

Vision-Force ACT: Multimodal Action Chunking for Robotic Bimanual Peg-in-Hole Assembly

Mingqi Chen¹, Meijie Chen¹, Qiang Li¹, Zhong Ming¹ and Chao Liu²

Abstract—Robotic bimanual peg-in-hole assembly requires accurate geometric alignment and compliant manipulation. Vision-based imitation learning is widely used in such kind of bimanual manipulation, which can efficiently locate and approach the target. However, visual observations could be ambiguous when the peg, hole or contact surface is occluded by the manipulators. This paper presents *Vision-Force ACT*, a multimodal extension of Action Chunking with Transformers (ACT) that incorporates bilateral wrist force/torque observations into both the conditional latent inference and the policy representation. The proposed method successfully outputs temporally coherent bimanual action chunks from synchronized images, joints state, and wrist F/T signals. A Mujoco-based simulated bimanual assembly platform equipped with 6-axis force/torque (F/T) sensor at both wrists is used in the experiment, and the proposed method is compared with the vision-only ACT baseline under identical training settings. The results show that Vision-Force ACT improves the task success rate from 40% to 68% with 50 demonstrations and from 60% to 78% with 200 demonstrations. The recorded F/T curves are also smoother than those of the vision-only policy and show closer agreement with the force demonstrations. These preliminary results support force sensing as an effective complement to visual observations for contact-rich bimanual imitation learning, while also motivating broader multi-task and real-robot validation.

I. INTRODUCTION

Bimanual manipulation is important for assembly tasks that require simultaneous stabilization, alignment, and insertion. In a typical bimanual peg-in-hole task, one manipulator stabilizes the receiving part, while the other aligns and inserts the mating part. Although the free-space motion can largely be guided by visual geometry, insertion requires precise inter-arm coordination, accurate relative positioning, and compliant responses to contact. Even small position or orientation errors may therefore cause edge contact, jamming, or insertion failure. End-to-end imitation learning avoids manually designing separate perception, alignment, contact-detection, and force-control modules. In particular, Action Chunking with Transformers (ACT) predicts temporally coherent action chunks instead of isolated actions, reducing the effective decision horizon and the accumulation of errors in fine-grained bimanual manipulation [1]. Its ability to jointly model multi-view images, robot proprioception, and coordinated bimanual actions makes ACT a suitable foundation for learning insertion skills from demonstrations.

However, standard ACT mainly conditions its predictions on visual observations and robot proprioception, making the physical interaction state only indirectly observable. This limitation becomes critical after contact, when the mating interface may be occluded by the manipulators, grippers, or objects. Moreover, correct insertion, edge contact, and jamming may produce similar visual and proprioceptive observations while requiring different, or even opposite, actions. A vision-only policy may recognize these conditions only after visible displacement or motion failure occurs. As a result, the robot may continue moving in the planned direction after improper contact, increasing the contact force and pose error before the problem becomes visible. In bimanual manipulation, the F/T signal at the inserting manipulator also depends on whether the other manipulator holds the receiving part firmly or allows it to slip. Bilateral six-axis F/T measurements directly indicate contact onset, loading direction, asymmetric resistance, and rotational misalignment, complementing the geometric information provided by vision. Previous visual-force learning methods have shown that interaction information can improve manipulation robustness [2], [3], [4] and reduce undesirable contact forces [5], [6]. Nevertheless, integrating bilateral F/T observations into ACT without losing its unified end-to-end action-chunking structure remains insufficiently studied.

This work develops a force-aware ACT policy for simulated bimanual peg-in-hole assembly, where one manipulator stabilizes or positions the receiving component while the other coordinates the insertion motion. Bilateral six-axis wrist F/T measurements are used as additional observations to provide contact information during action-chunk prediction. The key idea is to encode the F/T data as a sequence of features and use the resulting contact information in both the ACT latent posterior and the action decoder.

The main contributions of this work are as follows. First, we extend ACT with bilateral wrist F/T observations, enabling contact-aware action-chunk prediction while preserving the original CVAE Transformer policy structure. Second, we build a MuJoCo-based bimanual peg-in-hole assembly benchmark with virtual wrist F/T sensors and evaluate the proposed policy against a vision-only ACT baseline. The results show a clear improvement in task success rate and suggest that force observations are especially useful during contact-rich approach and insertion.

The remainder of the paper is organized as follows: Section II provides a review of related work, Section III proposes Vision-Force ACT, Section IV introduces the experimental setup, and Section V gives the results. Section VI discuss

*This research is supported by the Natural Science Foundation of Top Talent of SZTU (Grant No. GDRC202411).

¹Mingqi Chen, Meijie Chen, Qiang Li and Zhong Ming are with the Dexterous Robotics Lab, Shenzhen Technology University, Shenzhen, China

²Chao Liu is with Department of Robotics, LIRMM, CNRS- University of Montpellier, 34095 Montpellier, France

Corresponding author: Qiang Li (liqiang1@sztu.edu.cn)

TABLE I: Representative Learning Approaches for Contact-Rich Manipulation

Method	Vision	Contact sensing	Bimanual	Learning/control form	Main distinction
ACT [1]	Yes	No	Yes	CVAE Transformer imitation	Predicts temporally coherent action chunks from images and proprioception.
Bi-Touch [2]	Optional	Tactile images	Yes	Sim-to-real deep RL	Learns bimanual tactile behaviors with a tactile-domain translation pipeline.
3D-ViTac [4]	Yes	Dense tactile arrays	Yes	Visuo-tactile policy learning	Uses spatially dense tactile sensing for fine-grained two-hand manipulation.
MOMA-Force [5]	Yes	6-D F/T	No	Imitation + admittance control	Predicts motion and desired wrench for contact-rich mobile manipulation.
ForceMimic [6]	Limited	Force-motion capture	No	Hybrid force-motion imitation	Collects force-centric demonstrations and fits hybrid force-position primitives.
FoAR [7]	Yes	High-rate F/T	No	Contact-aware reactive policy	Modulates multimodal fusion using a future-contact predictor.
Vision-Force ACT	Yes	Bilateral 6-D F/T	Yes	CVAE Transformer imitation	Conditions bimanual action chunks and latent inference on wrist F/T signals.

the whole work and Section VII concludes the paper.

II. RELATED WORK

A. Imitation Learning for Bimanual Manipulation

Bimanual manipulation requires coordinated motion between two manipulators, especially in tasks involving grasping, alignment, and contact-rich insertion. Imitation learning provides an effective way to learn such behaviors from demonstrations without manually designing a complete sequence of perception, planning, and control modules. Transformer-based policies are suitable for this setting since they can model temporal dependencies among multi-view observations, proprioceptive states, and future actions [8], [9]. ACT combines a conditional variational autoencoder (CVAE) [10] with action chunking and temporal aggregation, allowing the policy to represent variability in demonstrations while predicting temporally coherent action sequences [1]. Compared with single-step action prediction, action chunking shortens the effective decision horizon and reduces compounding errors in fine-grained bimanual manipulation.

Standard ACT mainly relies on visual observations and robot proprioception. This is effective for many free-space motions, such as reaching and approaching, but it becomes less reliable once physical contact dominates the task state. In peg-in-hole assembly, correct insertion, edge contact, and jamming may look similar in camera images and joint states, while requiring different corrective actions. This motivates our extension of ACT with bilateral wrist force/torque observations. Instead of replacing ACT with a different policy class, we keep the compact CVAE Transformer action-chunking structure and study whether explicit contact observations can improve bimanual insertion.

B. Visual-Force and Visual-Tactile Manipulation

Force and tactile sensing provide direct information about physical interaction that is difficult to infer from images alone. For bimanual manipulation, Bi-Touch uses tactile observations and sim-to-real reinforcement learning to improve contact-rich behaviors [2], while 3D-ViTac shows that dense tactile sensing can support fine-grained two-hand manipulation [4]. Visuo-tactile Transformer models further

demonstrate that self-attention and cross-modal attention can be used to fuse visual and tactile features for manipulation [3]. These works show the value of contact sensing, but they often require specialized tactile sensors or adopt learning formulations different from ACT. Another line of work uses force/torque sensing for contact-rich manipulation. MOMA-Force combines visual imitation with force information for mobile manipulation and shows that force observations can improve robustness and reduce undesirable contact behavior [5]. ForceMimic collects force-motion demonstrations and learns contact-rich skills from force-centric demonstrations [6]. FoAR introduces force-aware reactive modulation for contact-rich manipulation using high-rate force/torque sensing [7]. These studies confirm that contact signals are useful for recognizing and correcting physical interaction states.

Table I summarizes representative learning approaches for contact-rich manipulation and clarifies the position of our method. Compared with ACT, Vision-Force ACT explicitly observes contact through bilateral wrist F/T sensors. Compared with tactile-based methods, it uses standard wrist force/torque measurements instead of dense tactile arrays. Compared with force-aware manipulation methods, it does not introduce a separate admittance controller, force primitive, or reactive modulation module. Instead, it retains the original CVAE Transformer policy structure and uses F/T measurements as an additional condition for both latent inference and action prediction. Therefore, our method occupies a complementary position: a bimanual action-chunking policy that incorporates contact sensing while preserving the simplicity of end-to-end ACT imitation learning.

III. METHOD

A. Problem Formulation

We consider bimanual peg-in-hole assembly with synchronized visual, proprioceptive, and wrist F/T observations. At time step t , the robot receives multi-view RGB images $I_t = \{I_t^c\}_{c=1}^C$, the bimanual joint state q_t , and bilateral wrist F/T measurements

$$f_t = [w_t^L; w_t^R] \in \mathbb{R}^{12}, \quad w_t = [F_x, F_y, F_z, \tau_x, \tau_y, \tau_z]. \quad (1)$$

Here, w_t^L and w_t^R denote the six-axis measurements from the left and right wrists. The goal is to learn a policy that predicts

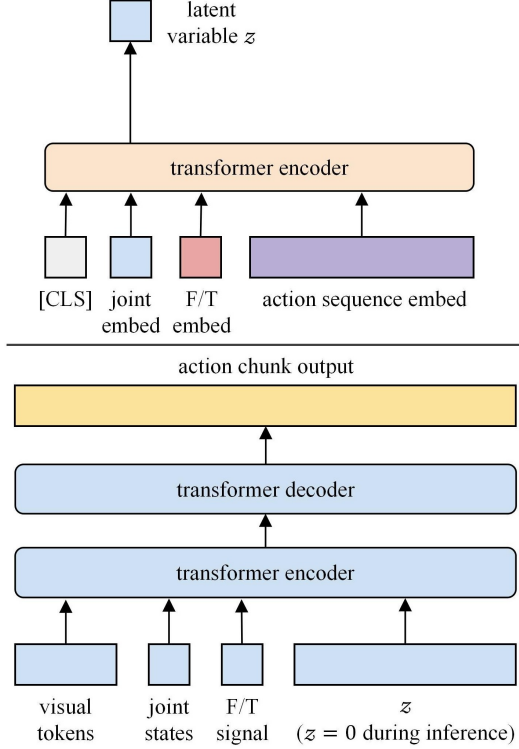


Fig. 1: Vision-Force ACT architecture. The bilateral F/T embedding is used as an explicit condition in both the latent encoder (upper part) and the action decoder (lower part). During inference, the latent variable is set to zero, while the F/T input remains active for contact-aware action prediction.

a future action chunk $\hat{A}_t = [\hat{a}_t, \dots, \hat{a}_{t+K-1}]$, where each action contains the joint-space commands of both manipulators.

Different from vision-only ACT, the proposed policy explicitly conditions action prediction on f_t . Thus, the policy can be written as

$$\hat{A}_t = \pi_{\theta}(I_t, q_t, f_t, z), \quad (2)$$

where z is the latent variable used by the ACT-style CVAE. During execution, overlapping action chunks are temporally aggregated following ACT [1]. The key difference is that the predicted chunk is conditioned not only on visual geometry and joint states, but also on the current contact-related F/T observation.

B. Force-Conditioned ACT

Fig. 1 shows the proposed Vision-Force ACT architecture. The visual images are encoded into visual tokens, and the joint state is embedded as a proprioceptive token. The bilateral F/T signal is normalized using training-set statistics and projected into an F/T token. This token is used in two places: the CVAE latent encoder and the Transformer policy decoder.

During training, the latent encoder receives the current joint embedding, the demonstrated action sequence embed-

TABLE II: Complementary Cues Across Assembly Phases

Phase	Visual cue	F/T cue
Pick/lift	Part pose and manipulator configuration	Load change from grasping or slippage
Approach	Relative pose between peg and hole	Contact onset, side, and normal direction
Insertion	Small and ambiguous image change	Resistance, edge contact, jamming, or release

ding, the class token, and the F/T embedding. It estimates the posterior distribution

$$q_{\phi}(z | A_t, q_t, f_t) = \mathcal{N}(\mu_{\phi}, \text{diag}(\sigma_{\phi}^2)). \quad (3)$$

This design makes the latent variable contact-aware, since the demonstration action chunk is explained together with the measured wrist interaction signal.

The lower part of Fig. 1 corresponds to action prediction. The Transformer policy encoder fuses the visual tokens, joint token, F/T token, and latent variable z . The Transformer decoder then outputs the bimanual action chunk:

$$\hat{A}_t = \pi_{\theta}(I_t, q_t, f_t, z). \quad (4)$$

At inference time, z is set to zero as in deterministic ACT deployment, while the F/T input remains active. Therefore, the policy still receives contact information during execution and can generate action chunks conditioned on the current wrist force/torque state.

C. Role of the Force Modality

The force modality is introduced to reduce the ambiguity of contact states in bimanual insertion. In free-space motion, visual observations provide the main information about object poses, manipulator configurations, and the relative position between the peg and the hole. After contact, however, the contact interface may be partly occluded, and different states such as correct insertion, edge contact, and jamming can produce similar images and proprioceptive observations. Wrist F/T measurements provide direct information about contact onset, loading direction, and abnormal resistance, which makes them complementary to visual observations.

Table II summarizes this complementarity across the main phases of the task. Vision is more important for global geometric guidance, while F/T sensing becomes more informative when physical contact dominates the task state.

This complementarity is especially important in bimanual assembly. The force measured at the inserting manipulator is affected not only by the local peg-hole contact, but also by whether the other manipulator holds the receiving part firmly or allows it to move. Therefore, the left and right wrist measurements are encoded separately before fusion, preserving bilateral contact information instead of reducing it to a single global force magnitude.

Within Vision-Force ACT, the F/T features are used in both the latent posterior and the action decoder. During training, they help the latent variable represent demonstration differences caused by contact conditions. During inference,

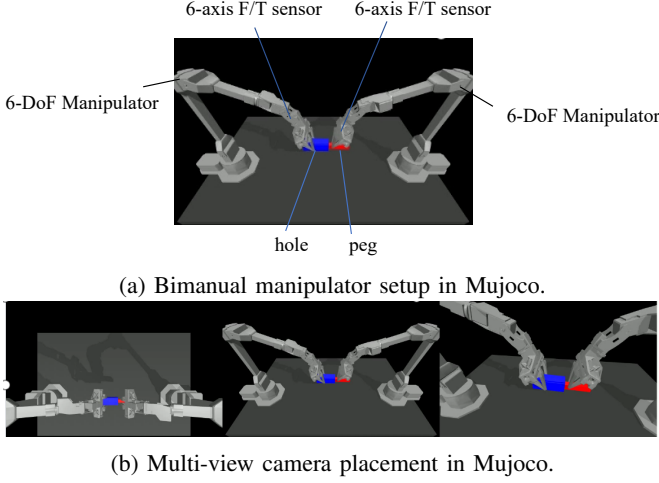


Fig. 2: Simulation setup in Mujoco, where ALOHA robot [1] with two 6-DoF manipulators is used.

they provide updated contact information for action prediction, allowing the policy to react to unexpected contact or jamming even when visual changes are small.

D. Training Objective

The network is trained with the standard ACT reconstruction and KL regularization losses:

$$\mathcal{L} = \frac{1}{K} \sum_{i=0}^{K-1} \|\mathbf{a}_{t+i} - \hat{\mathbf{a}}_{t+i}\|_1 + \beta D_{\text{KL}}(q_\phi(\mathbf{z} | \cdot) \| \mathcal{N}(0, \mathbf{I})). \quad (5)$$

where β is the weight of KL divergence loss, which is set to 10 in this work.

IV. EXPERIMENTAL SETUP

A. Bimanual Assembly Environment

A bimanual simulation for peg-in-hole assembly is built in Mujoco to evaluate the proposed method, as shown in Fig. 2. The manipulators used is the original ALOHA setup [1] with two 6-DoF manipulators with a single DoF gripper. Three camera views observe the workspace, while each wrist is equipped with a virtual six-axis force/torque sensor. Demonstrations contain synchronized images, joint states, action sequences, and bilateral F/T measurements. The task involves coordinated reaching, alignment, contact establishment, and insertion. Occlusion is common near the contact interface since the manipulators and grasped parts block the camera views.

The simulation provides repeatable multimodal data and avoids the safety risk of early-stage contact experiments. It also exposes two limitations relevant to later transfer: contact dynamics depend on friction and compliance parameters, and simulated rendering or sensing may not fully match real hardware.

B. Compared Policies and Metrics

In the experiments we compare the proposed Vision-Force ACT with vision-only ACT, where the observation are different as follows:

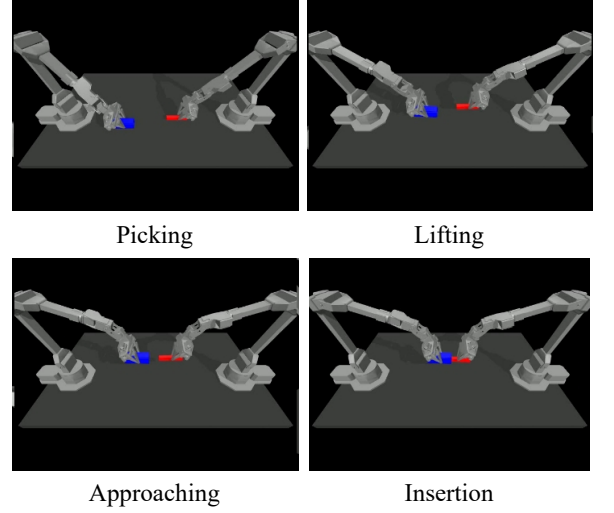


Fig. 3: A successful rollout via the proposed strategy, where the sub-processes are also shown.

- 1) **Vision ACT**: multi-view images and joint states only;
- 2) **Vision-Force ACT**: the same policy augmented with bilateral F/T tokens.

For fair comparison, both policies are trained under identical settings with either 50 or 200 demonstrations. Each demonstration contains synchronized three-view 480×640 images, 7×2 joint states, 6×2 F/T signals, and action chunks, with varied initial peg and hole poses. The network uses a ResNet18 visual backbone, 4 encoder layers, 7 decoder layers, a 512-dimensional latent space, batch size 8, learning rate 10^{-5} , 30000 epochs, and chunk size 50. Thus, the comparison isolates the contribution of the F/T modality.

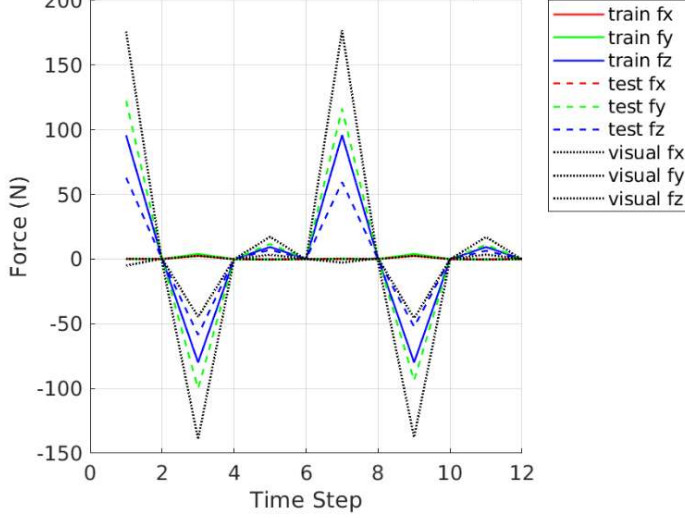
C. Controlled Evaluation Protocol

We evaluate vision-only ACT and the proposed Vision-Force ACT under the same task setting. For each training-data setting, the policy is tested in 50 trials, where the initial poses of the peg and the hole component are randomly varied within the same predefined range. Therefore, the “50 samples” and “200 samples” in Table III refer to the number of collected demonstrations, while the reported success rate is computed from 50 evaluation trials. The main evaluation metric is the task success rate.

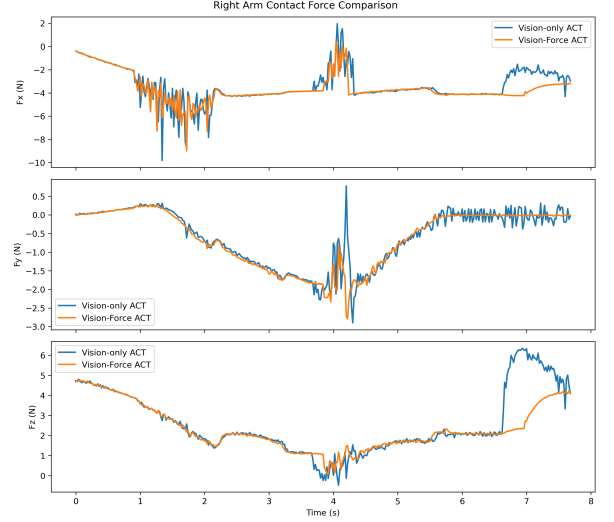
To determine success objectively, a virtual block is placed inside the hole at a predefined depth. A trial is counted as successful only when the peg reaches and contacts this virtual block, indicating that the peg has been inserted to the required position. Failures mainly include unsuccessful grasping, object slippage, alignment failure, jamming, and edge contact without completing insertion.

In addition to the success rate, we qualitatively compare the wrist force signals measured during execution. The comparison focuses on whether the force profiles are smooth and stable during contact-rich alignment and insertion.

Right-arm Force: Three Files Comparison



(a) Demonstration, vision-force test, and vision-only test force profiles.



(b) Force curves during real-time execution.

Fig. 4: Force behavior in the assembly task, which reports smoother and more compliant contact for the multimodal policy.

TABLE III: Success Rate under Different Training Data Sizes

Policy	Training data size	
	50 samples	200 samples
Vision ACT	40%	60%
Vision-Force ACT	68%	78%

TABLE IV: Stage-wise Completion Rates under 50-Sample Training Setting

Stage of completion	Vision ACT	Vision-Force ACT	Gain
Grasp	96%	98%	+2 pp
Lift	94%	94%	0 pp
Approach	68%	86%	+18 pp
Insertion	40%	68%	+28 pp

V. RESULTS

A. Quantitative Task Evaluation

Table III reports the overall task completion performance under two training-data sizes. With 50 demonstrations, Vision-Force ACT improves the success rate from 40% to 68%, corresponding to an absolute gain of 28 percentage points. When the number of demonstrations is increased to 200, both policies achieve higher success rates, while Vision-Force ACT still outperforms the vision-only baseline by 18 percentage points, improving from 60% to 78%. These results indicate that bilateral F/T observations consistently improve task completion, with a larger gain under the smaller-data setting.

Fig. 3 illustrates the overall execution process, including picking, lifting, approaching, and insertion. Since the two policies use the same robot platform and task procedure, their successful visual rollouts can appear similar. The difference between the two policies is therefore mainly reflected in the success rates and the wrist force profiles.

Table IV further reports the stage-wise completion rates under the 50-demonstration setting. The two policies perform similarly in the early grasping and lifting stages, while the gap becomes clear in the later contact-rich stages. Vision-Force ACT improves the approach-stage completion rate from 68% to 86% and the final insertion success rate from 40% to 68%. This suggests that force sensing mainly

contributes to contact-rich alignment and insertion, where visual observations alone are often insufficient to distinguish correct alignment, edge contact, and jamming.

B. Contact-Force Behavior

Fig. 4 compares the wrist force profiles during execution. The two policies produce similar average force magnitudes, indicating that the improvement does not simply come from reducing the overall contact load. However, Vision-Force ACT shows smaller force fluctuations. For the force norm, Vision ACT obtains 4.75 ± 1.03 N, while Vision-Force ACT obtains 4.67 ± 0.74 N. For the axial component F_z , the standard deviation also decreases from 1.56 N to 1.11 N, with comparable mean values of 2.59 N and 2.34 N.

These results indicate that the multimodal policy produces a more stable contact-force response, which is consistent with the smoother force curves in Fig. 4(b). The current analysis mainly reflects the force components, while the contribution of torque is not separately quantified and should be examined in future ablation studies.

VI. DISCUSSION

The results show that bilateral wrist F/T observations improve bimanual peg-in-hole assembly across different training-data sizes. The improvement is larger with 50 demonstrations and remains visible with 200 demonstrations,

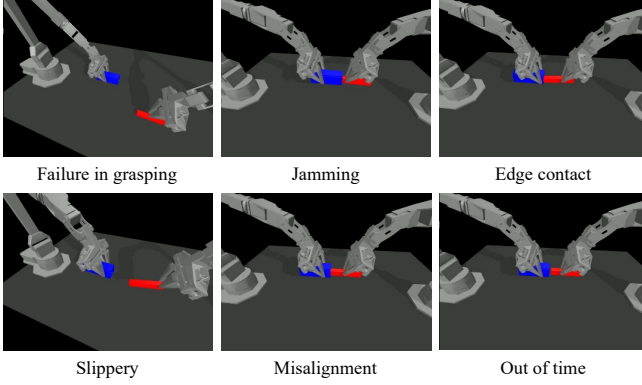


Fig. 5: Failure cases of the proposed method.

suggesting that force sensing provides complementary information rather than only compensating for insufficient training data. The stage-wise completion rates further indicate that both policies can usually complete grasping and lifting, while the main advantage of Vision-Force ACT appears during approaching and insertion. This supports the motivation that force sensing is more useful when visual observations become ambiguous due to occlusion, small pose errors, edge contact, or jamming.

Fig. 5 shows the main failure cases, including failure in grasping, object slippage, jamming, edge contact, misalignment, and out-of-time execution. These cases suggest that force observations alone cannot solve all assembly errors. Grasp failure and slippage are mainly related to unstable pre-insertion manipulation, while misalignment, edge contact, and jamming are caused by insufficient correction after contact. Therefore, the current method improves contact-state perception, but it does not fully address recovery from unfavorable contact.

A key limitation is that the current policy is still trained from demonstration trajectories without rich corrective behaviors, such as retracting, re-aligning, rotating, or reducing force after jamming. Thus, although F/T observations are used as policy inputs, the learned behavior is not an adaptive force-control process in the strict sense. The policy can condition action chunks on force observations, but it does not explicitly perform closed-loop force regulation or impedance/admittance adjustment. In addition, the present results mainly reflect the effect of force components, while the contribution of torque is not sufficiently isolated or quantified.

Future work will improve both data and control. Teleoperation can be used to collect demonstrations with explicit recovery behaviors, such as retracting, re-aligning, and retrying after edge contact or jamming. Vision-Force ACT can also be combined with a residual closed-loop module, where the learned policy predicts nominal actions and a force-based controller generates small corrective motions during contact. In addition, the task distribution should be expanded by varying peg-hole clearance, hole shape, object pose, friction, and compliance.

VII. CONCLUSION

This paper presented Vision-Force ACT, a multimodal extension of ACT for bimanual peg-in-hole assembly. The method incorporates bilateral wrist F/T observations into the ACT latent inference and action prediction process while keeping the original action-chunking formulation. In the simulated bimanual assembly task, Vision-Force ACT improves the success rate from 40% to 68% with 50 demonstrations and from 60% to 78% with 200 demonstrations. The stage-wise results further indicate that the improvement mainly occurs in the approach and insertion stages, where contact information is more important than in free-space motion. The force analysis also shows smaller force fluctuations than the vision-only baseline, suggesting a more stable contact response.

The current study also has clear limitations. The policy mainly benefits from force information, while the contribution of torque remains insufficiently evaluated. Moreover, the learned behavior is not yet an adaptive force-control process, since the demonstrations do not contain enough corrective actions after contact failures. Future work will focus on collecting teleoperation demonstrations with explicit recovery behaviors, introducing force-based residual or impedance/admittance correction, and evaluating the method under different peg-hole clearances, shapes, and physical parameters.

REFERENCES

- [1] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Proceedings of Robotics: Science and Systems*, 2023.
- [2] Y. Lin, A. Church, M. Yang, H. Li, J. Lloyd, D. Zhang, and N. F. Lepora, "Bi-touch: Bimanual tactile manipulation with sim-to-real deep reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 8, no. 9, pp. 5472–5479, 2023.
- [3] Y. Chen, M. V. der Merwe, A. Sipos, and N. Fazeli, "Visuo-tactile transformers for manipulation," in *Proceedings of the 6th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 205, 2023.
- [4] B. Huang, Y. Wang, X. Yang, Y. Luo, and Y. Li, "3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing," in *Proceedings of the 8th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 270, 2025, pp. 2557–2578.
- [5] T. Yang, Y. Jing, H. Wu, J. Xu, K. Sima, G. Chen, Q. Sima, and T. Kong, "Moma-force: Visual-force imitation for real-world mobile manipulation," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 6847–6852.
- [6] W. Liu, J. Wang, Y. Wang, W. Wang, and C. Lu, "Forcemimic: Force-centric imitation learning with force-motion capture system for contact-rich manipulation," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 1105–1112.
- [7] Z. He, H. Fang, J. Chen, H.-S. Fang, and C. Lu, "Foar: Force-aware reactive policy for contact-rich robotic manipulation," *IEEE Robotics and Automation Letters*, vol. 10, no. 6, pp. 5625–5632, 2025.
- [8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [9] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, vol. 44, no. 10-11, pp. 1684–1704, 2025.
- [10] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems*, vol. 28, 2015.