Proceedings of SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/2025.TJ22700

Dual-Stage Attention-Augmented ACT Model: Optimization and Validation for Bimanual Fine Manipulation Tasks

Shuhang Liang[1]*, Yuanlong Yue[1]

[1]Department of Automation, China University of Petroleum (Beijing), Beijing, China
**Corresponding Author. Email: 17303578185@163.com*

Abstract: Traditional Action Chunking with Transformers (ACT) exhibit limitations in multi-step precision manipulation tasks (e.g., battery insertion, nylon cable tie threading). Their single-stage attention mechanism struggles to capture cross-temporal dependencies, leading to significant deterioration of action coherence during critical phases in later task stages. Concurrently, the static fusion approach for multimodal features restricts adaptability in dynamic contact scenarios. To address these challenges, this paper proposes a dual-stage attention mechanism: A global self-attention layer establishes long-term dependencies in action sequences, while a conditional cross-attention layer dynamically filters critical contextual features (including latent variables and visual cues). These cascaded operations enable closed-loop optimization from global planning to local calibration. The hierarchical architecture integrates three functional modules: enhanced initial action sequence modeling at the encoder side, adaptive multimodal fusion through cross-module interfaces, and optimized end-effector control at the decoder side. Experimental results demonstrate that the improved model achieves 8% higher success rate in Cube Transfer tasks and 14% improvement in bimanual insertion tasks, with notable robustness in high-precision scenarios. Ablation studies confirm the necessity of hierarchical components: Removing encoder attention induces cumulative planning errors, disabling cross-module attention causes multimodal feature misalignment, and eliminating decoder attention exacerbates control instability. This research provides an extensible sequence modeling paradigm for complex manipulation tasks.

Keywords: Multi-step precision tasks, Dual-stage, Hierarchical architecture, Attention mechanisms

1. Introduction

Fine manipulation tasks such as battery insertion, cable threading, and precision assembly constitute the backbone of industrial automation and service robotics. While imitation learning has emerged as a powerful paradigm for acquiring manipulation skills from human demonstrations [1], existing methods still face persistent challenges in long-horizon task planning and dynamic multimodal perception during contact-rich operations. The Action Chunking with Transformers (ACT) approach [2], which addresses compounding errors through sequence prediction, reveals two critical limitations: First, its single-stage attention mechanism struggles to model long-term dependencies across action sequences. This deficiency becomes particularly pronounced in tasks requiring multi-phase coordination, such as bimanual nylon cable tie threading, where the initial grasping posture must align with the final insertion trajectory [3]. Second, the static fusion strategy for multimodal inputs (e.g., vision, force-torque) fails to adapt to environmental perturbations like sensor noise or occlusion, leading to cascading errors where millimeter-level deviations in early stages ultimately escalate into task failure [4].

Recent advancements in Transformer-based architectures, including RT-1 and BeT [5], have improved multi-task generalization but remain constrained by rigid attention patterns and fixed fusion strategies. For instance, ACT's homogeneous self-attention layers overly emphasize local temporal correlations at the expense of global task logic, while conventional multimodal fusion methods rely on simplistic concatenation [6] or static weighting. These limitations prove particularly detrimental in precision tasks like micro-assembly, where success hinges on both centimeter-level trajectory planning and real-time integration of force feedback.

To address these challenges, we propose a Hierarchical Dual-Stage Attention Mechanism for Complex Tasks (HiT-ACT), a novel architecture incorporating three key innovations: a task-conditioned dual-stage attention mechanism that models long-term dependencies through global self-attention while dynamically filtering critical features via conditional cross-attention; a hierarchical feature enhancement framework that iteratively refines action planning through encoder-decoder attention alignment; and latent variable-driven multimodal fusion, which extends the Conditional Variational Autoencoder (CVAE) framework [7] to dynamically adjust the weighting of visual and force-control signals.

Experimental validation in simulated and real-world tasks demonstrates that HiT-ACT significantly outperforms existing methods in both precision and robustness. In high-precision manipulation scenarios, our approach effectively mitigates error propagation through hierarchical attention, while dynamic multimodal fusion enables adaptive responses to environmental variations [8]. Ablation studies further confirm the necessity of each architectural component. Detailed quantitative comparisons and task-specific analyses are presented in Section IV.

2. Related works

2.1. Action chunking and sequence modeling

Action chunking techniques address long-horizon task planning by decomposing continuous operations into executable sub-sequences. Early studies employed sequence modeling methods based on Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks [9] to maintain temporal coherence in action sequences. For instance, end-to-end policy optimization methods optimized motion planning through temporal difference learning. However, due to gradient vanishing issues in long-term sequence modeling, such approaches struggle to capture distant dependencies effectively. The Action Chunking with Transformers (ACT) model [2] first introduced the Transformer architecture into action chunking, leveraging self-attention mechanisms to capture local temporal correlations and achieving notable progress in tabletop manipulation tasks. Nevertheless, its experimental validation was limited to single-arm short-cycle tasks (e.g., block pushing, door opening) and did not address bimanual coordination or precision contact scenarios. Recent extensions, such as Temporal Ensemble methods, improved long-term consistency by expanding historical action windows, but their fixed-length attention mechanisms still face computational efficiency bottlenecks in high-dimensional operational spaces.

2.2. Multimodal feature fusion

Multimodal perception plays a critical role in robotic fine manipulation tasks, and researchers have proposed various fusion methods to enhance environmental understanding and task adaptability. Traditional approaches often adopt static fusion strategies. For example, Levine et al. [10]

implemented visual-force signal concatenation via end-to-end policy learning to optimize robotic control. However, such methods fail to adapt to sensor noise or dynamic environmental changes. To improve adaptive fusion, Lee et al. [11] proposed attention-based modality weighting adjustment, enabling dynamic allocation of importance between visual and tactile signals. Yet, this method still suffers from response latency under real-time contact variations. Recent advances, such as self-supervised learning frameworks proposed by Gao et al. [12], employ latent variable modeling to align multimodal representations, significantly enhancing fusion robustness. Notably, the Conditional Variational Autoencoder (CVAE) framework demonstrates unique advantages in latent space fusion. Zhao et al. achieved visual-tactile representation alignment through variational inference, but their decoding process lacks task-specific conditional constraints. Additionally, recent studies like the Multimodal Transformer [13] utilize cross-modal attention to improve feature interaction, yet their high computational complexity limits applicability in high-frequency control tasks. In contrast, our work adopts a task-conditioned dynamic fusion strategy combined with hierarchical attention mechanisms, achieving a better balance between computational efficiency and environmental adaptability.

2.3. Attention-based manipulation policies

The application of Transformer architectures in robotic control has diversified rapidly. RT-1 achieves multi-task generalization through large-scale pretraining, but its dense attention mechanism results in quadratic computational cost growth with task complexity. Behavior Transformer (BeT) [5] enhances trajectory prediction accuracy via bidirectional attention but tends to generate physically infeasible solutions in tasks involving contact force constraints. The Dual-Attention method [14] decouples spatial and temporal attention for assembly tasks but fails to address feature propagation decay in multi-stage tasks. Recent studies like the Hierarchical Transformer employ a two-level architecture to separately handle macro task decomposition and micro-action generation, yet their rigid inter-layer interfaces struggle to adapt to dynamic environmental perturbations. In contrast, our proposed dual-stage attention mechanism achieves superior balance between computational efficiency and task adaptability through global-local closed-loop optimization.

3. Method

3.1. Action chunking transformer framework overview

The original ACT framework leverages transformers to learn human behaviors by utilizing their sequential data processing capabilities and long-range dependency modeling for imitation learning. Its core components include a visual encoder, action sequence encoder, and conditional decoder. The action sequence encoder encodes demonstrated action sequences and joint positions into a style variable z, capturing the distribution patterns of actions. The visual encoder extracts static visual features, while the conditional decoder synthesizes the style variable, static visual features, and joint positions to predict future action sequences. However, its single-stage attention mechanism struggles to model cross-timestep dependencies and dynamically adjust attention weights based on task conditions, which may reduce action coherence and limit performance in complex tasks.

3.2. Dual-stage attention mechanism

To address these limitations, HiT-ACT introduces a dual-stage attention mechanism, as illustrated in Figure 2, comprising two stages: a global planning layer and a local calibration layer.

Global Self-Attention Layer: This sequence modeling module captures long-term dependencies between input elements. By computing attention weights between every pair of elements in the sequence, it dynamically generates global contextual features. Formally, given an input sequence X ∈RL×D(where L is sequence length and D is the embedding dimension), the multi-head self-attention mechanism produces the output:

$$GlobalAttn(X) = Softmax\left(\frac{(W_Q X)(W_K X)^T}{\sqrt{D}}\right) W_V X \tag{1}$$

Where $W_Q, W_K, W_V \in R^{D\times D}$ are learnable projection matrices. The global attention layer effectively associates distant elements (e.g., initial and terminal states), ensuring current decisions account for future behaviors. This prevents the encoder from being dominated by local features and neglecting long-term dependencies. In HiT-ACT, this layer is deployed in the initial stages of both the encoder and decoder to encode global task logic.

Conditional Cross-Attention Layer: A dynamic feature filtering module that adaptively weights the input sequence X using external conditional signals C:

$$GondAttn(X, C) = Softmax(\frac{(W_C C)(W_K X)^T}{\sqrt{D}}) W_V X \tag{2}$$

where $W_C \in R^{Dc\times D}$ is a conditional projection matrix. This layer dynamically adjusts element-wise weights in X based on task conditions C , enabling closed-loop optimization from global planning to local calibration.

By hierarchically combining global self-attention and conditional cross-attention layers, HiT-ACT enhances adaptability to complex manipulation tasks. The dual-stage mechanism is deployed at multiple positions, forming a multi-level feature refinement system.

3.3. HiT-ACT hierarchical architecture design

HiT-ACT's hierarchical architecture, shown in Figure 1, achieves closed-loop optimization from global planning to local calibration through synergistic encoder, cross-module interface, and decoder designs.

Encoder Global Planning Module: The encoder input includes action sequences, joint positions, and a CLS token. Zero-initialized latent variables serve as task conditions, aligned via linear projection layers. Global dependency modeling is performed through dual-stage attention processing. During early training stages, zero initialization avoids prior bias in latent variables, ensuring the encoder learns critical features directly from input sequences. This effectively associates sequence elements and enriches global context for subsequent encoding.

Cross-Module Interface: Acts as a bridge between encoder and decoder, facilitating information transfer and interaction. It processes encoder outputs and latent variables generated from encoder outputs, incorporating dual-stage attention layers. Unlike traditional ACT models that merely pass latent variables, HiT-ACT's interface fuses multi-level features and dynamically adjusts encoder output weights based on latent dependencies, improving generalization across tasks.

Decoder Local Calibration Module: Refines preliminary decoding results from the transformer module. A dynamic weight generator produces time-varying weights based on the mean of decoder outputs, reflecting their importance at each timestep. These weights adjust the significance of latent variables, providing multi-level information to the decoder. The weighted latent variables and decoder outputs are fed into dual-stage attention layers. Since latent variables encapsulate action distribution patterns, they help the model contextualize current actions within the broader distribution, optimizing predictions to align with real-world action sequences.

Through enhanced encoder modeling, adaptive cross-module fusion, and decoder optimization, HiT-ACT achieves closed-loop optimization from global planning to local calibration, addressing challenges in complex manipulation tasks while improving performance and robustness.

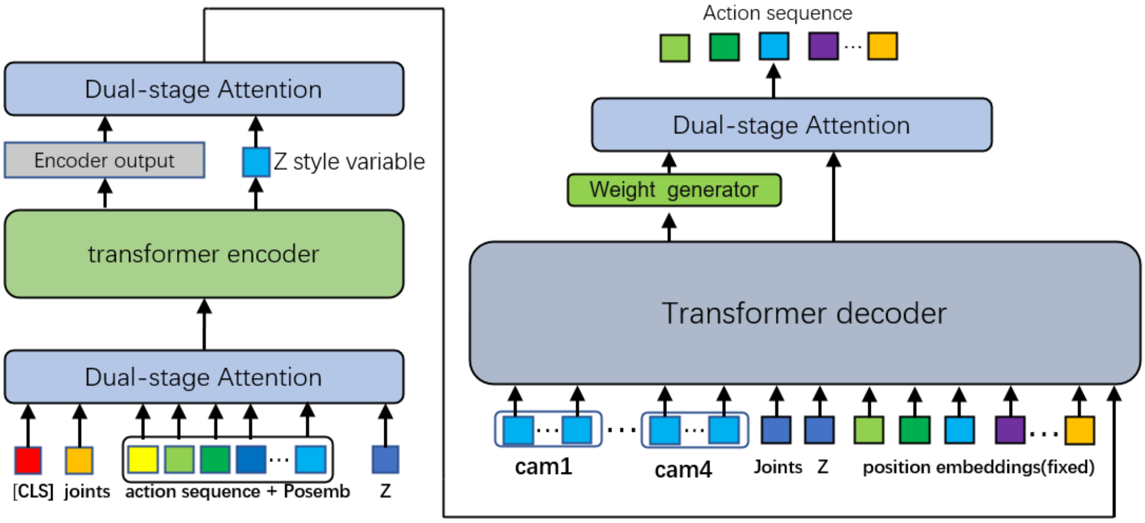

Figure 1: HiT-ACT splitting architecture

4. Experiments

4.1. Experimental setup

To validate the effectiveness of our proposed HiT-ACT model in bimanual fine manipulation tasks, we constructed two challenging simulation tasks in the MuJoCo environment: (1) Transfer Cube and (2) Bimanual Insertion. Both tasks involve dual-arm coordination and require high precision to evaluate the model's adaptability and stability in complex operations.

The simulation environment provides precise state feedback while incorporating real-world uncertainties. The initial positions of objects are randomized within a defined range to enhance the model's adaptability to varying conditions. Additionally, uniform random perturbations are applied during experiments to test robustness against environmental changes.

4.1.1. Transfer cube

In this task, the right arm must pick up a red cube from a tabletop and precisely place it into the gripper of the left arm. The clearance between the cube and the left gripper is approximately 1 cm, where minor errors may cause collisions or failure. This task evaluates fine motion planning and dual-arm coordination. We measure the success rates of different methods across subtasks (picking, transferring, placing) to analyze robustness.

4.1.2. Bimanual insertion

Here, the left and right arms grasp a socket and plug, respectively, and align them mid-air to complete the insertion. The clearance between the plug and socket is 5 mm, requiring high-precision alignment and force control to avoid failure or component damage. This task evaluates HiT-ACT's execution capability in precision alignment scenarios. Small perturbations are applied to the initial positions of the socket and plug to simulate real-world randomness and test adaptability to environmental variations.

4.2. Baseline comparison

To systematically analyze the core differences between HiT-ACT and prior imitation learning methods, we compare its architectural features with five baseline approaches. BC-ConvMLP, a widely used baseline, employs a convolutional neural network to process current image observations and directly concatenates visual features with joint position information for action prediction. BeT (Behavior Transformer), another Transformer-based architecture, differs from ACT in several key aspects: it does not utilize action chunking and instead predicts single-step actions directly from historical observations. Notably, BeT's visual encoder is pretrained independently from the control network, resulting in unoptimized integration between perception and control modules. Furthermore, BeT discretizes the action space by predicting a categorical distribution over discrete bins and generates final actions by superimposing continuous offsets. RT-1, also based on the Transformer architecture, predicts single-step actions using a fixed-length historical observation window, with a similarly discretized action space and output format. In contrast, VINN adopts a non-parametric approach that requires real-time access to the demonstration dataset during testing: it extracts visual features from new observations using a pretrained ResNet, retrieves the k-nearest historical observation samples, and generates actions via a weighted k-nearest neighbors algorithm. While VINN's visual feature extractor is based on a pretrained ResNet, it undergoes unsupervised fine-tuning on demonstration data. Finally, the original ACT model, serving as the foundational framework for this study, employs a single-stage attention mechanism for action prediction, providing a critical reference for our design. Quantitative comparisons of success rates across these methods are presented in Table 1.

4.3. Ablation studies

To investigate the contributions of different components in HiT-ACT, we conduct three ablation experiments by removing encoder attention, cross-module attention, and decoder attention, respectively, and record their impact on task success rates (Table 2). Analysis reveals that removing encoder attention prevents the model from fully extracting global features, leading to degraded long-term action planning and execution deviations. Removing cross-module attention reduces the fusion capability between latent variables and visual information, destabilizing multimodal information transmission and impairing adaptability to dynamic environments. Removing decoder attention compromises the final optimization of action sequences, weakening error correction in fine operations and resulting in unstable control during the late stages of tasks. Overall, the ablation results validate the importance of each module in HiT-ACT. Specifically, the three attention mechanisms collectively form a hierarchical information processing framework, enabling efficient modeling of long-term dependencies, optimized multimodal fusion, and enhanced control precision in complex tasks. These results further demonstrate that our method exhibits strong adaptability and robustness in fine manipulation tasks, offering new directions for future research in robotic intelligent control.

Table 1: Comparison of HiT-ACT with baseline algorithms

Method	Cube Transfer	Bimanual Insertion
BC-ConvMLP	1%	1%
BeT	27%	3%
RT-1	2%	1%
VINN	3%	1%
ACT	86%	32%
HiT-ACT(Ours)	94%	46%

Table 2: Ablation studies

Method	Cube Transfer	Bimanual Insertion
HiT-ACT	94%	46%
Ours w/o Encoder	91%	40%
Ours w/o Cross-Module	90%	33%
Ours w/o Decoder	88%	35%

5. Conclusions and future work

This study improves upon the original ACT framework by addressing its limitations in long-horizon tasks and proposes HiT-ACT, an optimized architecture based on a dual-stage attention mechanism. By integrating global planning with local calibration, we enhance the model's ability to model action sequences, enabling precise capture of long-term dependencies and dynamic adjustment under varying task conditions. Additionally, we introduce a cross-module information interaction mechanism to strengthen feature fusion capabilities, improving task coherence and generalization. Experimental results demonstrate that HiT-ACT achieves superior performance across multiple key metrics, not only increasing task success rates but also optimizing computational resource utilization, laying a solid foundation for broader applications.

While this work achieves significant advancements, several directions warrant further exploration. Future efforts will focus on refining the algorithm's architecture to enhance adaptability and robustness in complex environments. For instance, we aim to improve the conditional cross-attention layer to enable more efficient input sequence weighting and strengthen the model's responsiveness to diverse task conditions. We also plan to optimize the dynamic weight generation mechanism, allowing the decoder to better contextualize the role of current actions within the broader task, thereby improving the rationality and stability of outputs.

Furthermore, we will explore richer multimodal fusion strategies to enhance performance across diverse scenarios. For example, integrating visual, tactile, and auditory signals could equip the model with comprehensive environmental perception, improving its applicability in robotic interaction tasks. Additionally, we intend to incorporate self-supervised and transfer learning methods to boost training efficiency, enabling rapid adaptation to new tasks with limited data while reducing training costs and deployment overhead.

In practical applications, we envision deploying HiT-ACT to high-dynamic environments such as smart manufacturing, autonomous driving, and robotic manipulation. For resource-constrained settings, we will further optimize computational efficiency to ensure stable performance with minimal overhead. Integrating reinforcement learning techniques could empower the model to autonomously adapt to dynamic environments, enhancing long-term decision-making capabilities. To improve interpretability, we will investigate visualization methods to transparently illustrate decision-making processes, ensuring trustworthiness in real-world deployments.

In summary, this study not only successfully refines the ACT algorithm but also advances robotic imitation learning. As the technology evolves, we anticipate HiT-ACT will demonstrate exceptional performance in increasingly complex scenarios, offering novel insights and practical pathways for continued progress in related fields.

References

[1] Hussein A, Gaber M M, Elyan E, et al. Imitation learning: A survey of learning methods[J]. ACM Computing Surveys (CSUR), 2017, 50(2): 1-35.

[2] Zhao T Z, Kumar V, Levine S, et al. Learning fine-grained bimanual manipulation with low-cost hardware[J]. arXiv preprint arXiv:2304.13705, 2023.

[3] Mazzia V, Angarano S, Salvetti F, et al. Action transformer: A self-attention model for short-time pose-based human action recognition[J]. Pattern Recognition, 2022, 124: 108487.

[4] Wang C, Fan L, Sun J, et al. Mimicplay: Long-horizon imitation learning by watching human play[J]. arXiv preprint arXiv:2302.12422, 2023.

[5] Brohan A, Brown N, Carbajal J, et al. Rt-1: Robotics transformer for real-world control at scale[J]. arXiv preprint arXiv:2212.06817, 2022.

[6] Tan J, Tang J, Wang L, et al. Relaxed transformer decoders for direct action proposal generation[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 13526-13535.

[7] Kingma D P, Welling M. Auto-encoding variational bayes[EB/OL].(2013-12-20)

[8] Levine S, Finn C, Darrell T, et al. End-to-end training of deep visuomotor policies[J]. Journal of Machine Learning Research, 2016, 17(39): 1-40.

[9] Cheng J, Dong L, Lapata M. Long short-term memory-networks for machine reading[J]. arXiv preprint arXiv:1601.06733, 2016.

[10] Levine S , Pastor P , Krizhevsky A ,et al.Learning Hand-Eye Coordination for Robotic Grasping with Deep Learning and Large-Scale Data Collection[J].International Journal of Robotics Research, 2016, 37(4-5):421-436.DOI:10.1007/978-3-319-50115-4_16.

[11] LEE M A, ZHU Y, SRINIVASAN K, et al. Multimodal sensor fusion with differentiable attention for robotic manipulation[C]//IEEE International Conference on Robotics and Automation (ICRA). Montreal, Canada, 2019: 5595-5601.

[12] GAO J, LI Y, TRIVEDI S S. Self-supervised multimodal fusion for robotic grasping[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 141-148.

[13] Kroll A , Ranjan S , Lercher M J .A multimodal Transformer Network for protein-small molecule interactions enhances predictions of kinase inhibition and enzyme-substrate relationships[J].PLoS Computational Biology, 2024, 20(5).DOI:10.1371/journal.pcbi.1012100.

[14] UNCAPHER M R, WAGNER A D. Posterior parietal cortex and episodic encoding: Insights from fMRI subsequent memory effects and dual-attention theory[J]. Neurobiology of Learning and Memory, 2009, 91(2): 139-154. DOI: 10.1016/j.nlm.2008.10.011.