Re-ACT on Sensing: Force-Aware Action Chunking with Responsive Temporal Ensemble for Robot Manipulation

Seungho Yeom*
2020145157
Mechanical Engineering Department
Yonsei University
Seoul, Korea
duatmdgh3@yonsei.ac.kr

Joseph Park*
2020145021
Mechanical Engineering Department
Yonsei University
Seoul, Korea
iamjoseph1129@yonsei.ac.kr

Seunghun Han
2019145084
Mechanical Engineering Department
Yonsei University
Seoul, Korea
hansh326@yonsei.ac.kr

***Abstract*— The ability to respond to perturbations and adapt movements is essential for effective task execution, as humans do through multimodal sensing. However, most current imitation learning methods primarily focus on generating precise and smooth motions using action chunking, which often overlook responsiveness in action modification. To address this issue, we introduce Responsive Temporal Ensemble for Action Chunking with Transformer (Re-ACT), which enables robots to respond quickly to sudden external force deviation while maintaining smooth motions. Re-ACT achieves this by dynamically adjusting the weighting of action sequences based on force feedback and ensembling them to derive the final action for the current timestep. Our experimental results demonstrate that Re-ACT outperforms prior baselines in a responsive manipulation task by prioritizing actions generated when external force deviations are detected, while maintaining the ability to successfully complete the manipulation task. Code is available at https://github.com/iamjoseph1/Re-ACT.**

Keywords—F/T Sensors, Action Chunking, Temporal Ensemble, Imitation Learning

I. Introduction

The capacity for autonomous robotic systems to operate effectively in dynamic and unpredictable environments hinges on their ability to handle changing situations proficiently. This becomes especially critical in cases where disturbances arise during ongoing actions, a situation in which humans excel. Upon encountering external stimuli, people seamlessly integrate multimodal sensory information to execute precise movements, react swiftly, and resume interrupted tasks [1]. Simultaneously, they can prioritize relevant cues, such as focusing on touch or sound while ignoring distractions [2, 3], thereby enhancing their ability to detect and respond effectively to unexpected events. This intrinsic capability for responsive action modification, underpinned by sophisticated sensory processing, is essential for robust task performance in real-world scenarios.

In the realm of robot manipulation, substantial progress has been achieved through imitation learning, particularly in generating smooth and accurate trajectories. Methods employing action chunking, for example, have demonstrated considerable success in replicating complex motion patterns from human demonstrations. [4] However, a prevalent limitation in many current approaches is their primary focus on kinematic precision and motion fluidity. They often neglect the crucial aspect of responsiveness to unmodeled environmental interactions or sudden external perturbations. This oversight can lead to less robust robot behaviors in dynamic settings where deviations from anticipated conditions necessitate immediate adaptation.

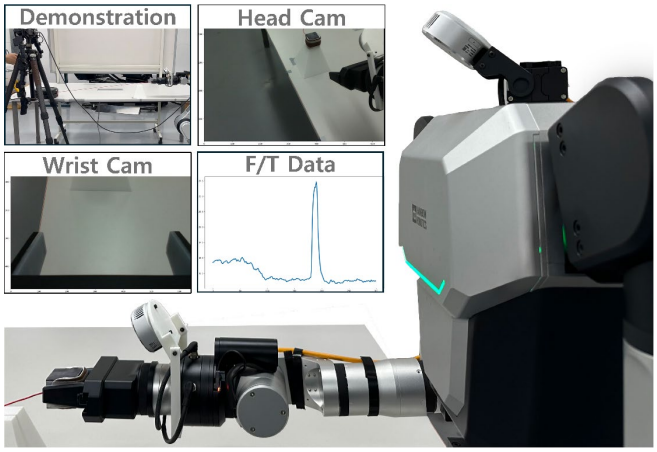

Fig. 1. Manipulation Task Setup with Multimodal Inputs.

To address this critical gap between precise motion generation and dynamic responsiveness, we introduce ***Re****sponsive Temporal Ensemble for* ***A****ction* ***C****hunking with* ***T****ransformer* (Re-ACT). Drawing inspiration from the human brain's capacity to selectively and dynamically respond to external signals, Re-ACT empowers robots to react quickly and appropriately to abrupt force deviations while maintaining the overall smoothness required for successful task completion. Our approach extends the foundational Action Chunking with Transformer (ACT) framework [4] by incorporating essential force/torque sensor data as an additional modality, alongside conventional visual and joint value inputs.

A core innovation within Re-ACT is the Responsive Temporal Ensemble (RTE) algorithm. While typical temporal ensembling techniques generally prioritize past predictions to achieve motion smoothing, which can inherently limit immediate responsiveness, RTE explicitly addresses this by dynamically adjusting the weighting of action sequences based on real-time force feedback. When no significant external stimulus is detected, RTE functions similarly to a conventional temporal ensemble, ensuring continuous, fluid motion. Conversely, upon the detection of a sudden external force, RTE dynamically increases the weight of the current timestep's predicted action, thereby facilitating a rapid and targeted response. To enhance the robustness of force signal processing, we further employ a Kalman filter for noise reduction and apply SoftMax weighting to emphasize the most relevant force signals.

*: Contributed Equally.

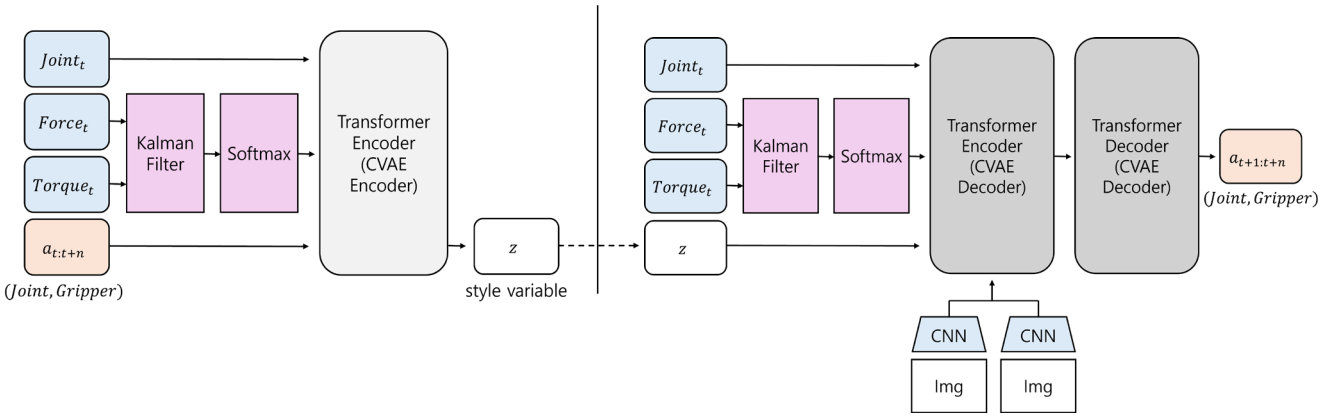

Fig. 2. Re-ACT's Network architecture. *Left:* The action sequence, consisting of n robot states (Joint target and gripper width), is encoded alongside the current joint values and the preprocessed Force/Torque signals with Kalman filter and SoftMax transformation by the CVAE encoder. This network is discarded at inference time. *Right:* The policy inputs are images from two viewpoints, the current joint values, and the measured Force/Torque. The policy predicts a n-sized action sequence.

We rigorously evaluated Re-ACT on a newly designed manipulation task setup in Fig. 1, where the robot interacts with an object under varying disturbance conditions by leveraging multimodal inputs. Our experimental results clearly demonstrate that Re-ACT surpasses prior baselines in its ability to respond to external force deviations, showing performance improvements of approximately 30-35%. Concurrently, Re-ACT maintains the necessary task completion rate, validating its dual objective of responsiveness and task efficacy. Our contributions are summarized as follows:

- Re-ACT: A novel multi-modal responsive model specifically engineered for robust robot manipulation in dynamic environments.
- Proposed Responsive Temporal Ensemble (RTE): A new algorithm that enhances responsiveness by dynamically adjusting action sequence weighting based on external stimuli.
- Applied SoftMax weighting to force/torque data to intelligently emphasize relevant sensor signals while suppressing noise and irrelevant activations, thereby improving performance.

II. Related Works

A. Incorporating F/T data to imitation learning

Learning frameworks that integrate force/torque (F/T) sensor feedback into imitation learning (IL) policies have demonstrated notable improvements in robot performance for manipulation tasks, especially in tasks that require contact-rich interactions. Comp-ACT integrates a compliance control mechanism based on the Action Chunking Transformer (ACT) [4], which learns to adjust compliance profiles among a set of discrete candidates while simultaneously tracking the trajectory derived from human demonstrations. This approach enables delicate manipulation by dynamically modulating compliance, resulting in more nuanced control [5]. Adaptive Compliance Policy (ACP), which is built upon the Diffusion Policy framework [6], learns continuous compliance profiles to reduce contact forces during task execution, thereby enhancing safety and adaptability [7]. ForceMimic adopts a hybrid force-position control strategy, dynamically switching between trajectory tracking and orthogonal force projection based on empirically determined force thresholds to effectively manage contact forces [8]. Although these approaches achieve higher task success rates compared to naive position-control methods that rely solely on visual input, they typically depend on dynamic impedance control architectures, which limits their applicability to robots that lack such impedance control hardware. To address this issue, a force-aware IL method that integrates tool–tissue interaction forces directly into the policy—without requiring an impedance control system—has been proposed [9]. However, this method is still limited to systems equipped with high-end force sensing, and its policy architecture largely follows the structure of prior methods [4], thereby constraining its generalizability and practical deployment across a wide range of robotic platforms.

B. Action chunking with temporal ensemble

Action chunking is a common method to avoid compounding errors in IL by predicting n-sized action sequences ("chunks") rather than single action for each timestep[1-4]. Due to this action chunking, there are n action candidates for each timestep and how to utilize these candidates to derive final single action that robot actually executes for each timestep is crucial for improving task performance. To address this issue, ACT [4] proposed Temporal Ensemble (TE) which ensembles candidates by using an weighted average over action candidates to ensure smooth transitions and robust control with following weighting schemes:

$$w_i = e^{-m*i}, \; m > 0 \tag{1}$$

Here, w_i is weight for the ith element of n-size action candidates for each timestep and m is a hyperparameter that controls the extent to which recently generated candidates are incorporated. As shown in Equation (1), TE always assigns more weight on candidates generated at earlier rollout stages for current timestep, so prioritizes them over the most recently generated candidate to derive final single action for the current timestep regardless of m. TE with upper weighting schemes has been shown to effectively achieve smooth transitions without any additional learnable parameters for the ensemble while avoiding compounding errors. Proleptic Temporal Ensemble (PTE) extended this idea to accelerate task execution by combining action sequences for future timesteps to return action for current timestep-preemptively predicting future actions and adjusting their execution timing [10]. This approach successfully enhanced execution speed for tasks. However, since it simply utilizes action sequences generated

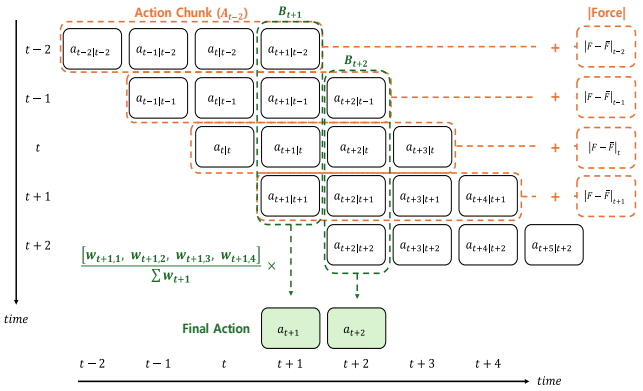

Fig. 3. Action Chunking with Responsive Temporal Ensemble (RTE). RTE combines action candidates ($\boldsymbol{B_{t+1}}$) with dynamically adjusted weights based on force feedback and produces the final action for each timestep.

for future timesteps from earlier rollout stages, it only expedites the execution timing of actions generated based on past observations under the assumption of no external disturbances. As a result, the policy cannot make decisions based on the most recent observations, making it difficult to adapt to variations that arise during task execution.

III. Methods

In this section, we introduce Re-ACT, which modifies ACT to sense and respond quickly to external force/torque (F/T) feedback. Re-ACT integrates F/T signals into the imitation learning network, allowing the robot to make decisions in response to external perturbations. The responsive temporal ensemble (RTE) further enhances the robot's reactivity by dynamically adjusting the action weighting based on sudden changes in force magnitude. Together, Re-ACT allows the robot to execute rapid and adaptive actions in tasks requiring fast reactions to external forces.

A. Imitation Learning Network

We built Re-ACT upon the ACT framework to learn position control in joint space. The learned policy, as the decoder of a conditional variational autoencoder (CVAE) [11], predicts a sequence of actions conditioned on the current observation, including RGB images, F/T sensor signals, and proprioception.

To mitigate the impact of sensor noise inherent in F/T signals, we integrated a Kalman filter that denoises raw F/T sensor data, ensuring more stable feature representations for imitation learning. Furthermore, to prevent overfitting to the magnitude of the F/T signals—especially given the variations in scale between expert demonstrations—we applied a SoftMax transformation to the denoised F/T signals. This SoftMax is applied independently to each 3-dimensional F/T vector per end effector, converting the raw signals into a normalized distribution. This transformation helps the policy focus on the relative differences between axes rather than their absolute magnitudes, which is crucial for learning generalizable and transferable policies during inference.

By incorporating the denoised and normalized F/T signals into the ACT, our method is able to leverage F/T signals as auxiliary feedback, resulting in improved performance on tasks based on perceived external forces. We illustrated full architecture of Re-ACT in Fig. 2.

Algorithm 1 Re-ACT Training

Given: Demo dataset D, chunk size k, weight β.
Let a_t, o_t represent action and observation at timestep t, $\bar{o}_t$ represent o_t without image observations.
Initialize encoder $q_\phi(z|\hat{a}_{t:t+k}, \bar{o}_t)$
Initialize encoder $\pi_\theta(\hat{a}_{t:t+k}|o_t, z)$
for iteration n = 1, 2, ... **do**
  Sample $o_t, a_{t:t+k}$ from D
  Sample z from $q_\phi(z|\hat{a}_{t:t+k}, \bar{o}_t)$
  Predict $\hat{a}_{t:t+k}$ with $\pi_\theta(\hat{a}_{t:t+k}|o_t, z)$
  $\mathcal{L}_{reconst} = MSE(\hat{a}_{t:t+k}, a_{t:t+k})$
  $\mathcal{L}_{reg} = D_{KL}(q_\phi(z|\hat{a}_{t:t+k}, \bar{o}_t) \,||\, \mathcal{N}(0, I)$
  Update θ, ϕ with ADAM and $\mathcal{L} = \mathcal{L}_{reconst} + \mathcal{L}_{reg}$

Algorithm 2 Re-ACT Inference

Given: trained π_θ, episode length T, weight m.
Initialize FIFO buffers $\mathcal{B}[0:T]$, where $\mathcal{B}[t]$, stores actions predicted *for* timestep t.
for timestep t = 1, 2, ...T **do**
  Predict $\hat{a}_{t:t+k}$ with $\pi_\theta(\hat{a}_{t:t+k}|o_t, z)$ where z = 0
  Add $\hat{a}_{t:t+k}$ to buffers $\mathcal{B}[t:t+k]$ respectively
  Obtain current step actions $A_t = \mathcal{B}[t]$
  Apply $a_t = \sum_i^k w_{t,i} B_t[i] \,/\, \sum_i^k w_{t,i}$, with $w_{t,i} = e^{(-m*i + \alpha * |F-\bar{F}|_{t-(i-k)})}$

B. Responsive Temporal Ensemble

To enhance 'responsiveness' in tasks that require quick reactions to sudden changes in external force, we propose the Responsive Temporal Ensemble (RTE) technique. RTE modifies the original TE mechanism in ACT by dynamically adjusting the weight assigned to each action candidate based on force feedback with following weighting schemes:

$$w_{t,i} = e^{-m*i+\alpha*|F-\bar{F}|_{t-(i-n)}},\ m > 0 \qquad (2)$$

Here, $|F - \bar{F}|_t$ is measured difference between force magnitude and a moving average of it at timestep t. We adopt a moving average to capture sudden changes in the measured force magnitude without any task-specific threshold. α is a hyperparameter that controls the extent to which real-time force feedback influences the weighting.

Unlike the fixed-weighted temporal ensemble in ACT that always prioritizes action candidates generated at earlier rollout stages, RTE assigns higher weights to action candidates produced at timesteps where the external force shows significant deviation. This approach effectively biases the policy towards more responsive and adaptive behaviors when sudden force changes occur, enabling the agent to better perform tasks requiring quick reactions to external force perturbations. As shown in Equation (2), RTE is based on the TE weighting scheme, so it introduces no learnable parameters. We illustrated process of Action Chunking with RTE in Fig. 3.

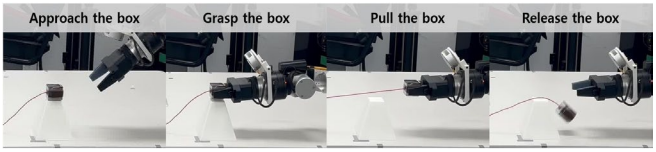

Fig. 4. Box Pick-and-Pull Procedure

IV. Experiment

A. Experimental Setup

We used the RB-Y1 mobile bi-manipulator with built-in F/T sensors in real-world experiments with a single-arm task. Three Realsense L515 cameras were attached to the top of

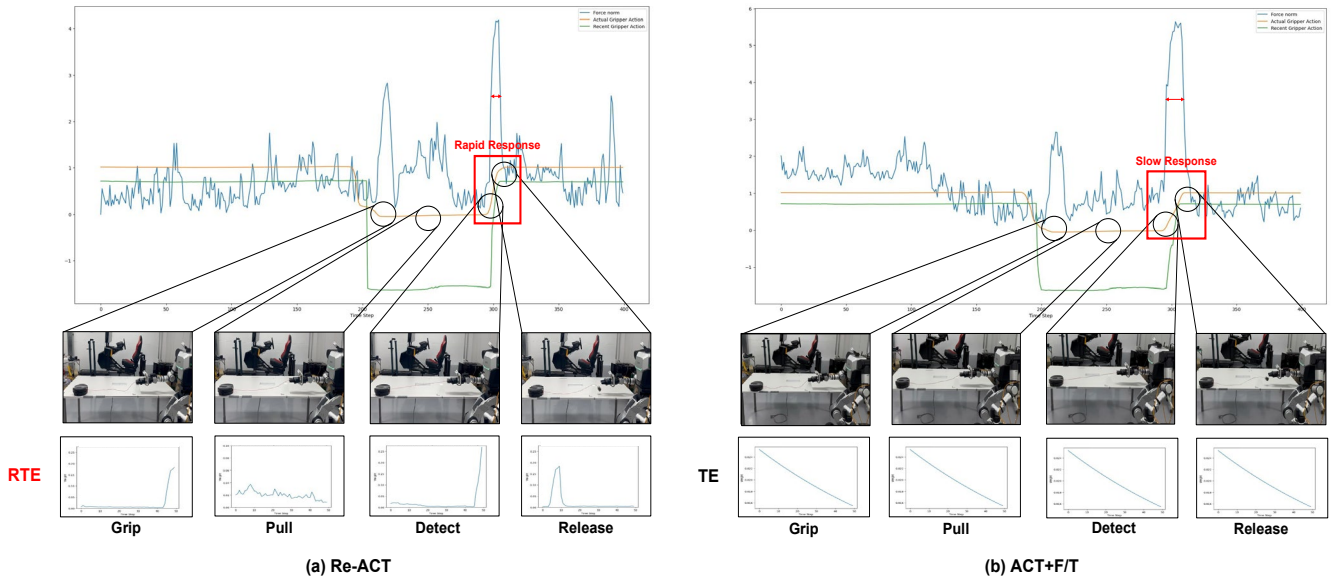

Fig. 5. Performance Analysis. *Top:* Force-Action Plot during rollout. Blue line is measured force magnitude, yellow line is actual gripper action, green line is most recently generated gripper action for each timestep. *Middle:* Scene image during rollout. *Bottom:* RTE / TE weights during rollout.

the torso and the wrist of each arm. For our single-arm task, only two cameras (one on the torso, one on the right wrist) were used for task execution.

B. Task

We designed a new task named 'Box Pick-and-Pull' to evaluate whether Re-ACT performs as intended. As shown in Fig. 4, the robot approaches a box tied to a heavy object with a rope and lifts and pulls the box. When it detects rope tension, it releases the box as soon as possible. The policy was trained from scratch for 30,000 epochs with 100 human demonstrations collected via teleoperation.

C. Results

We compared following three methods: 1) Re-ACT using RTE with 2) ACT using TE and 3) ACT with F/T sensor signals as observation (ACT+F/T) using TE. We illustrated each method's architecture in Fig. 6. Since the key point we want to evaluate with this task is the response speed of each method toward detected rope tension, we adopted the 'traveled distance of the heavy object' rather than success rate as the evaluation criterion to quantify the responsive-ness. We measured traveled distance over three different initial positions of the heavy object, which were designed to vary the moment (i.e., the timestep) at which the robot detects the rope tension during rollout. All methods were tested 10 times each.

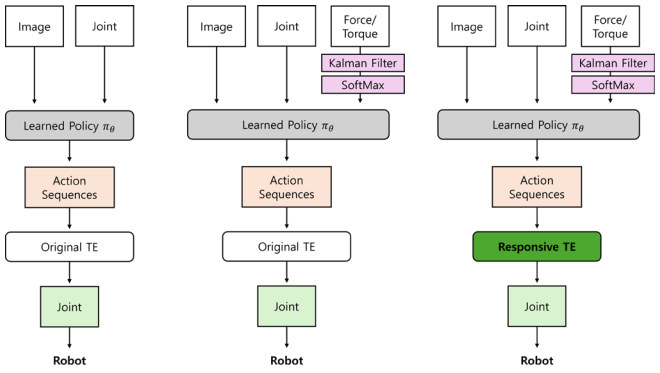

Fig. 6. Method Comparisons. *Left:* ACT with TE. *Middle:* ACT + F/T with TE. *Right:* Re-ACT with RTE

As shown in Fig. 7 and Table 1, All methods eventually released the box before the end of inference, but Re-ACT recorded the shortest traveled distance for all initial position. This result demonstrates Re-ACT showed the fastest response to detected rope tension and outperformed other methods.

We also illustrated both the actual gripper action executed by the robot and the action generated by the policy at the current timestep, as they change over time during the rollout, alongside the measured force magnitude, to analyze the performance of our method in a real-world setting in Fig. 5. Specifically, in Fig. 5(a), we show that Re-ACT exhibits a rapid change in the actual gripper action when a force peak occurs due to rope tension. This indicates that the RTE module effectively prioritizes the "Open the gripper" action—generated in response to the sudden detection of rope

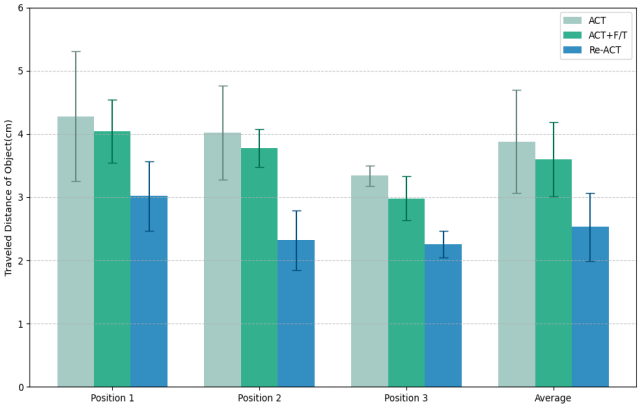

Fig. 7. Performance Comparison based on Traveled Distance. The average traveled distance (bar) and standard deviation (error bar) of 10 trials for each initial position and policy are shown.

TABLE I. PERFORMANCE COMPARISION BASED ON TRAVELED DISTANCE

Model	Obs. Type	TE	Pos. 1(cm)	Pos. 2(cm)	Pos. 3(cm)	Avg (cm)
ACT	Vision	TE	3.34	4.02	4.28	3.87
ACT+F/T	Vision+ F/T	TE	2.98	3.78	4.04	3.59
Re-ACT	**Vision+ F/T**	**RTE**	**2.26**	**2.32**	**3.02**	**2.53**

tension—over action candidates for the current timestep by assigning a higher weight to actions generated at the current timestep. This dynamic weighting enables a quick response to sudden force peak, allowing the robot to release the box almost immediately, which in turn leads to a rapid drop in the measured force magnitude. In contrast, as illustrated in Fig. 5(b), the ACT+F/T method which uses an original TE mechanism shows a much slower change in the actual gripper action following a force peak. This sluggish response is due to TE's fixed weight assignment that consistently prioritizes actions generated at earlier rollout stages rather than the current one. Consequently, the robot's releasing motion, which is in response to the force peak, is delayed, prolonging the rope tension and maintaining a high force magnitude for an extended period. These observations highlight the advantages of our method's fast response to sudden external force events compared to previous approaches, which can be critical for tasks involving delicate or dynamic physical interactions.

D. Ablations

Re-ACT employs SoftMax to focus on the relative size differences between axes of 3-dimensional force/torque signals, thus preventing overfitting to the absolute magnitude itself of each axis. Without SoftMax on F/T signals, Re-ACT could not consistently generate 'Open the gripper' action when rope tension is detected. We analyze this lack of reproducibility stems from the difference between the force signal values observed during training and those observed during inference for each individual axis. Each force/torque signal is 3-dimensional, so the measured value of a given axis can vary significantly between the training dataset and inference, even with small differences in the joint values of the robot. This leads to observations that are out-of-distribution, which the policy cannot handle robustly. However, SoftMax normalizes 3-dimensional signals, so policy receives relative size differences of the values measured on each axis. This helps policy deal with unseen force magnitudes during inference and leads to consistent responses to detected rope tension.

We also tested Re-ACT with different sets of α. When α is less than 1, robot's behavior is similar to that of ACT+F/T with TE, which means current observation is weakly reflected in the weight, and actions from earlier rollout stages are still prioritized. When α is greater than 1, RTE loses its smoothing ability due to strong reflection of the current observation in the weight, causing the robot to exhibit unstable and discontinuous behavior. This result shows the $\alpha = 1$ appropriately reflects the current observation in the weighting while maintaining the smoothing ability on action chunks.

V. Conclutsion

In this paper, we presented Re-ACT, a novel framework that bridges the gap between motion precision and dynamic responsiveness in robot manipulation, inspired by the human ability to adapt swiftly to external stimuli through multimodal sensing. Experimental results demonstrate that Re-ACT outperforms existing imitation learning baselines, achieving significantly enhanced responsiveness without compromising task success.

These findings underscore the importance of feedback-driven temporal adaptation in unpredictable environments and validate the role of multimodal fusion in enabling responsive robotic behavior. As future work, we aim to extend Re-ACT to a broader range of contact-rich manipulation tasks that require precise force application during dynamic motion, without relying on explicit force control. Promising candidates include drawing on a curved surface without damaging the object and wiping an unfixed whiteboard with a fragile erasing tool—tasks that demand delicate force modulation under changing contact conditions.

Acknowledgment

We thank the MLCS lab at Yonsei University for their support. We would also like to thank Su-Cheol Yoo for his helpful feedback throughout the entire process of this study as a domain mentor, Hyun-Jin Park and Jae-Beom Chae for their assistance with hardware setup and feedback on imitation learning.

References

[1] Trivedi H, Leonard JA, Ting LH, Stapley PJ. Postural responses to unexpected perturbations of balance during reaching. Exp Brain Res. 2010 Apr;202(2):485-91. doi: 10.1007/s00221-009-2135-4. Epub 2009 Dec 25. PMID: 20035321; PMCID: PMC4059204.

[2] Lavie N. Distracted and confused?: selective attention under load. Trends Cogn Sci. 2005 Feb;9(2):75-82. doi: 10.1016/j.tics.2004.12.004. PMID: 15668100.

[3] Iguchi Y, Hoshi Y, Tanosaki M, Taira M, Hashimoto I. Attention induces reciprocal activity in the human somatosensory cortex enhancing relevant- and suppressing irrelevant inputs from fingers. Clin Neurophysiol. 2005 May;116(5):1077-87. doi: 10.1016/j.clinph.2004.12.005. Epub 2005 Jan 22. PMID: 15826848.

[4] Zhao, T. Z., Kumar, V., Levine, S., & Finn, C. (2023). Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*.

[5] Kamijo, T., Beltran-Hernandez, C. C., & Hamaya, M. (2024, October). Learning variable compliance control from a few demonstrations for bimanual robot with haptic feedback teleoperation system. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 12663-12670). IEEE.

[6] Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., ... & Song, S. (2023). Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 02783649241273668.

[7] Hou, Yifan, et al. "Adaptive Compliance Policy: Learning Approximate Compliance for Diffusion Guided Control." *arXiv preprint arXiv:2410.09309* (2024).

[8] Liu, W., Wang, J., Wang, Y., Wang, W., & Lu, C. (2024). ForceMimic: Force-Centric Imitation Learning with Force-Motion Capture System for Contact-Rich Manipulation. *arXiv preprint arXiv:2410.07554*.

[9] Abdelaal, A. E., Fang, J., Reinhart, T. N., Mejia, J. A., Zhao, T. Z., Bohg, J., & Okamura, A. M. (2025). Force-Aware Autonomous Robotic Surgery. *arXiv preprint arXiv:2501.11742*.

[10] Park, H., Lim, D., Kim, S., & Park, S. (2024). Proleptic Temporal Ensemble for Improving the Speed of Robot Tasks Generated by Imitation Learning. *arXiv preprint arXiv:2410.16981*.

[11] Kim, J., Kong, J., & Son, J. (2021, July). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning* (pp. 5530-5540). PMLR.

Name	Contribution
Joseph Park	Implementation for all methods, Presentation preparation, Paper writing, and Experiments execution.
Seungho Yeom	Re-ACT model architecture design, Presentation preparation, Paper writing, and Experimental design and execution.
Seunghun Han	Experimental object design and Executing the experiments.