Tool-Action Comprehension: A Cross-Morphology Imitation Learning Framework for Bimanual Manipulation

Yechen Fan, *Graduate Student Member, IEEE*, Xinjie Zhang, Jianghao Zhao, Xiaohan Liu, Haibin Wu, Jinhua Ye, *Senior Member, IEEE*, and Gengfeng Zheng, *Senior Member, IEEE*

***Abstract*—Humans excel at bimanual manipulation, while robots struggle with bimanual tasks due to the high cost of data collection. To address this, we propose the tool-action comprehension (TAC) framework, a cross-morphology imitation learning method that directly acquires bimanual skills from human demonstration videos. TAC establishes a morphologically invariant representation by leveraging matching and stereo 3D reconstruction (MASt3R) and 3-D Gaussian Splatting for 3-D scene reconstruction and FoundationPose for 6-D tool pose alignment. Crucially, it employs grounded segment anything model (SAM) to decouple human hand morphology from visual observations, generating embodiment-agnostic observation–action representations without teleoperation devices or manual annotations. Inspired by human hierarchical planning, TAC decomposes tasks into high-level target prediction and low-level trajectory generation, unified through a conditional diffusion model to produce smooth and coordinated motions. Experiments demonstrate that TAC achieves an average success rate of 81.5%, outperforming the strongest baseline by an absolute margin of 62.2%, and enables scalable data collection, with novices collecting demonstrations 14.9% faster than experts.**

***Index Terms*—Bimanual manipulation, cross-morphology imitation learning, diffusion policy (DP).**

I. Introduction

HUMANS effortlessly perform complex bimanual tasks, from folding blankets to flipping eggs, demonstrating skills far beyond current robots. Developing such bimanual coordination is important for general-purpose robotic manipulation, yet most research remains constrained to simple actions, such as grasping or picking, with bimanual tool use beyond parallel grippers still a major challenge.

Although robotic manipulation has made significant progress recently, bimanual coordination remains a challenge. End-to-end state-to-action mapping approaches struggle to generalize in multimodal environments and unseen interactions, whereas humans collaborate efficiently through hierarchical task goal planning. This has driven the development of imitation learning, which decomposes complex tasks into predicting collaborative poses and synchronized trajectories. However, it relies on high-quality demonstration data, and collecting such data for bimanual tasks remains difficult [1], [2]. Teleoperation methods [3] suffer from latency, hardware dependence, and environmental constraints, while handheld teaching devices [4], [5] are costly and cumbersome. More critically, both methods depend on expert demonstrations, limiting large-scale data collection.

A promising solution is to leverage natural human demonstrations in everyday life [6], where people skillfully use tools without costly or specialized hardware. Such data are abundant and inexpensive, but disparities in morphology, perception, and control prevent direct transfer to robots, especially for bimanual manipulation. Existing cross-morphology methods [7], [8] attempt to bridge this gap often via kinematic retargeting, which suffers from complex structural mismatch.

To address these challenges, we propose the tool-action comprehension (TAC) framework, a cross-morphology imitation learning framework that learns directly from natural human tool-use videos, avoiding the reliance on costly teleoperation systems or wearable devices (see Fig. 1). It uses grounded segment anything model (SAM) to remove human hands from visual observations, focusing on tool-environment interactions to achieve stronger generalization. In addition, matching and stereo 3D reconstruction (MASt3R) and 3-D Gaussian Splatting

Received 26 November 2025; revised 27 February 2026 and 14 May 2026; accepted 12 July 2026. Recommended by Technical Editor R. Meattini and Senior Editor J. Ueda. This work was supported in part by the Fujian Provincial Science and Technology Major Project of China under Grant 2024HZ026020, in part by the Projects of Fujian University Industry-Academic Cooperation under Grant 2022H6016, in part by the Project of the Ministry of Industry and Information Technology under Grant 2024-RL-FZSX-01, in part by the Fujian Young Scientific and Technological Personnel Training under Grant 2025350124, and in part by the Projects of Strategic Emerging Industries and Future Industries of Fuzhou-Xiamen-Quanzhou National Independent Innovation Demonstration Zone. *(Corresponding authors: Jinhua Ye; Gengfeng Zheng.)*

Yechen Fan, Xinjie Zhang, Jianghao Zhao, Xiaohan Liu, Haibin Wu, and Jinhua Ye are with the School of Mechanical Engineering and Automation, Fuzhou University, Fuzhou 350116, China (e-mail: 240227014@fzu.edu.cn; zhangxinjie0907@163.com; zjh19157703940@163.com; 240227050@fzu.edu.cn; wuhb@fzu.edu.cn; yejinhua@fzu.edu.cn).

Gengfeng Zheng is with the Fujian Key Laboratory of Special Intelligent Equipment Safety Measurement and Control, Fujian Special Equipment Inspection and Research Institute, Fuzhou 350008, China (e-mail: 18862889747@163.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMECH.2026.3713779.

Digital Object Identifier 10.1109/TMECH.2026.3713779

Fig. 1. Overview of the TAC framework. By bridging the cross-morphology gap, TAC enables bimanual robots to learn complex collaborative skills directly from natural human demonstration videos.

are employed for multiview 3-D reconstruction, while FoundationPose extracts 6-D tool-pose trajectories, automatically generating high-quality tool-pose action labels without manual annotation or specialized hardware. At the policy learning level, the framework, inspired by hierarchical planning in human manipulation, decomposes each task into high-level short-horizon goal prediction and low-level trajectory generation. The high-level policy predicts short-term collaborative goals to resolve ambiguity, while the low-level policy generates smooth and feasible trajectories. These are unified in a conditional diffusion model, which generates high-quality continuous trajectories and achieves natural synchronization by jointly modeling the actions of both arms, thereby ensuring coordinated movements.

Our main contributions are summarized as follows:

1) We propose a novel tool-centric alignment pipeline that directly extracts high-fidelity tool trajectories from human videos, eliminating the complex structural mapping required by traditional kinematic retargeting.
2) We construct an embodiment-invariant visual representation via morphological decoupling, effectively bridging the perceptual domain gap between humans and robots.
3) We design a goal-to-trajectory hierarchical policy that decomposes complex bimanual tasks into short-horizon goal prediction and trajectory generation, improving robustness and generalization.
4) Experiments validate the effectiveness and usability of TAC for nonexpert data collection.

II. Related Works

A. Learning-Based Bimanual Manipulation

Enabling bimanual robots to collaborate like humans is essential for complex manipulation, yet the vast action space of bimanual systems makes learning highly challenging. Existing research mainly follows two directions: simplifying the action space via predefined parameterized actions [9], composable motion primitives [10], or role assignment between arms [11], and introducing prior knowledge through graph-based interarm modeling [12] or hierarchical structures [13]. While effective in specific tasks, these approaches depend heavily on task-specific priors, which severely limit generalization when contexts shift. In particular, many such methods adopt a discrete skill abstraction to structure long-horizon behaviors. While methods like BUDS [14] effectively segment tasks into discrete skills via temporal clustering, they typically rely on consistent robot proprioception and discrete switching. In contrast, TAC targets cross-morphology alignment by predicting tool-centric goals, providing spatial guidance to stabilize the low-level policy against domain gaps.

B. Cross-Morphology Policy Learning

Cross-morphology policy learning seeks to transfer skills across different robotic embodiments. Early work relied on multirobot datasets with embodiment-specific representations, achieving only limited generalization [15], [16]. More recent approaches exploit natural human videos, either by extracting object motions to render synthetic robot demonstrations [7] or inferring end-effector keyposes from hand tracking [8], but still require robot-specific data generation, such as kinematic simulation and real-world rollouts, for policy training and remain fragile due to the large perceptual and morphological gap between humans and robots. To mitigate this, unified transformer encoders have been proposed to map observations and actions into a shared discrete space [17], [18], [19], yet these methods demand large-scale models and datasets, incur high training costs, and still lack robust few-shot transfer to unseen morphologies. In contrast, TAC employs grounded SAM to decouple human hand regions and constructs a fully tool-centric observation–action space, achieving domain alignment beyond human–robot differences.

C. Data Collection for Policy Learning

High-quality data is essential for robotic learning and generalization, with existing approaches [7], [8] relying on simulation data, real robot data, and human data. Simulation is popular for its controllability and ease of collection [20], but the sim-to-real gap undermines performance in real settings. To address this, research has turned to teleoperated demonstrations, such as generalizable learning for low-cost, low-latency teleoperation (GELLO) [3], which lowers costs but remains platform-specific. Physical interfaces like universal manipulation interface (UMI) [4] and fast universal manipulation interface (FastUMI) [5] extend data acquisition, yet their expense and hardware dependence limit scalability. Human demonstrations offer a more promising alternative, though prior work often requires specialized tools or additional hardware [6], [31], adding deployment overhead. In contrast, our approach leverages only natural human demonstration videos in everyday environments, enabling low-cost, device-free, and scalable data collection.

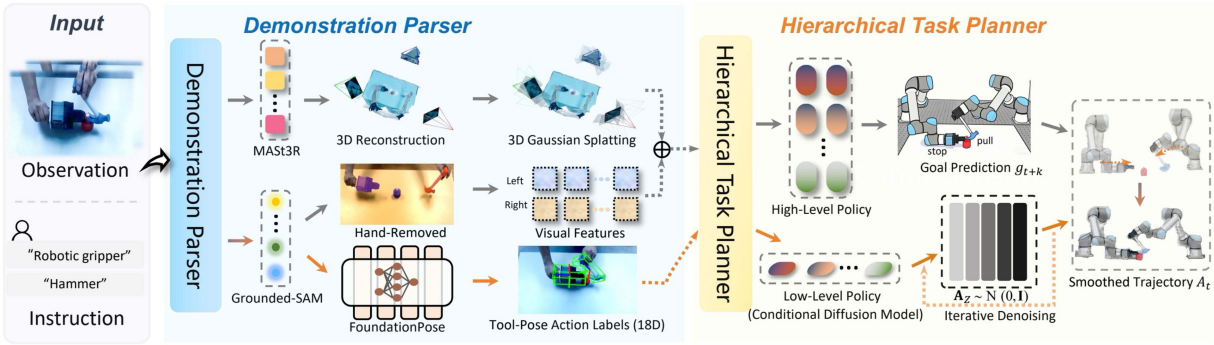

Fig. 2. Technical architecture of TAC. TAC takes visual observations and text prompts as inputs. The demonstration parser uses MASt3R and 3-D Gaussian Splatting for 3-D reconstruction, FoundationPose for pose estimation, and grounded SAM for morphological decoupling, automatically converting human videos into structured training data. The final training samples include hand-removed tool-centric observations and the corresponding 18-D tool-pose action labels. During deployment, the hierarchical planner first predicts short-term coordinated goals, and the low-level DP then generates specific, coordinated bimanual trajectories to execute the task.

III. Problem Formulation

We formulate bimanual coordination as a Markov decision process [21] to learn a policy $\pi : \mathcal{O}'_t \to \mathcal{A}_t$. The raw observation $\mathcal{O}_t$ comprises the RGB image $I_t \in \mathbb{R}^{H\times W\times 3}$ and current tool poses $\{X_{l,t}, X_{r,t}\} \in \mathrm{SE}(3)^2$. After morphological decoupling, the processed tool-centric observation $\mathcal{O}'_t$ comprises the tool-centric image I'_t and the tool poses and is provided to the policy. The action $\mathbf{a}_t \in \mathbb{R}^{18}$ concatenates the target poses of both tools ($[a_{l,t}^\top, a_{r,t}^\top]^\top$), where each pose is a 9-D vector (3-D translation + 6-D rotation, see Section IV-E). To mitigate the complexity of mapping high-dimensional observations to bimanual actions, we decompose the policy into hierarchical subtasks [22]: goal prediction and trajectory generation.

1) *High-level policy*: Coordinated goal pose prediction. We learn a high-level module g_θ, which, based on the processed observation $\mathcal{O}'_t$, predicts a short-horizon (k steps ahead) coordinated goal $g_{t+k} \in \mathrm{SE}(3) \times \mathrm{SE}(3)$ that both tools need to achieve. This goal provides the policy with a clear, short-term intent.
2) *Low-level policy*: Synchronized trajectory generation. We learn a low-level policy π_ϕ, which takes the processed observation $\mathcal{O}'_t$ and the high-level predicted goal g_{t+k} as conditions to generate a concrete bimanual action $\mathbf{a}_t = [a_{l,t}^\top, a_{r,t}^\top]^\top$ to drive the robot towards the goal.

Thus, the overall policy can be expressed as: $\mathbf{a}_t = \pi_\phi(\mathcal{O}'_t, g_\theta(\mathcal{O}'_t))$.

Our core challenge is to learn two policy modules from an offline dataset $\mathcal{D}$ containing only human demonstrations. This raises a key question: how can we bridge the morphology gap [23] and execute the learned policy on a robot when morphology and viewpoints are misaligned?

IV. Methodology

We propose the TAC framework, which consists of two core modules: a demonstration parser and a hierarchical task planner, as shown in Fig. 2.

A. Demonstration-to-Robot Alignment Pipeline

During data collection, we use three RGB cameras with different viewpoints to capture the entire process of a human performing a bimanual manipulation task with everyday tools, recording the video streams simultaneously. This process constitutes a human demonstration dataset $\mathcal{D}$ of synchronized frame-level samples, where each sample consists of multiview images and the corresponding two-handed tool poses

$$\mathcal{D} = \{(\mathcal{O}_i^{\text{cam}}, a_i^h)\}_{i=1}^N \tag{1}$$

where N is the total number of synchronized frames across all demonstration trajectories, $\mathcal{O}_i^{\text{cam}} = \{I_i^1, I_i^2, I_i^3\}$ is the observation from the human demonstration with $I_i^v \in \mathbb{R}^{H\times W\times 3}$ being the RGB image from the vth camera. The action label $a_i^h = \{T_{\text{tool},l,i}, T_{\text{tool},r,i}\}$ represents the 6-D poses of the left and right tools relative to the main camera's coordinate frame, belonging to the $\mathrm{SE}(3) \times \mathrm{SE}(3)$ space.

To recover accurate 3-D information and tool poses from 2-D images, our pipeline proceeds in steps. Crucially, to prevent human artifacts from affecting the scene geometry, we first apply the morphological decoupling module to mask out human hands from the raw observations. Utilizing these processed tool-centric images, we employ MASt3R [24] to compute dense pixel-level correspondences. Based on this, we construct a 3-D Gaussian Splatting representation S of the scene [25]. This method generates high-quality, differentiable scene models, which are fundamental for subsequent precise pose estimation.

Given this 3-D scene, we employ the 6-D pose estimation model FoundationPose [26], using the tool's 3-D mesh model M_{tool} as a prior, to accurately estimate the poses $\{T_{\text{tool},l,t}, T_{\text{tool},r,t}\}$ of the two tools in each frame relative to the main camera's coordinate system, thereby automatically generating high-quality action labels a_t^h

$$\{T_{\text{tool},l,t}, T_{\text{tool},r,t}\} = \text{FoundationPose}(\{I_t^v\}_{v=1}^3, M_{\text{tool}}). \tag{2}$$

These estimated tool poses serve as the supervised action labels for policy learning, enabling TAC to learn actions in a morphology-invariant tool-pose space without robot joint-angle or teleoperated end-effector annotations.

This pipeline aligns the action space by extracting high-quality tool pose labels from raw video. However, a significant domain shift remains within the observation data itself: the raw images contain the human operator's hands, which may cause the policy network to erroneously associate actions with

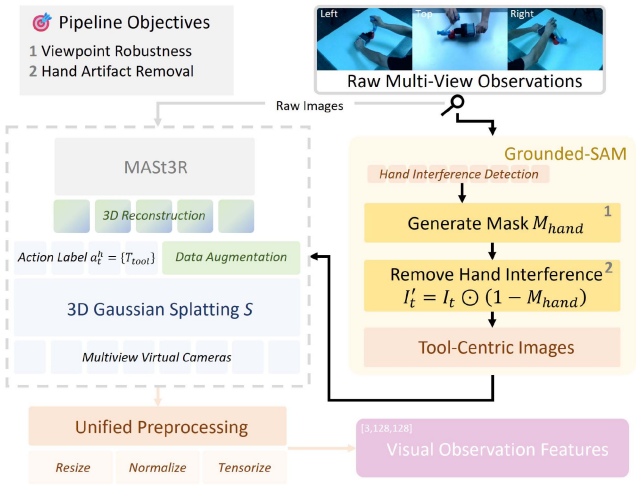

Fig. 3. Visual augmentation pipeline. 1) *Morphological decoupling*: Grounded SAM removes human hand artifacts to align the observation space. 2) *Geometric augmentation*: The hand-removed scene is reconstructed via MASt3R and 3-D Gaussian Splatting to synthesize novel viewpoints, enhancing policy resilience to camera perturbations.

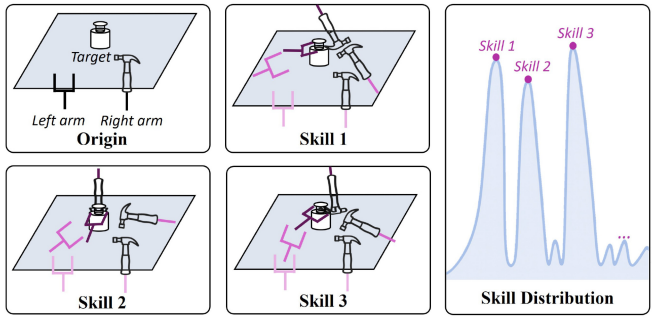

Fig. 4. Multimodal action hierarchical planning. To address the multiple valid solutions in complex tasks, TAC adopts hierarchical planning to first select a clear subgoal, transforming multimodal decision-making into a unimodal trajectory generation problem and enabling stable, unambiguous action output.

morphological features of the human hand rather than the actual interaction between the tool and the environment.

B. Morphological Decoupling for Policy Learning

To align the visual observation space across human–robot morphological differences in policy learning, we use grounded SAM [27] during training data preprocessing to generate a binary mask M_{hand} of the human hand, thereby enabling an embodiment-invariant scene representation. We then apply this mask to the original image to eliminate hand features

$$I'_t = I_t \odot (1 - M_{\text{hand}}) \tag{3}$$

where $\odot$ denotes element-wise multiplication. The processed image I'_t has the pixel values in the hand region set to zero, encouraging the policy network to focus on task-relevant tool-object interactions rather than hand-specific visual cues.

C. Viewpoint Data Augmentation

To enable the policy to handle various unknown camera viewpoints during deployment, we design a 3-D scene-based data augmentation method (see Fig. 3). The core of this method is leveraging the high-fidelity Gaussian Splatting scene representation S reconstructed in Section IV-A. This 3-D model-based augmentation provides more geometrically consistent viewpoint variations than conventional 2-D image augmentation.

Specifically, we leverage the continuous rendering capability of Gaussian Splatting to synthesize photorealistic RGB images from virtual camera poses $T_{\text{cam,new}}$. For each trajectory, virtual viewpoints are sampled around the calibrated real camera poses within a bounded perturbation range, rather than from unconstrained random poses, to mimic realistic deployment-time camera variations while preserving task visibility. The rendered views are combined with the original observations as augmented training samples. Since S is reconstructed from hand-masked images, the synthesized virtual views inherently maintain the embodiment-invariant (tool-centric) property, eliminating the need for repeated segmentation. Simultaneously, through a standard coordinate transformation, we convert the original tool pose labels to the corresponding virtual camera frame, ensuring that each rendered observation is paired with a geometrically consistent action label. By repeating this process, we automatically generate a large-scale dataset with Gaussian Splatting-based augmentation, encouraging the policy network to learn robust 3-D spatial representations and enhancing its resilience to real-world camera instability.

D. Goal-to-Trajectory Policy Learning

As shown in Fig. 4, inspired by human hierarchical planning, we decompose end-to-end imitation learning into a two-stage policy. The high-level module is a deterministic model trained via behavior cloning to predict short-term goals, while the low-level policy is a conditional diffusion model [28] that leverages its strength in producing smooth, diverse, and high-quality trajectories, crucial for fine-grained bimanual manipulation.

During policy learning and inference, the raw observation $\mathcal{O}_t$ consists of the RGB image I_t from its main camera and the current tool poses $\{X_{l,t}, X_{r,t}\}$ from proprioception. To align with our data processing pipeline, this raw observation is preprocessed. The processed observation, denoted as $\mathcal{O}'_t$, includes the tool-centric image I'_t [as per (3)] and the tool poses, ensuring a morphologically consistent input space during both training and inference.

The high-level policy module g_θ understands the "next step" of the task by predicting the coordinated goal pose g_{t+k} for the two tools k steps into the future based on the processed observation $\mathcal{O}'_t$

$$g_{t+k} = g_\theta(\mathcal{O}'_t) \tag{4}$$

where $g_{t+k} \in \text{SE}(3) \times \text{SE}(3)$. The prediction horizon k is a key hyperparameter balancing goal predictability with planning depth. We set it to a fixed value in our experiments. This module is trained via behavior cloning [21], with the objective of minimizing the mean squared error between the predicted goal

and the true future pose a^h_{t+k} from the human demonstration data

$$\mathcal{L}_{\text{goal}}(\theta) = \mathbb{E}_{(\mathcal{O}'_t, a^h_{t+k})\sim\mathcal{D}}\left[\|g_\theta(\mathcal{O}'_t) - a^h_{t+k}\|_2^2\right]. \quad (5)$$

The low-level policy π_ϕ, a conditional diffusion model [30], generates a future H-step smooth and synchronized action trajectory $A_t = \{\mathbf{a}_t, \ldots, \mathbf{a}_{t+H-1}\}$ based on the processed observation $\mathcal{O}'_t$ and the predicted goal g_{t+k}.

In the forward diffusion process, for a real trajectory segment A_0 from human demonstrations, we gradually add Gaussian noise over Z steps. At any step z, a noisy trajectory A_z can be sampled directly

$$A_z = \sqrt{\bar{\alpha}_z}A_0 + \sqrt{1-\bar{\alpha}_z}\epsilon, \quad \text{where} \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (6)$$

with $\bar{\alpha}_z = \prod_{i=1}^{z}(1-\beta_i)$ derived from a predefined noise schedule $\{\beta_z\}_{z=1}^{Z}$. As z increases, the original data A_0 is gradually overwhelmed by noise.

In the reverse process, the network π_ϕ learns to predict the added noise ϵ from the noisy trajectory A_z, the time step z, and conditional information c_t

$$\hat{\epsilon} = \pi_\phi(A_z, z, c_t) \quad (7)$$

where $c_t = \text{Concat}[\text{Encoder}(\mathcal{O}'_t), g_{t+k}]$ concatenates visual features with the high-level goal. The training objective minimizes the prediction error

$$\mathcal{L}_{\text{diffusion}}(\phi) = \mathbb{E}_{A_0,\epsilon,z,c_t}\left[\|\epsilon - \pi_\phi(\sqrt{\bar{\alpha}_z}A_0 + \sqrt{1-\bar{\alpha}_z}\epsilon, z, c_t)\|_2^2\right]. \quad (8)$$

During inference, we start with pure noise $A_Z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and iteratively denoise it over Z steps

$$A_{z-1} = \frac{1}{\sqrt{\alpha_z}}\left(A_z - \frac{1-\alpha_z}{\sqrt{1-\bar{\alpha}_z}}\pi_\phi(A_z, z, c_t)\right) + \sigma_z\epsilon' \quad (9)$$

where the sampling is made stochastic by adding Gaussian noise $\epsilon' \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ for steps $z > 1$ (with $\epsilon' = \mathbf{0}$ for the final step $z = 1$). The noise variance is given by $\sigma_z^2 = \frac{1-\bar{\alpha}_{z-1}}{1-\bar{\alpha}_z}\beta_z$.

For practical execution, we adopt a receding-horizon execution strategy, sequentially tracking the generated tool-pose trajectory and replanning from updated observations

$$\mathbf{a}_{\text{exec}} = \mathbf{a}_t. \quad (10)$$

This replanning mechanism, conditioned on the latest processed observation $\mathcal{O}'_t$, allows the robot to rapidly adapt to environmental changes, such as a shifted target (see Fig. 5).

E. Action Representation and Execution

The low-level policy outputs a future action trajectory, whose actions are sequentially executed as complete 18-D bimanual commands. Each command is the concatenation of two 9-D vectors, one for each arm $(a_{l,t}, a_{r,t})$

$$\mathbf{a}_t = [a_{l,t}^\top, a_{r,t}^\top]^\top \in \mathbb{R}^{18}. \quad (11)$$

Each 9-D subvector comprises a 3-D translation $\mathbf{p}_{\text{arm}}$ and a 6-D continuous rotation representation $(\mathbf{c}_{1,\text{arm}}, \mathbf{c}_{2,\text{arm}})$, corresponding to the first two columns of the target rotation matrix.

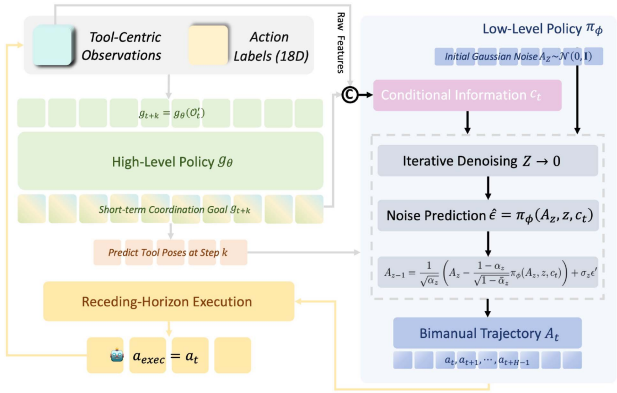

Fig. 5. Inference pipeline of the hierarchical task planner. The framework integrates high-level goal planning with low-level diffusion-based control. The high-level policy g_θ predicts a short-term coordination goal g_{t+k} from the processed tool-centric observation $\mathcal{O}'_t$. The low-level policy π_ϕ constructs conditioning information c_t and generates a bimanual trajectory A_t via iterative denoising. Finally, receding-horizon execution tracks the predicted tool-pose trajectory to achieve closed-loop control.

To convert the network's unconstrained output into valid rotation matrices ($R_{l,t}, R_{r,t} \in \text{SO}(3)$), the Gram–Schmidt orthonormalization process is applied independently to the 6-D rotation components of each arm's action subvector.

During deployment, the policy generates the target tool poses for both arms, $T^{\text{cam}}_{\text{tool},l}$ and $T^{\text{cam}}_{\text{tool},r}$, in the camera frame. We transform these into executable commands for each end-effector in the robot's base frame using precalibrated transforms

$$T^{\text{base}}_{\text{eef, arm}} = T^{\text{base}}_{\text{cam}}\, T^{\text{cam}}_{\text{tool, arm}}\, (T^{\text{eef}}_{\text{tool, arm}})^{-1} \quad (12)$$

where $T^{\text{base}}_{\text{cam}}$ is the fixed hand-eye calibration matrix and $T^{\text{eef}}_{\text{tool, arm}}$ is the tool-to-end-effector transformation for the respective arm. The resulting $T^{\text{base}}_{\text{eef, arm}}$ is used as the target end-effector pose and tracked by the robot-side controller via inverse kinematics and Cartesian servoing. Robot joint commands are, therefore, generated online during execution rather than used as supervised training labels.

V. Policy Evaluation and Results

In this section, we evaluate the performance of the TAC framework in bimanual tasks. A series of experiments evaluates task success rate, robustness, trajectory quality, and data efficiency and compares TAC with state-of-the-art (SOTA) methods.

A. Experimental Setup

Platform setup: The experimental platform consists of two universal robots 5e(UR5e) arms, each equipped with a 3D-printed task-specific tool. The workspace is observed by three Intel RealSense RGB-D cameras: one D435 at the top center and two D415 at the front left and right [see Fig. 6(a)]. System calibration was performed in two stages: extrinsic parameters were computed using a 6×9 ChArUco board (0.04 m square size), followed by fine-tuning the robot base transformation to align the scene point cloud with the robot's URDF model. Visual data are sampled at 30 Hz and synchronized with 125 Hz robot

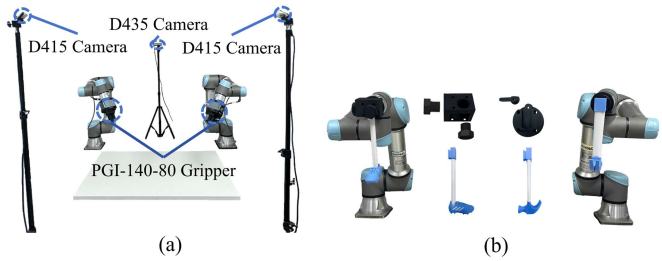

Fig. 6. Hardware setup. (a) Symmetrical dual UR5e arms and three fixed Intel RealSense cameras. (b) Quick-change tool adapter compliant with the ISO 9409-1-50-4-M6 flange standard.

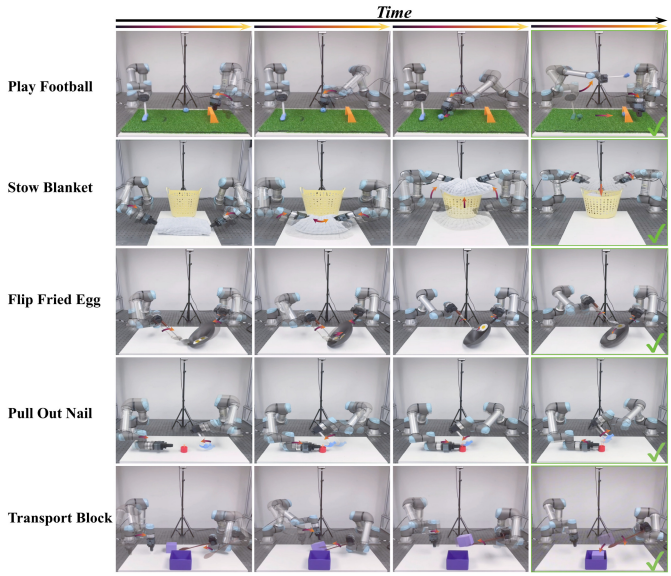

Fig. 7. Five bimanual manipulation tasks under TAC.

TABLE I
TASK CHARACTERISTICS AND NUMBER OF EVALUATIONS

Task Name	Precision	Dynamics	Dexterity	Contact	Gravity	Demos	Trials
Play Football	✓	✓	✓	✓	✓	80	28
Stow Blanket	✓	×	✓	✓	✓	80	18
Flip Fried Egg	✓	✓	✓	✓	✓	90	26
Pull Out Nail	✓	×	✓	✓	×	80	22
Transport Block	✓	×	✓	✓	✓	70	25

proprioception data. During the data preprocessing, the box and text thresholds for Grounded SAM were both set to 0.3. For rapid tool switching, we employ a quick-change device compliant with the ISO 9409-1-50-4-M6 standard [36] [see Fig. 6(b)]. In addition, Polycam is used to 3D-scan the tools, generating high-fidelity textured mesh models as reliable references for high-precision 6-D pose estimation with FoundationPose.

Evaluation tasks: As shown in Fig. 7, we evaluate the TAC framework on five challenging bimanual collaborative manipulation tasks that involve precision, dexterity, contact, and object interactions: Flip fried egg, pull out nail, transport block, play football, and stow blanket. Task difficulty is increased by randomizing the initial and goal positions of objects (see Fig. 8). The properties, data, and evaluation settings of all tasks are summarized in Table I.

Evaluation metrics: In the experiments, success rate is used as the primary evaluation metric. To comprehensively assess generalization, we introduce three types of perturbation:

1) unseen objects, where novel items not used in training are introduced during evaluation;
2) camera disturbance, where the camera is randomly shaken to simulate visual disturbances; and
3) human interference, where the object is physically perturbed during robot execution, as shown in Fig. 9.

Baseline methods: We compare TAC with four representative SOTA baselines. Action chunking with transformers (ACT) [29] utilizes a CVAE and transformer decoder to predict action sequences. Diffusion policy (DP) [30] maps visual inputs directly to continuous actions, 3-D diffusion policy (DP3) [32] extends DP to 3-D observations, achieving better spatial reasoning, and hierarchical diffusion policy (HDP) [33] adopts a high–low hierarchical planning scheme similar to TAC, making it a particularly relevant baseline. For fairness, all baseline methods were trained on demonstration data collected via GELLO teleoperation by the same professional operator.

Evaluation of the data collection paradigm: To evaluate the data collection paradigm, we recruited four volunteers (two with teleoperation experience and two without) to collect data using natural human demonstration (TAC), GELLO [3], SpaceMouse [34], and Xbox [35], as shown in Fig. 10. The XBOX modality required collaboration between two operators. For fairness, a fixed number of demonstration videos was collected for each task under all methods, and DP was used as the training model. EDR denotes the percentage of collected demonstrations that are valid for policy training. SR denotes the average task success rate of the trained policy under the same robot rollout protocol, while operator expertise is analyzed through collection time and EDR.

Implementation details: The policy inputs RGB images resized to 128×128 and proprioceptive data, using a 2-frame observation history and a 16-step action horizon. The visual encoder utilizes a pretrained ResNet-18. For noise prediction, we employ a high-capacity 1-D Conditional U-Net (channels: $[512, 1024, 2048]$, kernel size: 5). The model was trained on a single NVIDIA RTX 4090 GPU using the AdamW optimizer (LR $= 10^{-4}$, weight decay 10^{-6}) with a batch size of 16 and a cosine decay schedule. The diffusion process used a squared-cosine noise schedule with 100 steps. Training ran for up to 3050 epochs with early stopping to prevent overfitting, and inference employed a 16-step DDIM sampler. To ensure a fair comparison, all image-based methods used the same ResNet-18 visual backbone, while DP3 retained its original point-cloud encoder. All methods were trained to convergence using the same number of task demonstrations.

Quantitative validation of data pipeline: Since obtaining ground truth 6-D poses from human videos is infeasible, we validated the accuracy of our annotation pipeline through a robot-mounted experiment. By rigidly fixing the tools to the UR5e end-effector and utilizing the robot's high-precision forward kinematics as ground truth, we evaluated FoundationPose on 100 diverse test samples covering the workspace with randomized positions and orientations. The pipeline achieved an

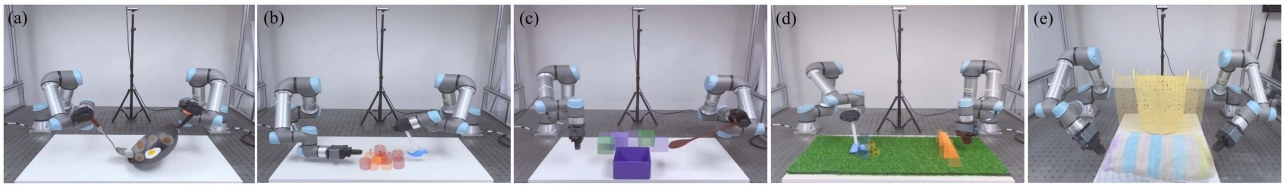

Fig. 8. Generalization tests across tasks. (a) *Flip fried egg*: Unseen egg types with randomized initial positions. (b) *Pull out nail*: Unseen nail and base colors and sizes, with randomized base initial positions. (c) *Transport block*: Unseen block colors with randomized initial positions. (d) *Play football*: Unseen ball colors with randomized initial positions of the ball and goal. (e) *Stow blanket*: Unseen blanket types with randomized initial positions of the blanket and basket.

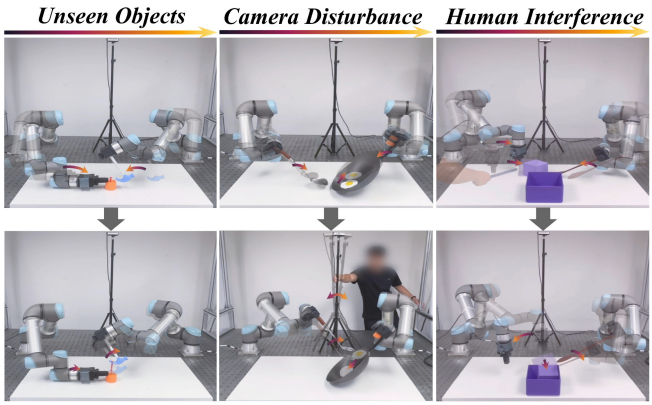

Fig. 9. Three perturbation tests in evaluation. From left to right, they include unseen objects, camera disturbance, and human interference by moving the target object during execution.

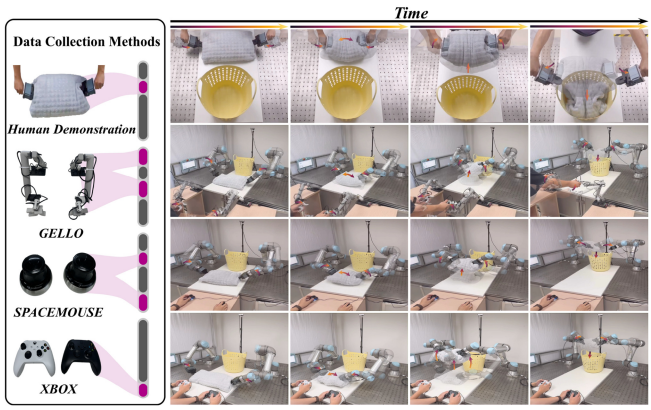

Fig. 10. Comparison of data collection methods.

average translation error of 5.2 mm and a rotation error of 2.8°. This accuracy falls well within the tolerance of closed-loop manipulation tasks, confirming the fidelity of our automated data generation.

B. Core Capabilities and Effectiveness of TAC

Based on Table II, under standard conditions, TAC achieved an average success rate of 81.5% across five tasks, significantly outperforming all baseline methods, including DP (5.9%), DP3 (15.1%), and HDP (19.3%). While ACT outperformed DP in stow blanket (16.7%) due to the temporal consistency of action chunking, it performed poorly in the highly dynamic play football task (0.0%), which requires rapid adaptation to changing object states.

TABLE II
PERFORMANCE COMPARISON OF POLICIES (%)

Policy/Task	PF[1]	SB[2]	FE[3]	PN[4]	TB[5]	SR (%)[6]
Standard						
ACT	0.0	16.7	0.0	9.1	4.0	6.0 ± 5.8
DP	3.6	11.1	0.0	9.1	8.0	5.9 ± 4.2
DP3	10.7	27.8	3.8	18.2	20.0	15.1 ± 6.4
HDP	14.3	33.3	7.7	22.7	24.0	19.3 ± 7.1
TAC (ours)	**75.0**	**88.9**	**80.8**	**81.8**	**84.0**	**81.5 ± 7.0**
Unseen Objects						
ACT	0.0	5.6	0.0	4.5	0.0	2.0 ± 2.5
DP	0.0	5.6	0.0	4.5	4.0	2.5 ± 2.8
DP3	7.1	22.2	0.0	13.6	12.0	10.1 ± 5.4
HDP	10.7	27.8	3.8	18.2	16.0	14.3 ± 6.3
TAC (ours)	**71.4**	**83.3**	**73.1**	**68.2**	**72.0**	**73.1 ± 8.0**
Human Interference						
ACT	0.0	0.0	0.0	0.0	0.0	0.0 ± 0.0
DP	0.0	0.0	0.0	0.0	0.0	0.0 ± 0.0
DP3	3.6	5.6	0.0	4.5	0.0	2.5 ± 2.8
HDP	7.1	11.1	0.0	4.5	4.0	5.0 ± 3.9
TAC (ours)	**75.0**	**83.3**	**76.9**	**72.7**	**76.0**	**76.5 ± 7.6**
Camera Disturbance						
ACT	0.0	0.0	0.0	0.0	0.0	0.0 ± 0.0
DP	0.0	0.0	0.0	0.0	0.0	0.0 ± 0.0
DP3	0.0	5.6	0.0	0.0	0.0	0.8 ± 1.6
HDP	3.6	11.1	0.0	4.5	0.0	3.4 ± 3.3
TAC (ours)	**67.9**	**88.9**	**80.8**	**77.3**	**80.0**	**78.2 ± 7.4**

[1]PF: Play Football; [2]SB: Stow Blanket; [3]FE: Flip Fried Egg; [4]PN: Pull Out Nail; [5]TB: Transport Block; [6]SR: average task success rate during evaluation.
The bold values indicate the best performance among all compared methods.

TAC's superiority lies in overcoming the inherent limitations of teleoperation-based data collection. In tasks, such as play football and flip fried egg, teleoperation data suffers from latency and nonintuitive control mappings, producing low-quality demonstrations with pauses and jitters. Consequently, all baseline methods fail, while TAC learns directly from natural human demonstrations, capturing smooth, agile, and precisely timed strategies essential for success.

In addition, as shown in the pull out nail task in Table III, under the identical conditions detailed in Section V with 80 demonstration videos, TAC achieved a success rate of 81.8% and an effective demonstration rate of 97.5%, highlighting its robustness and effectiveness in leveraging limited demonstration data.

C. Generalization and Robustness of TAC

We evaluated the generalization and robustness of TAC through unseen object tests. As shown in Table II, TAC achieved

TABLE III
COMPARISON OF DATA COLLECTION PARADIGMS

Data Source	Time (s)[1]	Ease	Accuracy	EDR (%)[2]	SR (%)[3]
Play Football					
GELLO	36 ± 6	Medium	Medium	65.0	3.6 ± 6.9
SPACEMOUSE	48 ± 12	Low	Low	51.3	0.0 ± 0.0
XBOX	60 ± 15	Low	Low	42.3	0.0 ± 0.0
TAC (ours)	**19 ± 3**	High	High	**98.8**	**75.0 ± 16.0**
Stow Blanket					
GELLO	32 ± 5	High	Medium	72.5	11.1 ± 14.5
SPACEMOUSE	40 ± 10	Medium	Low	61.3	11.1 ± 14.5
XBOX	47 ± 12	Low	Low	55.0	5.6 ± 10.6
TAC (ours)	**16 ± 2**	High	High	**95.0**	**88.9 ± 14.5**
Flip Fried Egg					
GELLO	21 ± 4	Medium	Low	58.9	0.0 ± 0.0
SPACEMOUSE	29 ± 8	Low	Low	43.3	0.0 ± 0.0
XBOX	38 ± 10	Low	Low	35.7	0.0 ± 0.0
TAC (ours)	**12 ± 2**	High	High	**96.7**	**80.8 ± 15.1**
Pull Out Nail					
GELLO	23 ± 4	High	Medium	78.8	9.1 ± 12.0
SPACEMOUSE	30 ± 8	Medium	Low	67.5	4.5 ± 8.7
XBOX	40 ± 10	Medium	Medium	71.3	9.1 ± 12.0
TAC (ours)	**13 ± 2**	High	High	**97.5**	**81.8 ± 16.1**
Transport Block					
GELLO	35 ± 5	Medium	Medium	68.6	8.0 ± 10.6
SPACEMOUSE	46 ± 11	Low	Low	57.1	4.0 ± 7.7
XBOX	55 ± 14	Low	Low	45.7	4.0 ± 7.7
TAC (ours)	**17 ± 3**	High	High	**94.3**	**84.0 ± 14.4**

[1]Time: average collection time per demonstration (mean ± std); [2]EDR: percentage of collected demonstrations valid for policy training; [3]SR: average task success rate during evaluation.
The bold values indicate the best performance among all compared methods.

an average success rate of 73.1% in these tests, significantly outperforming all baseline methods, especially HDP, which dropped from 19.3% to 14.3%. In contrast, ACT dropped to 0% under human interference due to its inability to recover from deviations during chunk execution. TAC's advantage lies in its focus on tool-environment interactions rather than object appearance, allowing it to learn more precise operational strategies. For instance, in transport block (72.0%) and pull out nail (68.2%), the policy successfully adapts to novel colors and sizes. This suggests that by removing hand-related visual variance, TAC forces the network to focus on geometric affordances, aligning the tool with the object's structural topology (e.g., cylindrical nail heads) rather than overfitting to specific textures.

In dynamic disturbance tests, TAC demonstrated strong robustness. For example, in the camera disturbance test, TAC achieved a success rate of 78.2%, nearly identical to the 81.5% under interference-free conditions. This stability is fundamentally enabled by our Gaussian Splatting-based data augmentation. By training on synthesized views that mimic physical jitter, the policy captures the intrinsic geometric relationship between the tool and the environment. In contrast, 2-D baselines suffer a severe performance degradation under such interference, as they rely on fragile pixel correspondences that break under camera shaking, resulting in success rates below 4%. TAC also showed strong resilience to human interference. When the nail base was shifted during evaluation, TAC still achieved a success rate of 72.7%, thanks to its continuous replanning mechanism, allowing it to quickly adapt to new target positions.

Furthermore, a comparison between TAC and the strongest baseline, HDP, under three perturbation conditions—unseen objects, human interference, and camera disturbance (see Fig. 11) shows that TAC consistently outperforms HDP in both average and maximum success rates across all conditions. The gap is most pronounced under camera disturbance, where HDP nearly fails due to its reliance on 2-D images, while TAC remains stable thanks to its 3-D scene representation.

TABLE IV
QUANTITATIVE ANALYSIS OF TRAJECTORY SMOOTHNESS

Policy	Mean Jerk ($\times 10^2$ m/s^3)	Action Var ($\times 10^{-3}$)
ACT	2.18 ± 0.55	1.75 ± 0.38
DP	4.52 ± 1.12	3.85 ± 0.92
DP3	3.10 ± 0.85	2.40 ± 0.65
HDP	2.25 ± 0.60	1.88 ± 0.45
TAC (ours)	**1.65 ± 0.42**	**1.20 ± 0.30**

The bold values indicate the best performance among all compared methods.

D. Trajectory Quality and Smoothness

The quality of generated motion trajectories is a key dimension of policy evaluation. We visualized the bimanual trajectories of four methods on the pull out nail task under standard conditions without external perturbations. As shown in Fig. 12(a) and (b), DP and DP3 produce noisy and jittery trajectories, reflecting the instability of end-to-end mappings in high-dimensional action spaces. Fig. 12(c) shows that HDP achieves some improvement, but its trajectories remain hesitant and indirect at critical collaboration points. In contrast, TAC in Fig. 12(d) generates smooth and clear trajectories, with the two arms naturally synchronized to effectively stabilize the base and successfully extract the nail.

Quantitatively, as shown in Table IV, we evaluated trajectory smoothness using mean jerk and action variance metrics. While DP3 (3.10) improves upon DP due to enhanced 3-D spatial stability, and HDP (2.25) mitigates oscillations through hierarchical planning, TAC achieves superior performance across both metrics, attaining the lowest mean jerk (1.65) and action variance (1.20). The reduced jerk indicates smoother transitions with fewer abrupt acceleration changes, while the lower action variance reflects more stable and consistent control signals over time. Together, these improvements lead to a substantial reduction in high-frequency jitter, by 63% compared to DP and 24% compared to ACT, demonstrating the stability of the proposed approach.

E. Efficiency and Usability of the Data Collection Paradigm

Beyond policy performance, Table III evaluates TAC's data collection efficiency and final outcomes. TAC consistently achieves higher efficiency than teleoperation devices across all five tasks, requiring substantially less demonstration time (e.g., 17 s versus 46 s with SPACEMOUSE in transport block), and also outperforms GELLO, with larger gains over six-DoF devices such as SPACEMOUSE and XBOX. These efficiency improvements do not compromise data quality, as TAC produces the highest effective data rate through a simple procedure and yields the strongest policy performance. Moreover, TAC significantly reduces dependence on operator expertise: while novices

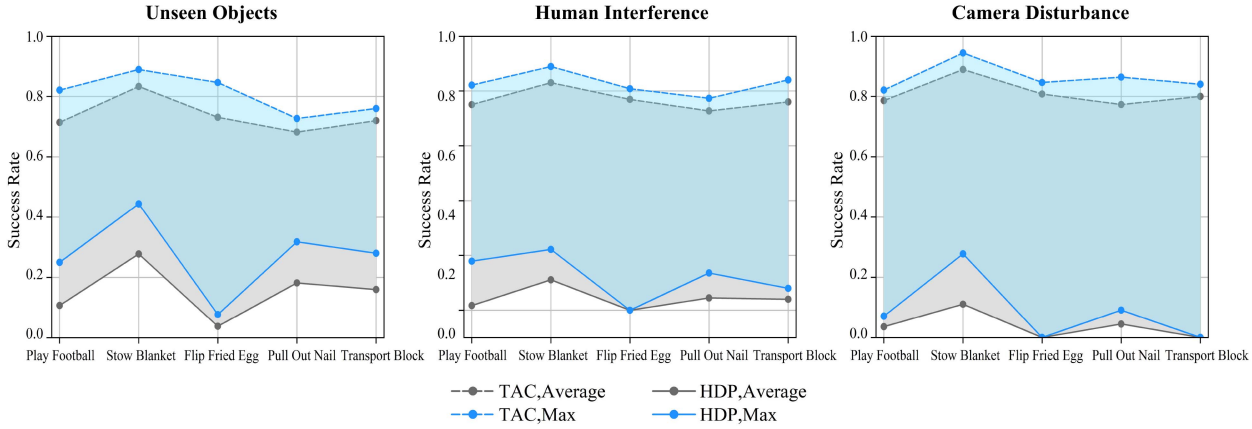

Fig. 11. Robustness comparison between TAC and HDP under three dynamic perturbations.

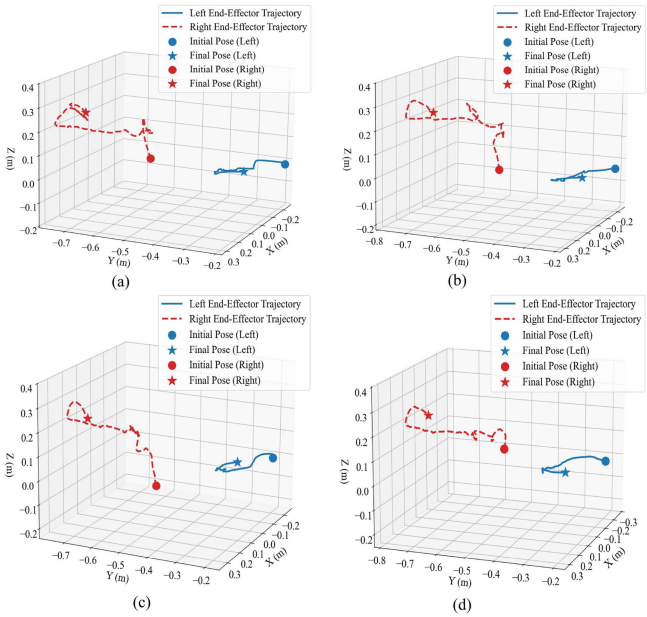

Fig. 12. Comparison of motion trajectories in the pull out nail task across four methods. Notably, ACT yields smooth trajectories similar to HDP and is omitted for layout clarity. (a) DP. (b) DP3. (c) HDP. (d) TAC.

TABLE V
AVERAGE EFFICIENCY AND QUALITY OF DATA COLLECTION BY USER EXPERTISE

Data Source	User Type	Time (s)		EDR (%)	
GELLO	Expert	29.4	↑ 128%	68.8	↓ 58.6%
	Novice	67.1		28.5	
SPACEMOUSE	Expert	38.6	↑ 142.8%	56.1	↓ 85.3%
	Novice	93.7		8.2	
XBOX	Expert	48.0	↑ 169.8%	50.0	↓ 74.6%
	Novice	129.5		12.7	
TAC (ours)	Expert	15.4	↓ 14.9%	96.5	↓ 23.2%
	Novice	13.1		74.1	↓ 23.2%

[1]EDR: percentage of collected demonstrations valid for policy training.
The bold values indicate the best performance among all compared methods.

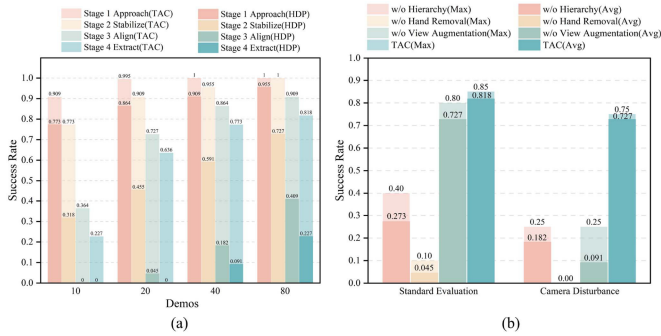

Fig. 13. Quantitative analysis of data efficiency and ablation studies. (a) Performance comparison between TAC and HDP under stage-wise evaluation. (b) Ablation results of three model variants.

using GELLO, SPACEMOUSE, and XBOX required 128%, 142.8%, and 169.8% more time than experts and achieved lower effective data rates (see Table V), novices using TAC completed tasks 14.9% faster than experts with a 74.1% effective data rate, demonstrating robustness to operator skill and enabling efficient, high-quality data collection.

F. Performance Under Data Scarcity

To demonstrate the advantages of TAC in data quality and sample efficiency, we conducted data-scarcity experiments on the pull out nail task. By using different amounts of training data (10, 20, 40, and 80 demonstrations), we compared TAC with the strongest baseline, HDP, while maintaining the 22 evaluation trials consistent with Table I. To analyze the results in depth, we broke the task down into four consecutive subgoals: Stage 1 (approach), Stage 2 (stabilize), Stage 3 (align), and Stage 4 (remove).

Fig. 13(a) clearly reveals the significant performance gap between TAC and HDP. In the case of extremely limited data (only 20 demonstrations), while HDP performed well in the simple approach stage (86.4%), it experienced a sharp drop in the critical alignment stage, with a success rate of only 4.5%. This indicates that teleoperation data lacks sufficient fine-grained manipulation details, making it difficult to support

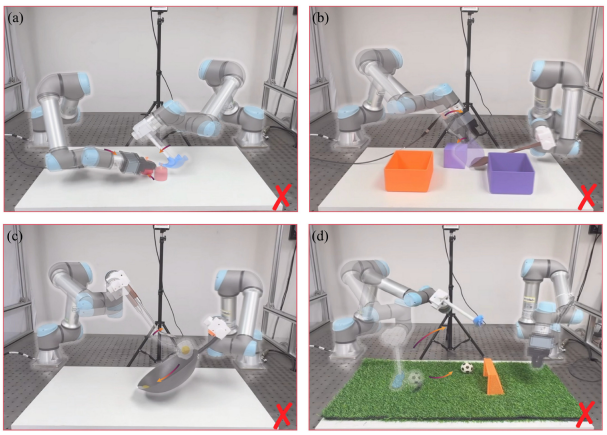

Fig. 14. Typical failure cases of the TAC framework. (a) Suboptimal initial state leads to an incomplete flip. (b) Perception errors prevent the hammer from accurately aligning with the nail head. (c) High-level policy selects an unstable pushing point, causing the block to topple. (d) Out-of-distribution ball position leads the high-level policy to choose a suboptimal contact point, resulting in a deviated kick.

precise operations. In contrast, under the same data conditions, TAC maintained robust performance in the alignment stage (with a success rate of 72.7%), ultimately achieving an overall success rate of 63.6%. This result suggests that natural human demonstrations provide smoother and more informative interaction trajectories for learning contact-rich behaviors, enabling TAC to learn complex interaction strategies even with very limited demonstrations.

G. Key Ablation Studies

To verify the necessity of each key design component in our framework, we conducted ablation studies on the pull out nail task, comparing the full TAC framework with three variants: one removing the hierarchical structure (w/o hierarchy), another retaining human hand features (w/o hand removal), and the last omitting view augmentation (w/o view augmentation). All models were trained on the same set of 80 demonstration videos. As shown in Fig. 13(b), removing the hierarchical structure significantly reduced the success rate from 81.8% to 27.3%, highlighting the importance of hierarchical decomposition for complex tasks. Retaining human hand features caused morphological confusion, reducing the success rate to 4.5% and causing complete failure under camera perturbations. The variant without view augmentation achieved 72.7% success under standard conditions but dropped to 9.1% under perturbations. In contrast, the full model with view augmentation maintained a 72.7% success rate even under perturbations, demonstrating the crucial role of view augmentation in enhancing robustness.

VI. Limitations and Future Work

Our method still exhibits limitations under challenging conditions. As shown in Fig. 14, failures mainly arise from inaccurate 6-D tool pose estimation under occlusion or motion blur, suboptimal high-level decisions in complex initial states, and limited generalization to out-of-distribution object configurations. These issues are partly caused by the dependence on FoundationPose and the limited coverage of human demonstrations. Furthermore, TAC's tool-centric design is less suitable for fine-grained gripper manipulation, such as pinching or handling small objects. Future work could investigate seamless switching between tool-centric and hand-centric manipulation, similar to UMI. The current framework also assumes rigid tools and, therefore, cannot readily handle deformable tools. More robust pose estimation and improved camera placement may further reduce failures caused by textureless surfaces, reflections, and occlusions.

VII. Conclusion

In this work, we present TAC, a novel imitation learning framework that tackles the challenges of bimanual manipulation and data scarcity by learning directly from natural human videos. TAC employs a hierarchical policy that decouples goal prediction from trajectory generation, combined with a tool-centric observation–action space and a multiview 3-D perception front-end, enabling precise, coordinated, and generalizable policies. Our key contribution is a new paradigm for data collection and policy learning that removes reliance on costly teleoperation hardware, lowering the barrier to imitation learning. Experiments demonstrate that TAC not only achieves an average success rate of 81.5%, exceeding the best baseline (HDP) by an absolute margin of 62.2%, but also maintains success rates of 73.1%, 76.5%, and 78.2% under unseen-object, human-interference, and camera-disturbance settings, respectively. Importantly, it transforms robot demonstration from an expert-dependent procedure into an intuitive process accessible to all, enabling novices to collect high-quality data 14.9% faster than experts, thereby substantially improving efficiency without compromising quality. We believe TAC represents a step toward scalable and low-cost imitation learning for complex robotic manipulation.

References

[1] S. Stepputtis, M. Bandari, S. Schaal, and H. B. Amor, "A system for imitation learning of contact-rich bimanual manipulation policies," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 11810–11817.
[2] X. Xu, M. You, H. Zhou, Z. Qian, and B. He, "Robot imitation learning from image-only observation without real-world interaction," *IEEE/ASME Trans. Mechatron.*, vol. 28, no. 3, pp. 1234–1244, Jun. 2023.
[3] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel, "GELLO: A general, low-cost, and intuitive teleoperation framework for robot manipulators," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2024, pp. 12156–12163.
[4] C. Chi et al., "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," in *Proc. Robot.: Sci. Syst.*, 2024.
[5] Zhaxizhuoma, "FastUMI: A scalable and hardware-independent universal manipulation interface with dataset," in *Proc. Mach. Learn. Res.*, 2025, vol. 305, pp. 3069–3093.
[6] Z. Lu, N. Wang, and C. Yang, "A constrained DMPs framework for robot skills learning and generalization from human demonstrations," *IEEE/ASME Trans. Mechatron.*, vol. 26, no. 6, pp. 3265–3275, Dec. 2021.
[7] J. Yu et al., "Real2Render2Real: Scaling robot data without dynamics simulation or robot hardware," in *Proc. Mach. Learn. Res.*, 2025, vol. 305, pp. 547–577.
[8] H. Zhou et al., "You only teach once: Learn one-shot bimanual robotic manipulation from video demonstrations," in *Proc. Robot.: Sci. Syst.*, 2025.

[9] J. Liu, H. Sim, C. Li, K. C. Tan, and F. Chen, "BiRP: Learning robot generalized bimanual coordination using relative parameterization method on human demonstration," in *Proc. 62nd IEEE Conf. Decis. Control*, 2023, pp. 8300–8305.
[10] M. Deniša, A. Gams, A. Ude, and T. Petrič, "Learning compliant movement primitives through demonstration and statistical generalization," *IEEE/ASME Trans. Mechatron.*, vol. 21, no. 5, pp. 2581–2594, Oct. 2015.
[11] J. Grannen, Y. Wu, B. Vu, and D. Sadigh, "Stabilize to act: Learning to coordinate for bimanual manipulation," in *Proc. Mach. Learn. Res.*, 2023, vol. 229, pp. 563–576.
[12] P. Zhou et al., "Bimanual deformable bag manipulation using a structure-of-interest based neural dynamics model," *IEEE/ASME Trans. Mechatron.*, vol. 30, no. 5, pp. 3254–3265, Oct. 2025.
[13] T. Yu et al., "ManiGaussian: General robotic bimanual manipulation with hierarchical Gaussian world model," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2025, pp. 12232–12239.
[14] Y. Zhu, P. Stone, and Y. Zhu, "Bottom-up skill discovery from unsegmented demonstrations for long-horizon robot manipulation," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 4126–4133, Apr. 2022.
[15] H. R. Walke et al., "BridgeData V2: A dataset for robot learning at scale," in *Proc. Mach. Learn. Res.*, 2023, vol. 229, pp. 1723–1736.
[16] H.-S. Fang et al., "Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2024, pp. 653–660.
[17] A. Brohan et al., "RT-1: Robotics transformer for real-world control at scale," in *Proc. Robot.: Sci. Syst.*, 2023.
[18] A. Brohan et al., "RT-2: Vision-language-action models transfer web knowledge to robotic control," in *Proc. Mach. Learn. Res.*, 2023, vol. 229, pp. 2165–2183.
[19] O. Model et al., "Octo: An open-source generalist robot policy," in *Proc. Robot.: Sci. Syst.*, 2024.
[20] H. Choi et al., "On the use of simulation in robotics: Opportunities, challenges, and suggestions for moving forward," *Proc. Nat. Acad. Sci.*, vol. 118, no. 1, 2021, Art. no. e1907856118.
[21] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robot. Auton. Syst.*, vol. 57, no. 5, pp. 469–483, 2009.
[22] T. D. Kulkarni, K. R. Narasimhan, A. Saeedi, and J. B. Tenenbaum, "Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, vol. 29, pp. 3675–3683.
[23] M. E. Taylor and P. Stone, "Transfer learning for reinforcement learning domains: A survey," *J. Mach. Learn. Res.*, vol. 10, no. 7, pp. 1633–1685, 2009.
[24] V. Leroy, Y. Cabon, and J. Revaud, "Grounding image matching in 3D with MASt3R," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 71–91.
[25] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian splatting for real-time radiance field rendering," *ACM Trans. Graph.*, vol. 42, no. 4, 2023, Art. no. 139.
[26] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "FoundationPose: Unified 6D pose estimation and tracking of novel objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 17868–17879.
[27] T. Ren et al., "Grounded SAM: Assembling open-world models for diverse visual tasks," 2024, *arXiv:2401.14159*.
[28] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 6840–6851, 2020.
[29] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Proc. Robot.: Sci. Syst.*, 2023.
[30] C. Chi et al., "Diffusion policy: Visuomotor policy learning via action diffusion," *Int. J. Robot. Res.*, vol. 44, vol. 10, pp. 1684–1704, 2025.
[31] M. Neunert et al., "Fast nonlinear model predictive control for unified trajectory optimization and tracking," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2016, pp. 1398–1404.
[32] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3D Diffusion Policy: Generalizable visuomotor policy learning via simple 3D representations," in *Proc. Robot.: Sci. Syst.*, 2024.
[33] D. Wang, C. Liu, F. Chang, and Y. Xu, "Hierarchical diffusion policy: Manipulation trajectory generation via contact guidance," *IEEE Trans. Robot.*, vol. 41, pp. 2086–2104, 2025.
[34] V. Dhat, N. Walker, and M. Cakmak, "Using 3D mice to control robot manipulators," in *Proc. ACM/IEEE Int. Conf. Hum.-Robot Interaction*, 2024, pp. 896–900.
[35] G. Lu, T. Yu, H. Deng, S. S. Chen, Y. Tang, and Z. Wang, "AnyBimanual: Transferring unimanual policy for general bimanual manipulation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2025, pp. 13662–13672.
[36] H. Chen, C. Zhu, S. Liu, Y. Li, and K. R. Driggs-Campbell, "Tool-as-interface: Learning robot policies from observing human tool use," in *Proc. Mach. Learn. Res.*, 2025, vol. 305, pp. 2695–2713.

Yechen Fan (Graduate Student Member, IEEE) received the B.S. degree in mechanical design, manufacturing, and automation from Jinling Institute of Technology, Nanjing, China, in 2024. He is currently working toward the M.S. degree in mechanical engineering with the School of Mechanical Engineering and Automation, Fuzhou University, Fuzhou, China.

His research focuses on robotic manipulation through multimodal robot learning, including simulation, teleoperation, and human video.

Xinjie Zhang received the B.S. degree in mechatronics engineering from Anhui University of Science and Technology, Huainan, China, in 2024. He is currently working toward the M.S. degree in mechanical engineering with the School of Mechanical Engineering and Automation, Fuzhou University, Fuzhou, China.

His research interests include UAV control and ground station design.

Jianghao Zhao received the B.S. degree in mechanical design and manufacturing and automation from Hangzhou Normal University, Hangzhou, China, in 2024. He is currently working toward the M.S. degree in mechanical manufacturing and automation with the School of Mechanical Engineering and Automation, Fuzhou University, Fuzhou, China.

His research interests include robot thermal protection and UAV under-rotor bevel gear transmission.

Xiaohan Liu received the B.S. degree in mechanical design and manufacturing and automation from Quanzhou Normal University, Quanzhou, China, in 2024. She is currently working toward the M.S. degree in mechanical engineering with the School of Mechanical Engineering and Automation, Fuzhou University, Fuzhou, China.

Her research interests include safety in human–robot interaction.

Haibin Wu received the M.S. degree in machine design from Liaoning Technical University, Fuxin, China, in 1999, and the Ph.D. degree in mechanical engineering from Zhejiang University, Hangzhou, China, in 2002.

Since 2002, he has been with the School of Mechanical Engineering and Automation, Fuzhou University, Fuzhou, China. He is currently a Professor with the School of Mechanical Engineering and Automation, Fuzhou University. His main research topics are safe human–robot interaction, robot control, and robot skin.

Jinhua Ye (Senior Member, IEEE) received the B.S. degree in mechanical engineering and automation from Fuzhou University, Fuzhou, China, in 2005, the M.S. degree in mechatronics engineering from Fuzhou University, Fuzhou, China, in 2008, and the Ph.D. degree in mechatronics engineering from the South China University of Technology, Guangzhou, China, in 2014.

Since 2014, he has been with the School of Mechanical Engineering and Automation, Fuzhou University, Fuzhou, China. He is currently an Associate Professor with the School of Mechanical Engineering and Automation, Fuzhou University. His research interests mainly include human–robot safety interaction, robot contact perception, and robot welding.

Gengfeng Zheng (Senior Member, IEEE) received the B.S. degree in bioelectronic information engineering from Jiangsu University, Zhenjiang, China, in 2006, the M.S. degree in electromechanical engineering from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China, in 2008, and the Ph.D. degree in electromechanical engineering from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China, in 2011.

He joined Fujian Special Equipment Inspection and Research Institute in 2011 and was an Engineer in 2012, a Senior Engineer in 2013, and a Technical Director in 2016. He is currently the Head of Fujian Key Laboratory of Special Intelligent Equipment Safety Measurement and Control and Fujian Engineering Research Center of Special Robots for Safe Operations. His research interests are in the areas of special-purpose robots, human–robot integration, intelligent operation, and emergency response.