meaningful and consistent with the underlying task geometry.

IX. Training and Deployment of Imitation Policies

A. Reproduction of Baselines

For DemoGen [93], we synthesize trajectories by editing the 3D point cloud in one-shot demonstration: objects are translated across grid cells (aligned with Sec. VIII-A standards) and rotated (for orientation-sensitive tasks) to match our BiDemoSyn spatial diversity for fair comparison. Edited point clouds exhibit artifacts (*e.g.*, misaligned edges, self-occluded regions) due to perspective distortions, degrading observation fidelity. For YOTO [107], we reduce data scale to 1/10 of BiDemoSyn to accommodate its time-consuming physical replay. Even at this scale, YOTO requires 388/777/115/259/129/100 replays across all six tasks (*e.g.*, `plugpen` needs 388 replays, taking about 11 hours), versus BiDemoSyn's 3-hour synthesis for 10× data. Note that the efficiency recorded here is the time consumed by continuous collection, including a lot of non-collection time for transition and error recovery. The spatial sampling of YOTO is relaxed to coarse left-right object placements, sacrificing grid uniformity.

B. Implementation of Imitation Policies

Both DP3 [98] and EquiBot [94] (refer Fig. 14) ingest segmented 3D object point clouds but differ in network architectures. Specifically, DP3 employs specially designed multiple MLPs to encode point clouds into latent features, then directly regressing dual-arm keyposes. EquiBot enhances robustness via a SIM(3)-equivariant PointNet variant: point clouds undergo rotation/scale-invariant feature extraction before action prediction. This equips EquiBot with improved generalization to unseen object variations, achieving 5–12% higher success rates than DP3 under identical training data and settings. Both policies operate in open-loop, aligning with synthesized demonstration formats. Below, we give a more detailed description of the visuomotor policy modification for the dual-arm action prediction.

Spaces of Observation and Action. We adopt a 13-dimensional proprioception vector and a 7-dimensional action space for each robot arm, respectively. The proprioception data for each arm consists of the following information: a 3-dimensional end-effector position, a 6-dimensional vector denoting end-effector orientation (represented by two columns of the end-effector rotation matrix), a 3-dimensional vector indicating the direction of gravity, and a scalar that represents the degree to which the gripper is opened. The action space for each arm consists of the following information: a 3-dimensional vector for the end-effector position offset, a 3-dimensional vector for the end-effector angular velocity in axis-angle format, and a scalar denoting the gripper action. For all bimanual tasks, the observation horizon is set to 1, so we only use the initial state observation of left arm as one of the network inputs. And the initial state of right arm is always fixed in each task. For the number of action steps (also the length of the predicted horizon), we simplify it and set the prediction length to the number of keyposes K, which can be extracted from the start and end actions of each Atomic Execution Primitives (AEPs) during task deconstruction.

Network Architecture. For DP3, it is a variant of Diffusion Policy [16] with a simpler point cloud encoder. It also designs a two-layer MLP to encode robot proprioceptive states before concatenating with the observation representation. For

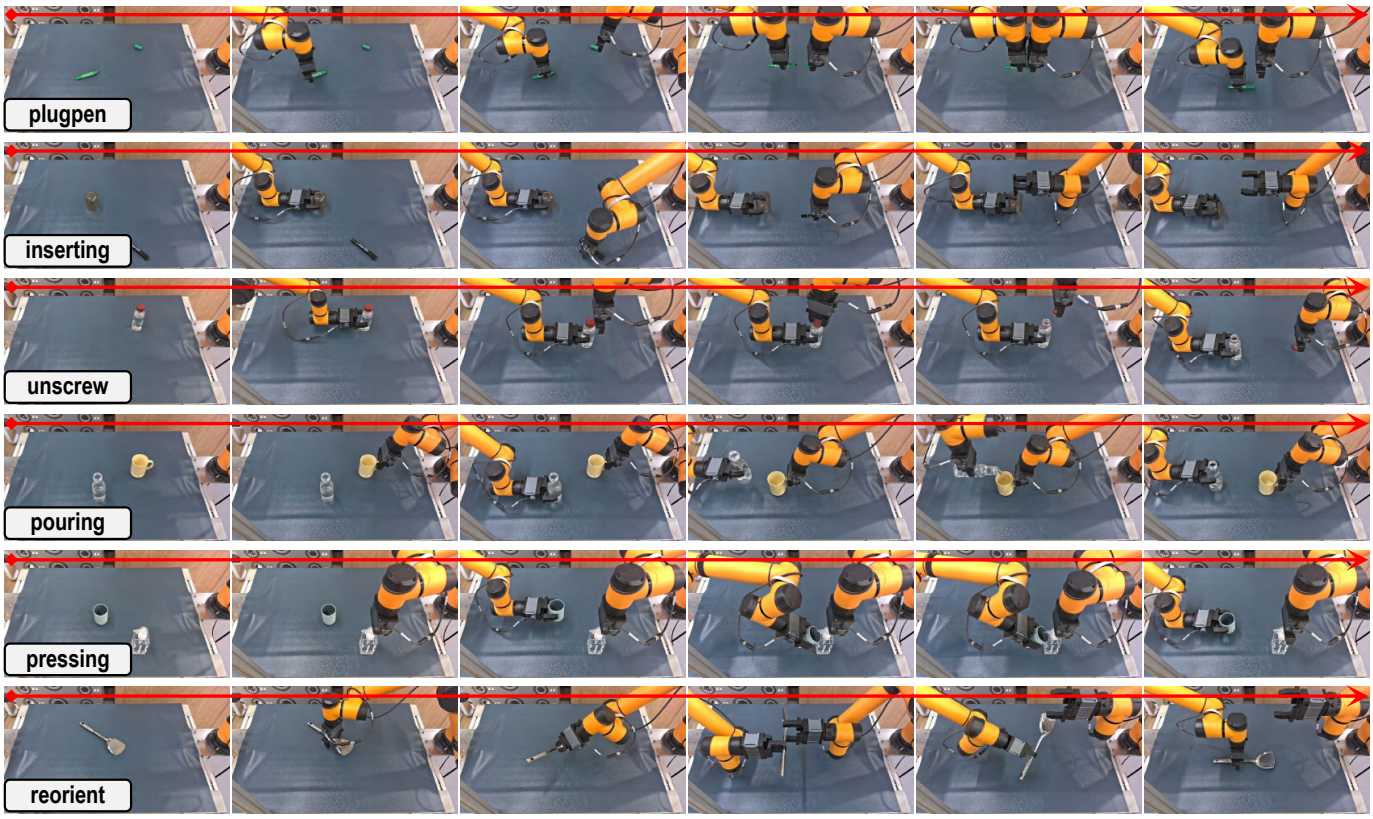

Fig. 15: Qualitative real robot rollout samples for defined six bimanual tasks. Best to view after zooming in.

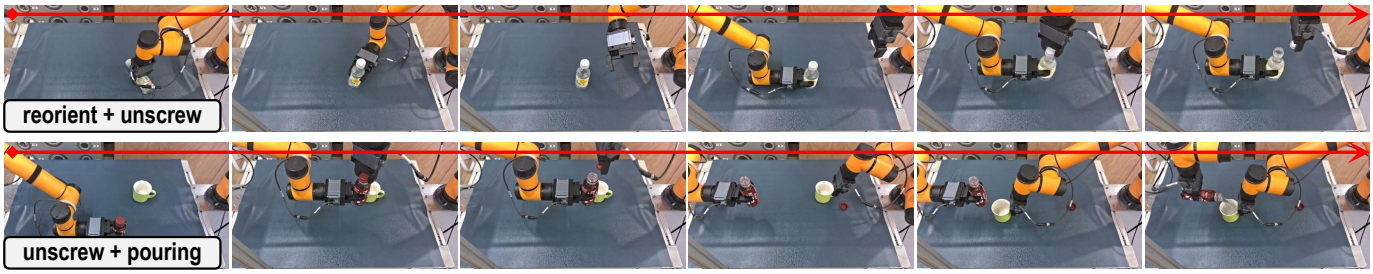

Fig. 16: Qualitative real robot rollout examples for two newly added long-horizon tasks (`pouring` and `reorient+unscrew`).

EquiBot, we use a SIM(3)-equivariant PointNet++ [95] with 4 layers and hidden dimensionality 128 as the feature encoder. For the noise prediction network, we inherits hyperparameters from the original Diffusion Policy. Specifically, to optimize for inference speed in all experiments, we use the DDIM scheduler [79] with 8 denoising steps, instead of the DDPM scheduler [35] which performs up to 100 denoising steps.

Sampling of Point Cloud. As we all known, setting the number of points to sample in the point cloud observation is a key hyperparameter to consider when designing an architecture that takes point cloud inputs. In our experiments, we found out that using 1024 points is sufficient for all tasks and policies. In particular, we have tried increasing the number of point clouds to 2048 or more, but the evaluation improvement in each task is minimal, and this will also cause the storage occupied by the training observation data to be too large and the training time cost to increase. Therefore, reducing the number of points to 1024 can make training faster without hurting performance. And all our policy models can be trained on a GeForce RTX 3090 Ti with 24 GB of memory.

Training and Evaluation. When conducting the in-distribution (ID) and out-of-distribution (OOD) evaluations with full demonstrations, we train all methods for 500 and 1,000 epochs, respectively. Otherwise, when using the 1/10 training data (especially the YOTO), we train all methods of the ID and OOD evaluations for 2,000 and 4,000 epochs, respectively. For all experiments, the batch-size is always set to 64. We only evaluate the last one checkpoint saved at the end of training. For every evaluation in the real world, we run the policy in a randomly initialized placement of objects, and record the average success rate achieved by the policy.

In addition, we have adopted the object-centric point cloud input. At inference time, we also need to preprocess the binocular RGB observations to obtain the point cloud of manipulated objects. This core design relies on the still rapidly developing capabilities of vision foundation models (VFMs).

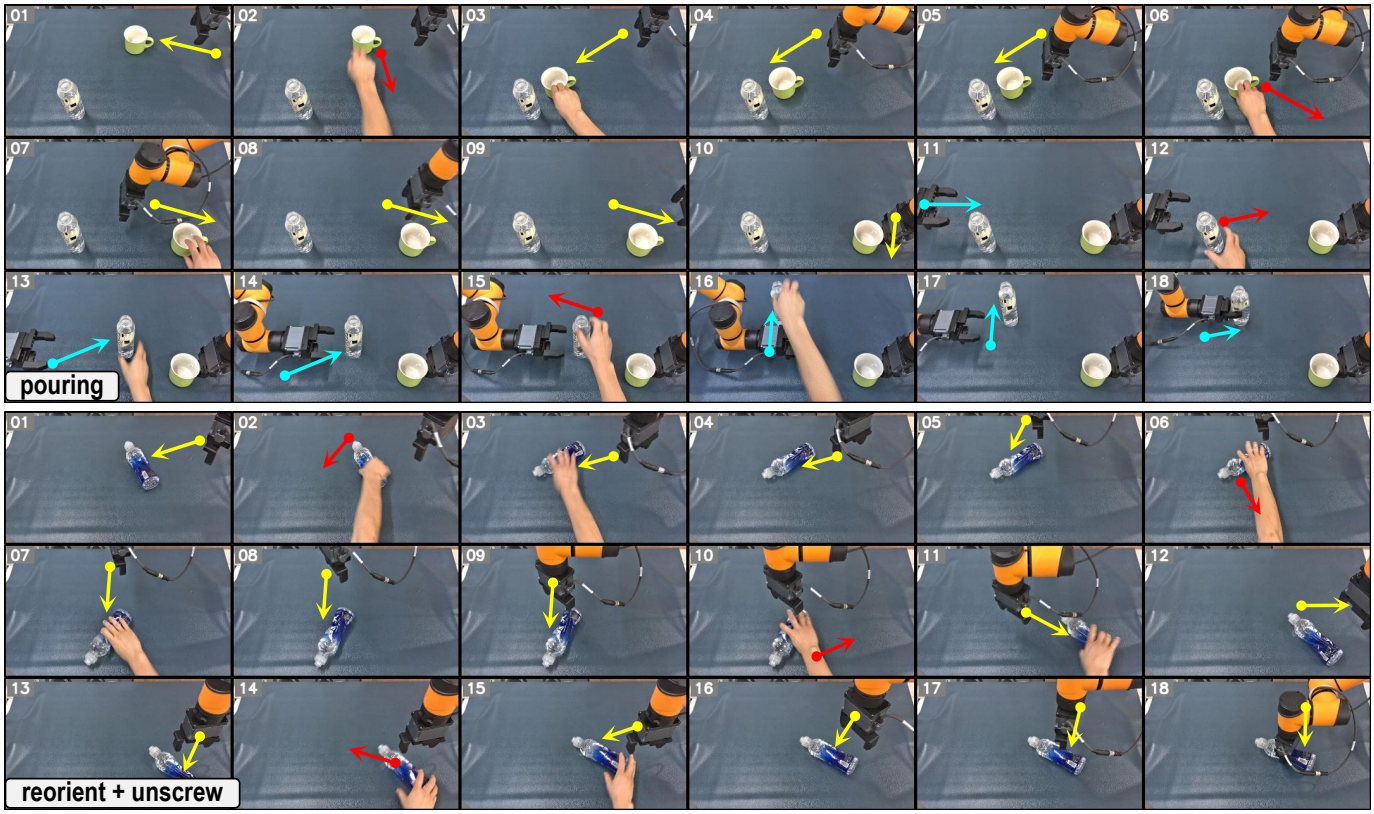

Fig. 17: Qualitative examples of applying dynamic interferences during the pre-grasping execution. From top to bottom, they are segments of the dynamic closed-loop grasping phase of tasks `pouring` and `reorient+unscrew`, where each object is manually disturbed from one to three times. The red arrow indicates the direction of the manually moved object (interfering). The cyan arrow and yellow arrow indicate the movement direction of the left and right robotic arms (chasing), respectively.

Here we leverage SOTA open vocabulary detection method Florence-2 [89] and segmentation method SAM2 [74] to automatically extract object masks and then filter out corresponding point clouds. Despite this, occasionally we may fail to segment desired objects accurately, and in these special cases we will manually correct the masks. These cases are not counted as failed evaluation trails due to not involving significant elements of bimanual robot manipulation. Because we believe that the next generation of VFMs can alleviate these problems, or we can directly address them through domain adaptation, test-time adaptation, or adjusting input prompts.

X. More Results Visualization and Analysis

A. More Examples from Real Rollouts

Fig. 15 supplements more qualitative results with additional real-robot execution visualizations, highlighting nuanced state transitions and task-specific challenges. These visualizations underscore BiDemoSyn's ability to handle real-world complexities (such as mechanical tolerances and imperfect perception), while also exposing limitations in dynamic force modulation (*e.g.*, over-pressing bottles or lids). Please refer our **Supplemental Videos** which provide frame-by-frame analyses of these executions, further dissecting critical phase and offering insights into policy decision-making under uncertainty.

B. Two More Complex Long-Horizon Tasks

To further evaluate the scalability of BiDemoSyn beyond the introduced six bimanual tasks, we add two new long-horizon tasks that require sequential manipulation steps and accumulate more physical error throughout the execution:

- `reorient+unscrew`: The robot must first upright a lying down bottle (reorient phase), then perform a dual-arm unscrewing motion to remove the cap (unscrew phase). This task stresses both precise estimation of bottle orientation and robust reuse of multi-step, contact-rich interaction patterns.
- `unscrew+pouring`: The fisrt phaset is the original unscrew task. After unscrewing the bottle cap, the robot must lift a mug cup, perform dual-arm coordination to tilt the bottle, and pour water into the cup. This composition integrates two high-precision subtasks and requires stable arm coordination across multiple transitions.

To utilize BiDemoSyn, these two newly added tasks were synthesized following the same three-stage pipeline described in the main text. We briefly summarize the configurations: for `reorient+unscrew`, we collected initial observations for 4 bottle instances, each placed at 7 orientations across 36 grid cells, yielding **1008 synthesized demonstrations**; for `unscrew+pouring`, we collected initial observations for 2 bottle and 2 cup instances, each placed on 18 grid cells, yield-

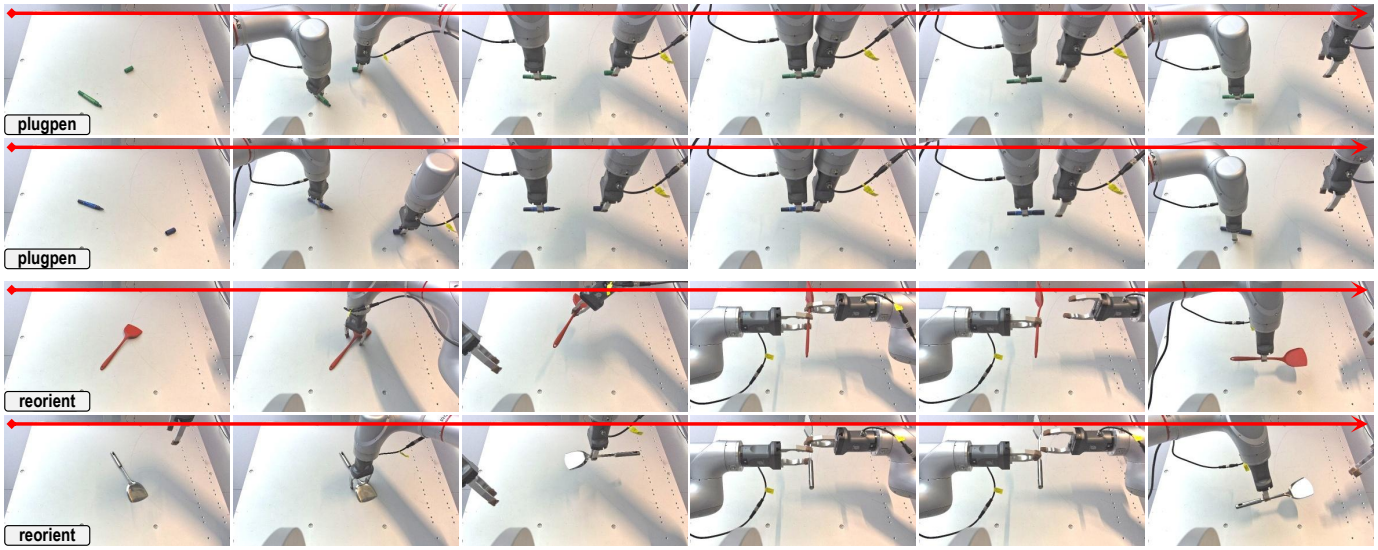

Fig. 18: Qualitative real robot rollout samples for two tasks (`plugpen` and `reorient`) on the auxiliary dual-arm platform.

ing **1296 synthesized demonstrations**. Importantly, both tasks introduce *longer temporal horizons, sequential coordination, and increased compounding error*, thus providing a stronger stress test than the original six tasks.

Then, using the synthesized datasets, we trained visuomotor policies for each long-horizon task based on the adapted EquiBot architecture. Across both tasks, the trained policies exhibited high robustness to spatial variation (different placements and bottle orientations), reliable execution of multi-step dual-arm coordination, and consistent task success under mild external perturbations introduced during pre-grasp interpolation. In Fig. 16, we present some qualitative results with additional real-robot execution visualizations of these two new tasks. And in Fig. 17, we provide illustrative examples of continuous dynamic interference scenarios among tasks `pouring` and `reorient+unscrew`. Full dynamic demonstrations and rollouts of all eight tasks (previous 6 tasks and the new 2 tasks) are included in the **Supplemental Videos**. These results further validate BiDemoSyn's ability to rapidly scale to new, long-horizon tasks with minimal additional engineering, and reinforce its suitability as a practical framework for synthesizing large-scale, real-world training data.

C. Cross-Embodiment Transferability Testing

Beyond training and evaluating policies on the primary dual-arm platform (Fig. 9 top), we further investigate the cross-embodiment transferability of policies trained with BiDemoSyn. Specifically, we deploy the learned visuomotor policies on a distinct humanoid-style dual-arm robot with different kinematic structures, arm mounting configurations, and workspace geometry (Fig. 9 bottom). Importantly, no additional demonstrations, fine-tuning, or embodiment-specific retraining are introduced.

We evaluate two representative bimanual tasks—`plugpen` and `reorient`—which respectively emphasize precise dual-arm alignment and large-range end-effector rotation. As illustrated in Fig. 18, policies trained on the original platform can be directly executed on the new robot after a simple coordinate transformation of the predicted 6-DoF end-effector poses, without requiring explicit calibration of dynamic or contact parameters. Across extensive trials, both tasks achieve stable and consistent execution performance, comparable to those reported on the source platform. To further assess robustness, we introduce external perturbations during the pre-grasp stage (similar to Fig. 17), including manual displacement and rotation of target objects. As shown in Fig. 19, the policy can reliably recover from such disturbances and successfully complete the task, demonstrating that the transferred policy preserves not only task semantics but also robustness to moderate environmental variations. These qualitative results highlight a key advantage of BiDemoSyn: by adopting object-centric observations and embodiment-agnostic action representations, the resulting policies exhibit strong zero-shot transferability across heterogeneous robot embodiments, suggesting promising scalability to broader robotic platforms. More rollouts on this humanoid platform are in our **Supplemental Videos**.

D. Statistics and Analysis of Failed Cases

While policies trained with BiDemoSyn achieve high success rates, failure analysis reveals systematic challenges. Note that although the policy we trained is executed end-to-end (it establishes an implicit mapping from observations to action outputs), we can still align the analysis from the stage where its failure cases are located to find the core cause of the error. Finally, using the experimental results of EquiBot under out-of-distribution evaluations, we categorize failures into five types according to the task execution logic (mainly from the design module of BiDemoSyn):

(1) Visual Perception Errors (15%): Inaccurate object segmentation (*e.g.*, mistaking the tabletop as part of the marker pen) leads to misaligned grasps.

(2) Orientation Estimation Errors (32%): Incorrect object pose estimation (*e.g.*, misjudging a spoon's concave direction) causes malformed trajectories (*e.g.*, flipping to the wrong side).

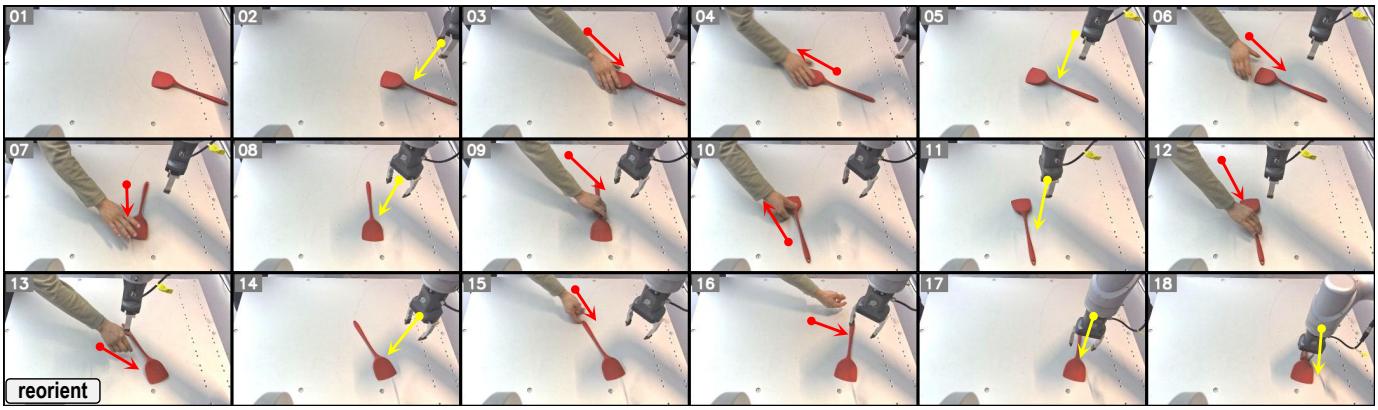

Fig. 19: Qualitative examples of applying dynamic interferences in pre-grasping for the task `reorient` on the auxiliary dual-arm platform, where the object is manually disturbed up to five times. The red arrow indicates the direction of the manually moved object (interfering). The yellow arrow indicates the movement direction of the right robotic arm (chasing).

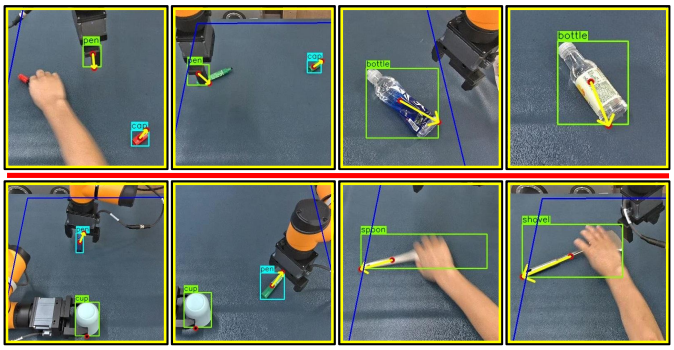

Fig. 20: *(Top)* Failure cases in the task `plugpen` that wrongly detected the marker body, and the task `reorient+unscrew` that incorrectly estimated the lying bottle's orientation. *(Bottom)* Failure cases in the task `inserting` where the robotic arm's end effector occasionally obscures the marker pen, and the task `reorient` where human interference causes the hand to partially obscure the manipulated spoon/shovel.

(3) Alignment Failures (18%): Vision-guided initial frame misalignment (*e.g.*, bottle nozzle offset by >1cm) propagates errors to subsequent steps (*e.g.*, spilling during `pouring`).

(4) Initial Grasp Failures (28%): Collisions during pre-grasp motions (*e.g.*, gripper nudging objects) or unstable grasps (*e.g.*, slippage in `reorient`) prevent task initiation.

(5) Error Propagation (7%): Cumulative errors from earlier stages compound into irreversible failures (*e.g.*, slightly misaligned pluging between pen body and pen cap in `plugpen`).

As can be seen, the orientation estimation and initial grasp failures dominate, reflecting two core challenges: (1) current pose estimators struggle with symmetric or textureless objects (*e.g.*, metal spoon or shovel), and (2) gripper-centric path planning lacks fine-grained contact modeling (*e.g.*, avoiding pre-touch collisions for irregular shapes). Addressing these requires advances in category-agnostic pose estimation and short-horizon contact optimization, which are critical directions for our future work. Some visual failure examples are shown in Fig. 20, which provides a foundation for a full understanding of the BiDemoSyn system and its robust adaptation to allow for dynamic pre-grasping. More dynamic rollouts can be found in our **Supplementary Videos**.

E. VFM-based Object Anchoring in Clutter

To evaluate the applicability of BiDemoSyn under more realistic and visually complex environments, we additionally examine its object-anchoring capability when target objects are placed in cluttered scenes with distractors and partial occlusions. Although the main experiments intentionally isolate environmental variables to study the core challenge of real-world demonstration synthesis, BiDemoSyn is fully compatible with more complex perception conditions due to its reliance on VFMs for object segmentation.

We utilize Florence-2 [89] and SAM2 [74] to detect and segment task-relevant items using either text queries or exemplar-based prompts. Fig. 21 illustrates representative examples where both bottle and mug are correctly segmented out of surrounding tools, utensils, packaging, and nonrigid paper balls. In cluttered tabletop scenes, the VFM consistently produces high-quality masks, even when distractors share similar color similarity and semantic ambiguity. Qualitatively, the predicted masks remain stable across moderate occlusions, enabling the system to isolate the target object with minimal ambiguity and providing reliable anchors for subsequent pose inference and variant-block adaptation in BiDemoSyn.

Because BiDemoSyn's variant-block adaptation depends only on reliable 2D masks along with centroid- and principal-axis–based orientation inference, accurate object anchoring is the only requirement for extending the system to cluttered environments. Our experiments show that the VFM-based perception pipeline provides robust segmentation to support pose-aligned trajectory modulation and collision-aware grasp adjustments, even when the background is visually nonuniform. However, while VFMs significantly enhance robustness to visual complexity, extreme occlusions or heavy inter-object entanglement may still degrade performance. Extending BiDemoSyn to handle these cases will require multi-view sensing, active view planning, or integrating VLM-driven object-query disambiguation. Nevertheless, these initial results demonstrate

Fig. 21: **Top**: Example cluttered tabletop scenes used to evaluate the perception front-end of BiDemoSyn. Each scene includes the target objects for the `pouring` task—two bottles (text prompts are *blue bottle* and *green bottle*) and two mugs (text prompts *blue cup* and *green cup*)—surrounded by visually distracting household items. Two additional tool-like distractors (*brush*, *dustpan*) are intentionally introduced to support future extensions such as dual-arm `sweeping` tasks. **Bottom**: High-quality 2D object masks produced by Florence-2 + SAM2 for all queried categories.

that *VFM-based anchoring provides a practical and reliable foundation for deploying BiDemoSyn beyond clean tabletop settings*, thereby enabling future extensions to household, warehouse, and mobile manipulation scenarios.

XI. Summary, Reflection and Outlook

This work introduces BiDemoSyn, a real-world demonstration synthesis framework that addresses a central bottleneck in bimanual imitation learning: how to scale high-quality, physically grounded data collection without reliance on simulation or extensive teleoperation. By decomposing a single kinesthetic demonstration into invariant coordination structures and object-dependent adaptations, and by coupling vision-guided alignment with lightweight trajectory optimization, BiDemoSyn enables the rapid generation of diverse, feasible bimanual demonstrations directly in the physical world. Extensive experiments across six contact-rich tasks demonstrate its efficiency, scalability, and strong generalization to novel object poses and geometries.

Beyond the original one-shot setting, we further show that BiDemoSyn naturally extends to *few-shot demonstration synthesis*, where additional demonstrations primarily enhance out-of-distribution generalization without altering the core pipeline. Moreover, we demonstrate *zero-shot cross-embodiment policy deployment*, where policies trained on BiDemoSyn data transfer effectively to a distinct humanoid-style dual-arm robot, preserving high success rates across representative tasks. These results highlight the importance of object-centric observations and embodiment-agnostic action representations, and suggest that BiDemoSyn supports not only scalable data generation, but also robust policy reuse across hardware platforms.

Despite these advances, several limitations remain. BiDemoSyn currently assumes quasi-static task execution and rigid or semi-rigid objects, making highly dynamic interactions and deformable object manipulation challenging. Failure cases are primarily attributed to perception noise—particularly orientation estimation for small or symmetric objects—and occasional contact inaccuracies during grasping and insertion. Addressing these issues will likely require tighter perception–control coupling, category-agnostic orientation reasoning, and more expressive contact-aware planners.

Looking forward, we see multiple promising directions. Future work will explore extending BiDemoSyn to deformable and articulated objects, incorporating dynamic and multi-view perception, and integrating closed-loop feedback into the synthesis pipeline. More broadly, we envision BiDemoSyn as a foundation for hybrid data-centric learning paradigms, combining one-shot, few-shot, and synthesized demonstrations to balance data efficiency with maximal generalization. By lowering the barrier to acquiring large-scale, real-world bimanual data, this work aims to accelerate progress toward practical, general-purpose robotic manipulation.