

InterACT: Inter-dependency Aware Action Chunking with Hierarchical Attention Transformers for Bimanual Manipulation

Anonymous Author(s)

Affiliation

Address

email

Abstract: We present InterACT: Inter-dependency aware Action Chunking with Hierarchical Attention Transformers, a novel imitation learning framework for bimanual manipulation that integrates hierarchical attention to capture inter-dependencies between dual-arm joint states and visual inputs. InterACT consists of a Hierarchical Attention Encoder and a Multi-arm Decoder, both designed to enhance information aggregation and coordination. The encoder processes multimodal inputs through segment-wise and cross-segment attention mechanisms, while the decoder leverages synchronization blocks to refine individual action predictions, providing the counterpart’s prediction as context. Our experiments on a variety of simulated and real-world bimanual manipulation tasks demonstrate that InterACT significantly outperforms existing methods. Detailed ablation studies validate the contributions of key components of our work, including the impact of CLS tokens, cross-segment encoders, and synchronization blocks.

Keywords: Imitation Learning, Bimanual Manipulation

1 Introduction

Bimanual manipulation tasks, such as unscrewing a bottle cap or connecting two electrical cables, presents significant challenges due to the need for high precision coordination. Traditional approaches often rely on high-end robots and precise sensors, which can be expensive and require meticulous calibration [1, 2, 3]. However, recent advances in learning-based approaches offer the potential to perform such complex tasks using low-cost hardware.

The ALOHA and the Action Chunking with Transformers (ACT) framework has shown that low-cost systems can achieve high-precision tasks that were traditionally only possible with expensive setups. The ACT addresses the compounding error problem in imitation learning by predicting sequences of actions rather than single steps, thereby reducing the task’s effective horizon and mitigating errors over time [4].

Despite these recent advances, it remains a challenge for bimanual robotics to ensure robust and accurate coordination between two arms in a dynamic environment. In this work, we propose InterACT: a new policy for bimanual manipulation that emphasizes inter-dependencies between two arms by utilizing hierarchical attention mechanisms. In our designs, multimodal inputs are encoded through segment-wise and cross-segment encoders which handle the complex relationships between different segments in a manner similar to how long documents are processed in NLP [5]. This combines the proprioceptive data of the robot arm joints and the visual features of the camera in a coherent latent space that allows for coordinated detail-oriented and smooth action execution.

Our method goes beyond existing approaches by focusing on the hierarchical nature of bimanual tasks, ensuring each arm operates independently while coordinating seamlessly with its counterpart

36 and the environment. Building on the latest advances in attention mechanisms, this work presents a
37 compelling solution and features improved performance in bimanual manipulation tasks.

38 **Contributions:**

- 39 1. **Hierarchical Attention Mechanism for Bimanual Manipulation:** We adapt a novel
40 hierarchical attention mechanism that effectively captures inter-dependencies between
41 dual-arm joint states and visual inputs. This mechanism processes multi-modal inputs
42 through segment-wise encoding followed by cross-segment encoding, which captures inter-
43 dependencies and allows for precise and coordinated bimanual manipulation.
- 44 2. **Comprehensive Evaluation:** We conduct extensive experiments on multiple simulated
45 and real-world implemented tasks. The results highlight the effectiveness of our approach
46 compared to previous methods.
- 47 3. **Ablation Studies:** We perform several ablation studies to assess contributions from dif-
48 ferent components of our framework, namely the classification (CLS) tokens, the cross-
49 segment encoder, and the synchronization block. These studies provide insight into the
50 importance of each element to realize bimanual manipulation that is robust and efficient.

51 **2 Related Works**

52 **2.1 Bimanual Manipulation**

53 Bimanual manipulation involves two robotic arms performing tasks that require dexterity and syn-
54 chronization. Inspired by the natural ability of humans to perform such tasks, researchers have
55 been keen on modeling these skills in robots. Prevailing methodologies, including classical control
56 methods, reinforcement learning, and imitation learning, have significantly advanced the field.

57 Early research primarily relied on classical control methods, focusing on predefined trajectories and
58 high-fidelity models to achieve coordinated movements [1, 2, 3]. However, these methods often
59 required extensive calibration and were less adaptable to dynamic environments.

60 The introduction of reinforcement learning (RL) made bimanual manipulation more adaptable and
61 robust. RL-based approaches have proven effective in handling complex tasks, outperforming clas-
62 sical methods, and generalizing across different scenarios [6, 7, 8, 9]. These methods leverage RL’s
63 ability to learn from interactions with the environment, improving performance in varied and unpre-
64 dictable conditions.

65 Imitation learning has emerged as a prominent method for teaching robots bimanual tasks through
66 human demonstrations. This approach enables robots to mimic complex human actions, facilitating
67 the execution of intricate tasks. Research has demonstrated the effectiveness of imitation learning
68 in training robots for coordinated dual-arm manipulation using hierarchical skill learning and force-
69 based techniques [10, 11, 12, 13, 14].

70 Recent advancements have focused on integrating machine learning models to enhance imitation
71 learning. Frameworks like ALOHA have shown that low-cost systems can perform high-precision
72 tasks using imitation learning techniques traditionally reserved for expensive setups [4, 15]. Sim-
73 ilarly, the Action Chunking with Transformers (ACT) algorithm addresses the compounding error
74 problem in imitation learning by predicting sequences of actions, improving accuracy and efficiency
75 [4].

76 Despite these advancements, bimanual manipulation remains challenging, especially in dynamic and
77 unstructured environments. Recent research has explored integrating additional multi-modal data,
78 such as language or sensory feedback, to enhance the robustness and efficiency of bimanual manip-
79 ulation systems [16, 17]. These efforts hold promise for developing more adaptable and efficient
80 robotic systems capable of performing a wide range of complex manipulation tasks.

81 2.2 Hierarchical Attention Mechanisms

82 Hierarchical attention [18] mechanisms have gained prominence for their ability to process and inte-
83 grate multi-modal inputs. These mechanisms aggregate information at multiple levels of granularity,
84 making them well-suited for tasks requiring both local and global context understanding [19, 20].

85 Hierarchical attention has shown considerable success in other domains, such as natural language
86 processing (NLP), where hierarchical attention transformers have been effectively used for long
87 document classification [5, 21, 22]. These models focus on different parts of the input text at varying
88 levels of abstraction, enhancing the ability to handle long and complex documents. Extensions of
89 foundational attention mechanisms [23], such as the Hierarchical Attention Network (HAN) [18] and
90 hierarchical representations in BERT [24], further demonstrate their potential in managing multi-
91 layered information.

92 In robotic manipulation, the concept of hierarchy has been explored to manage the complexity of
93 long-horizon tasks by breaking down problems into manageable sub-tasks [25]. As proven in long
94 document classification, leveraging segment-wise and cross-segment attention mechanisms, hier-
95 archical attention models can capture dependencies within and across segments [5]. This makes
96 hierarchical attention transformers particularly appealing for use in bimanual robotic manipulation
97 tasks, where precise coordination between arms is crucial.

98 In this research, we explore a different utility of hierarchical attention in which inter/intra-segment
99 attention supports a more robust extraction of inter-dependencies between the actions of the two
100 arms in handling a single task, leading to more coordinated actions in complex bimanual manipula-
101 tion.

102 3 InterACT: Inter-dependency Aware Action Chunking with Hierarchical 103 Attention Transformer

104 InterACT builds upon the Action Chunking with Transformers (ACT) framework [4], enhancing it
105 with hierarchical attention mechanisms to capture inter-dependencies between dual-arm joint states
106 and visual inputs. This section details the components and workflow of the InterACT framework,
107 focusing on the encoder and decoder structures within an imitation learning framework.

108 The ACT framework leverages transformer architecture to predict future steps in bimanual manip-
109 ulation tasks, effectively handling sequences of actions by capturing temporal dependencies. How-
110 ever, it does not explicitly model inter-dependencies between dual-arm joint states and visual inputs,
111 which can limit its performance in complex manipulation tasks.

112 InterACT extends the ACT framework by introducing hierarchical attention transformers, consisting
113 of two main components: **Hierarchical Attention Encoder** and **Multi-arm Decoder**.

114 3.1 Hierarchical Attention Encoder

115 Segments are defined as individual groups of data inputs that are processed independently before
116 being integrated. In this context, segments include the joint states of each arm and visual features
117 at a specific timestep. Hierarchical Attention Encoder processes input segments and captures both
118 intra-segment and inter-segment dependencies through a hierarchical attention mechanism, which
119 consists of segment-wise encoder and cross-segment encoder layers. A detailed pipeline for the
120 Hierarchical Attention Encoder is illustrated in Figure 1 and Algorithm 1.

121 **Input:** Each joint in each arm is embedded into a single token through a linear layer. Visual features
122 are extracted from RGB images using ResNet18 backbones, which convert and flatten the images
123 along the spatial dimension to form a sequence of feature tokens. The joint embeddings from both
124 arms and the visual embeddings of image frames from multiple cameras are concatenated to form a
125 combined sequence. Multiple classification (CLS) tokens are prepended to each segment, allowing
126 the model to summarize segment information during attention processing. Positional embeddings

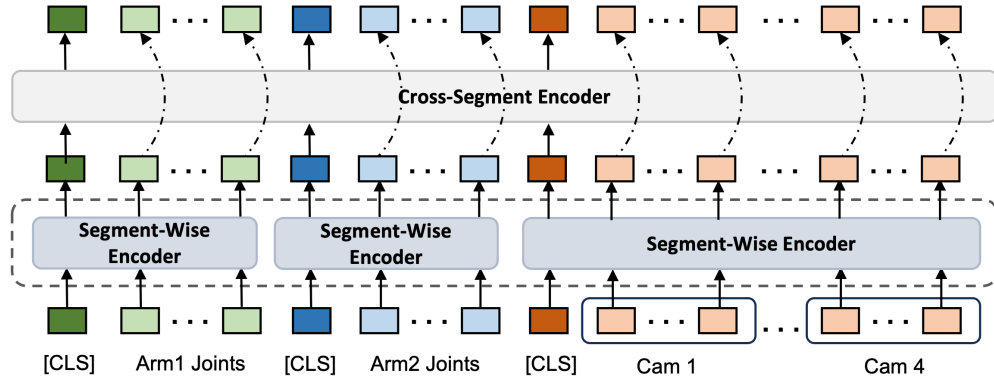


Figure 1: **Architecture of the Hierarchical Attention Encoder.** The encoder consists of segment-wise encoders that independently process each input segments (joint states of each individual arm and visual features at the current timestep). This is followed by a cross-segment encoder that integrates information across segments (across robot arms).

are then added to the tokens to retain sequence information, ensuring the model can understand the order of the tokens within each segment. The input for each decoder includes relevant segments and CLS tokens to facilitate cross-attention mechanisms.

Segment-wise Encoder: Each segment passes through multiple layers of multihead self-attention to aggregate information within each segment. The self-attention mechanism allows each token within the segment to attend to every other token, including the CLS tokens, thereby capturing intra-segment dependencies. This aggregation of information enables the model to effectively leverage the relationships within different parts of the segment [5], such as the connections between joints within the same arm and the relationships within the features from visual inputs. Importantly, the segment-wise encoder uses shared weights across all segments in each layer, ensuring uniform transformation and processing.

Cross-segment Encoder: The Cross-segment Encoder handles CLS tokens from each segment to capture inter-segment relationships. This component allows the model to discern and combine information across different segments [5] (e.g., joint states of both arms and visual features). The cross-segment encoder employs shared weights across all segments to maintain consistent modeling of these relationships.

Output: The original visual feature (input to the encoder, excluding CLS tokens) replaces the processed visual features to retain detailed spatial-temporal information.

3.2 Multi-arm Decoder

Similar to multi-task learning in NLP [26, 27], the Multi-arm Decoder comprises two parallel paths of decoder blocks, each dedicated to processing the encoded states and target tokens to generate the predicted actions for one of the arms. This section details the components and workflow of the Multi-arm Decoder, highlighting how it employs enriched embeddings from the encoder to coordinate actions between both arms effectively. A detailed pipeline for the Multi-arm Decoder is illustrated in Figure 2 and Algorithm 2.

Input: Encoded states from the Hierarchical Attention Encoder are used as the context for decoding. Target tokens, initialized with positional embeddings, serve as the starting point for generating actions. The input for each decoder includes relevant segments and CLS tokens to facilitate cross-attention mechanisms.

Algorithm 1: Hierarchical Attention Encoder

```
1: Given: Demo dataset  $\mathcal{D}$ , segment-wise encoder  $E_{\text{seg}}$ , cross-segment encoder  $E_{\text{cross}}$ , CLS tokens CLS,
   number of layers  $L$ .
2: Let  $S_i$  represent the input segment at index  $i$ ,  $\text{CLS}_i$  represent the CLS tokens for segment  $i$ , and  $S_{\text{visual}}$ 
   represent the visual features.
3: Initialize segment-wise encoder  $E_{\text{seg}}$ 
4: Initialize cross-segment encoder  $E_{\text{cross}}$ 
5: for each segment  $S_i$  in  $\mathcal{D}$  do
6:   Prepend CLS tokens:  $S'_i \leftarrow [\text{CLS}_i, S_i]$ 
7:   Add positional encoding to  $S'_i$ 
8:   for each layer  $l = 1, 2, \dots, L$  do
9:      $S'_i \leftarrow E_{\text{seg}}(S'_i)$ 
10:  end for
11: end for
12: Extract CLS tokens:  $\text{CLS}_{\text{tokens}} \leftarrow [\text{CLS}_1, \text{CLS}_2, \dots, \text{CLS}_N]$ 
13: Add positional encoding to  $\text{CLS}_{\text{tokens}}$ 
14: for each layer  $l = 1, 2, \dots, L$  do
15:    $\text{CLS}_{\text{tokens}} \leftarrow E_{\text{cross}}(\text{CLS}_{\text{tokens}})$ 
16: end for
17: Replace processed visual features  $S'_{\text{visual}}$  with original visual features:  $S'_{\text{visual}} \leftarrow S_{\text{visual}}$ 
18: Return final encoded states  $\{\text{CLS}_{\text{arm1}}, S'_{\text{arm1}}, \text{CLS}_{\text{arm2}}, S'_{\text{arm2}}, \text{CLS}_{\text{visual}}, S'_{\text{visual}}\}$ 
```

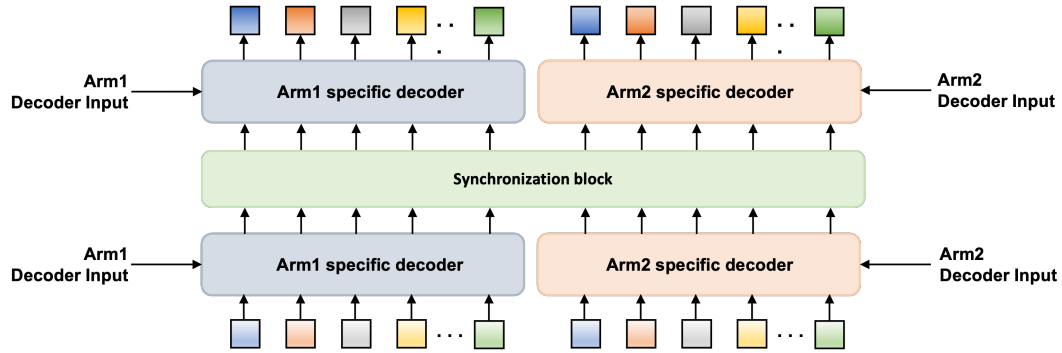


Figure 2: **Architecture of the Multi-arm Decoder.** The decoder consists of Arm1 and Arm2 specific decoders that process the input segments independently. The synchronization block allows for information sharing between the two decoders, ensuring coordinated and synchronized predictions for both arms. This mechanism is critical for achieving effective bimanual manipulation by integrating contextual information from both arms.

156 **Arm Specific Decoders:** The Arm Specific Decoders are responsible for generating intermediate
157 decoder outputs for each arm. The Arm1 Specific Decoder takes in Arm1 segments, Arm2 CLS
158 tokens, and visual information, while the Arm2 Specific Decoder takes in Arm2 segments, Arm1
159 CLS tokens, and visual information. Each layer processes these inputs through a cross-attention
160 mechanism with the target tokens, incorporating contextual information from both joint states and
161 visual features. This design enables each arm to use contextual information from the other arm
162 ensuring synchronized actions.

163 **Synchronization Block:** To enhance coordination between both arms, before the final layer of each
164 decoders, the intermediate outputs from the two decoders are concatenated and processed through
165 a synchronization block, which uses self-attention to integrate the shared information. This step
166 ensures that both arms leverage the combined context from the other arm and visual inputs before
167 the final decoder block. The use of attention mechanisms for sharing information across multiple
168 decoders has also been explored in multi-task learning, demonstrating its effectiveness in improving
169 performance and coherence within tasks [28]. The output of the synchronization block is then input
170 to the arm specific decoders for the final outputs.

Algorithm 2: Multi-arm Decoder

```
1: Given: Target tokens with positional embeddings  $T_{\text{pos}}$ , number of layers  $L$ .
2: Let  $D_{\text{arm1}}$ ,  $D_{\text{arm2}}$  represent decoders for Arm1 and Arm2 respectively.
3: Let  $D_{\text{sync}}$  represent Synchronization block.
4: Initialize cross-attention blocks for  $D_{\text{arm1}}$  and  $D_{\text{arm2}}$ 
5: Initialize synchronization block (multihead self-attention)
6: Arm1 Specific Decoder:
7: for each layer  $l = 1, 2, \dots, L$  do
8:    $\text{Arm1}_{\text{input}} \leftarrow \{S_{\text{arm1}}, \text{CLS}_{\text{arm2}}, \text{CLS}_{\text{visual}}, S_{\text{visual}}\}$ 
9:    $\text{Arm1}_{\text{output}} \leftarrow D_{\text{arm1}}(\text{Arm1}_{\text{input}}, T_{\text{pos}})$ 
10: end for
11: Arm2 Specific Decoder:
12: for each layer  $l = 1, 2, \dots, L$  do
13:    $\text{Arm2}_{\text{input}} \leftarrow \{\text{CLS}_{\text{arm1}}, S_{\text{arm2}}, \text{CLS}_{\text{visual}}, S_{\text{visual}}\}$ 
14:    $\text{Arm2}_{\text{output}} \leftarrow D_{\text{arm2}}(\text{Arm2}_{\text{input}}, T_{\text{pos}})$ 
15: end for
16: Synchronization Block:
17:  $\text{Concatenated}_{\text{output}} \leftarrow \{\text{Arm1}_{\text{output}}, \text{Arm2}_{\text{output}}\}$ 
18:  $\text{Shared}_{\text{output}} \leftarrow D_{\text{sync}}(\text{Concatenated}_{\text{output}})$ 
19: Split Shared Output:
20:  $\text{Shared}_{\text{output\_arm1}}, \text{Shared}_{\text{output\_arm2}} \leftarrow \text{Split}(\text{Shared}_{\text{output}})$ 
21: Arm Specific Decoder
22:  $\text{Arm1}_{\text{final\_output}} \leftarrow D_{\text{arm1}}(\text{Arm1}_{\text{input}}, \text{Shared}_{\text{output\_arm1}})$ 
23:  $\text{Arm2}_{\text{final\_output}} \leftarrow D_{\text{arm2}}(\text{Arm2}_{\text{input}}, \text{Shared}_{\text{output\_arm2}})$ 
24: Return  $\text{Arm1}_{\text{final\_output}}, \text{Arm2}_{\text{final\_output}}$ 
```

3.3 Training and Evaluation

InterACT is trained using an end-to-end imitation learning framework adapted from the ACT algorithm. The training process involves collecting high-quality human demonstrations through a teleoperation system, capturing joint positions and RGB images at 50Hz. The collected data is pre-processed to extract joint states and visual features using ResNet18 backbones, converting the RGB images into feature tokens. Each joint state and visual feature is tokenized, with multiple CLS tokens prepended to summarize each segment’s information. Positional embeddings are added to retain sequence information. Action chunking is implemented to predict sequences of actions rather than single steps, reducing the task’s effective horizon and mitigating compounding errors. Additionally, a temporal ensemble method is employed to improve the temporal consistency and robustness of the action predictions by weighing predictions over multiple time steps [4].

Evaluation is conducted on both simulated and real-world tasks, measuring success rates to assess the model’s performance in generating accurate and coordinated actions.

4 Experiments and Results

For the real-robot setup, we modified the ALOHA 2 [29] setup by adjusting the height of the top camera to improve the visibility of the tabletop environment. This adjustment ensures that the camera captures a more comprehensive view of the workspace, which is essential for accurately tracking the bimanual manipulation tasks. Our robot setup is illustrated in Appendix C.

To evaluate our model, we conducted experiments on three simulation tasks and six real-world tasks: **Transfer Cube** and **Peg Insertion** along with **Slide Ziploc** and **Thread Velcro** are tasks adapted from ACT [4]. We introduce five new tasks: one simulation task and four real-world tasks. The simulation task is **Slot Insertion**, where both arms need to grab each side of a long peg together and place it in a slot on the table. The new real-world tasks include **Insert Plug**, **Click Pen**, **Sweep**, and **Unscrew Cap**. Detailed task definitions are provided in Appendix A.

We focus our comparisons exclusively on ACT as ACT already outperforms BC-ConvMLP [30, 31], BeT [32], and VINN [33] by a large margin in bimanual manipulation tasks [4].

	Transfer Cube (Sim)			Peg Insertion (Sim)			Slide Ziploc (Real)			Thread Velcro (Real)		
	Touch	Lift	Transfer	Grasp	Contact	Insert	Grasp	Pinch	Open	Lift	Grasp	Insert
ACT	82	60	50	76	66	20	96	92	88	88	42	16
InterACT (Ours)	98	88	82	88	78	44	96	92	92	94	56	20

Table 1: Success rate (%) for tasks adapted from ACT, comparing ACT and InterACT. Coordination subtasks are indicated in bold.

	Slot insertion (Sim)		Insert Plug (Real)		Click Pen (Real)		Sweep (Real)		Unscrew cap (Real)	
	Lift	Insert	Grasp	Insert	Grasp	Click	Grasp	Sweep	Touch	Unscrew
ACT	96	88	92	30	92	56	88	42	84	60
InterACT (Ours)	100	100	92	42	94	62	92	52	88	62

Table 2: Success rate (%) for our original simulation and real-world tasks, comparing ACT and InterACT. Coordination subtasks are indicated in bold.

4.1 Results

The results of our experiments are summarized in Tables 1 and 2. Our InterACT model shows superior performance compared to ACT on all simulated and real-world tasks. In the simulated tasks, InterACT outperformed ACT significantly, particularly in the *Transfer Cube* and *Peg Insertion* tasks where coordination and precision are crucial. The success rates for the "Transfer" stages in the *Transfer Cube* task, as well as the "Insert" stages in the *Peg Insertion* task, were notably higher with InterACT, demonstrating the effectiveness of our method in tasks that require coordination between the two arms. Moreover, in newly introduced tasks such as *Slot Insertion* and *Insert Plug*, InterACT also demonstrated higher performance. The *Slot Insertion* task, which requires precise coordination between two arms to carry the peg and adjust for alignment, showed a 100% success rate with InterACT, compared to 88% with ACT. Similarly, in the *Insert Plug* task, InterACT achieved better results in the coordination subtask "Insert". This highlights the robustness of our model in handling tasks that require precise coordination between the two arms.

Overall, the experimental results validate the effectiveness of our hierarchical attention framework. By improving coordination and precision in bimanual manipulation tasks, our InterACT model provides a more robust solution for complex bimanual manipulation challenges in both simulation and real-world scenarios.

4.2 Ablation Studies

In this section, we perform ablation studies to evaluate the contributions of different components of the InterACT framework. Specifically, we focus on the impact of CLS tokens, the cross-segment encoder, and the synchronization block in the decoder. Detailed results of these ablation studies are provided in Appendix B.

Impact of CLS Tokens: To assess the impact of CLS tokens, we conducted experiments with and without CLS tokens as input to the decoder. The results, summarized in Table 3, showed no significant difference in the easier *Transfer Cube* task. However, there were notable improvements in the success rates of the more complex *Peg Insertion* task when CLS tokens were included. The aggregated information in the CLS tokens enhances the model's ability to generate accurate and coordinated actions, particularly in tasks requiring higher precision and synchronization.

Impact of Cross-segment Encoder: The cross-segment encoder captures inter-segment dependencies, allowing the model to effectively integrate information from different joints across the two arms as well as the camera frames. The results indicate that removing the cross-segment encoder significantly decreases performance in complex tasks. For example, in the *Slot Insertion* task, the success rate for the coordination subtask "Insert" dropped to 24% from 44% without the cross-segment

	Transfer Cube			Peg Insertion			Slot Insertion	
	Touch	Lift	Transfer	Grasp	Contact	Insert	Lift	Insert
InterACT (no CLS Tokens)	98	84	84	70	68	22	100	86
InterACT (no CS Encoder)	80	72	72	84	80	24	100	98
InterACT (no Sync Block)	74	54	54	90	86	30	100	100
InterACT (Ours)	98	88	84	88	78	44	100	100

Table 3: Success rate (%) for simulation tasks under different conditions, comparing ACT and InterACT model with InterACT model without CLS tokens, cross-segment (CS) encoder, and synchronization block. Coordination subtasks are indicated in bold.

230 encoder. This highlights the importance of capturing inter-segment dependencies for generating
231 accurate and coordinated actions.

232 **Impact of Synchronization Block:** The synchronization block enhances coordination between the
233 two arms by sharing contextual information during decoding. This is crucial for synchronized and
234 efficient bimanual manipulation, especially for predicting a sequence of actions. The results show
235 that the removal of the synchronization block leads to a significant drop in performance across all
236 tasks, particularly in the *Transfer Cube* and *Peg Insertion* tasks. This demonstrates the necessity of
237 the synchronization block for achieving coordinated and synchronized actions.

238 The ablation studies clearly illustrate that all components of the proposed InterACT frame-
239 work—CLS tokens, cross-segment encoder, and synchronization block—play a critical role in
240 achieving high success rates in coordination tasks. The highest performance for the final coordina-
241 tion task is achieved when all components are utilized, underscoring the importance of the holistic
242 integration of these elements in the InterACT framework.

243 5 Conclusion and Future Work

244 In this paper, we presented InterACT, a framework for robust bimanual manipulation, which in-
245 tegrates hierarchical attention transformers to capture inter-dependencies between dual-arm joint
246 states and visual inputs. The key contributions of our work include the development of a Hierar-
247 chical Attention Encoder and a Multi-arm Decoder. The Hierarchical Attention Encoder aggregates
248 intra-segment information using a segment-wise encoder and integrates inter-segment dependencies
249 through a cross-segment encoder. The Multi-arm Decoder ensures coordinated action sequence gen-
250 eration for each arm through synchronization blocks. Our experimental results, obtained from both
251 simulated and real-world tasks, demonstrate the superior performance of InterACT compared to the
252 baseline ACT framework. The use of CLS tokens, cross-segment encoder and the synchronizatin
253 block significantly enhances the model’s ability to generate accurate and coordinated actions, lead-
254 ing to higher success rates in bimanual manipulation tasks. The ablation studies further highlight
255 the importance of these components in our framework.

256 While InterACT has shown promising directions in integrating robotic arm joints and visual inputs,
257 it has not yet explored integration with other modalities. Future work will explore integrating ad-
258 ditional multi-modal data, such as text or sensory feedback, to further improve the robustness and
259 efficiency of bimanual manipulation tasks. Additionally, addressing the hierarchical nature of tasks
260 will be crucial for better task decomposition and execution. These enhancements could leverage the
261 flexible attention mechanisms demonstrated in this work to manage the added complexity and data
262 integration.

References

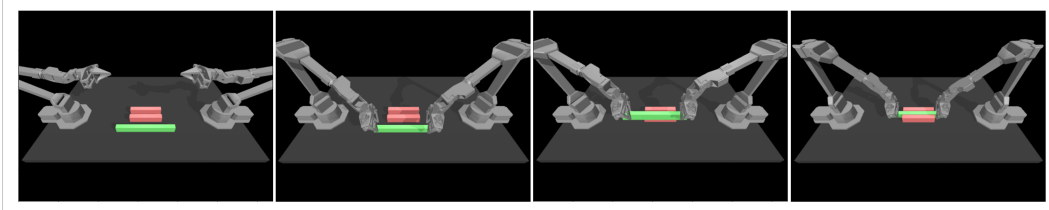
- [1] S. S. Mirrazavi Salehian, N. Figueroa, and A. Billard. A unified framework for coordinated multi-arm motion planning. *The International Journal of Robotics Research*, 37(10):1205–1232, 2018.
- [2] Y. Koga and J.-C. Latombe. Experiments in dual-arm manipulation planning. In *Proceedings 1992 IEEE International Conference on Robotics and Automation*, pages 2238–2239. IEEE Computer Society, 1992.
- [3] C. Smith, Y. Karayiannidis, L. Nalpantidis, X. Gratal, P. Qi, D. V. Dimarogonas, and D. Kragic. Dual arm manipulation—a survey. *Robotics and Autonomous systems*, 60(10):1340–1353, 2012.
- [4] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *RSS*, 2023.
- [5] I. Chalkidis, X. Dai, M. Fergadiotis, P. Malakasiotis, and D. Elliott. An exploration of hierarchical attention transformers for efficient long document classification. *arXiv preprint arXiv:2210.05529*, 2022.
- [6] Y. Chen, T. Wu, S. Wang, X. Feng, J. Jiang, Z. Lu, S. McAleer, H. Dong, S.-C. Zhu, and Y. Yang. Towards human-level bimanual dexterous manipulation with reinforcement learning. *Advances in Neural Information Processing Systems*, 35:5150–5163, 2022.
- [7] S. Kataoka, S. K. S. Ghasemipour, D. Freeman, and I. Mordatch. Bi-manual manipulation and attachment via sim-to-real reinforcement learning. *arXiv preprint arXiv:2203.08277*, 2022.
- [8] R. Chitnis, S. Tulsiani, S. Gupta, and A. Gupta. Efficient bimanual manipulation using learned task schemas. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1149–1155. IEEE, 2020.
- [9] K. S. Luck and H. B. Amor. Extracting bimanual synergies with reinforcement learning. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4805–4812. IEEE, 2017.
- [10] O. Kroemer, C. Daniel, G. Neumann, H. Van Hoof, and J. Peters. Towards learning hierarchical skills for multi-phase manipulation tasks. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 1503–1510. IEEE, 2015.
- [11] A. X. Lee, H. Lu, A. Gupta, S. Levine, and P. Abbeel. Learning force-based manipulation of deformable objects from multiple demonstrations. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 177–184. IEEE, 2015.
- [12] T. Asfour, P. Azad, F. Gyarfas, and R. Dillmann. Imitation learning of dual-arm manipulation tasks in humanoid robots. *International journal of humanoid robotics*, 5(02):183–202, 2008.
- [13] J. Silvério, L. Rozo, S. Calinon, and D. G. Caldwell. Learning bimanual end-effector poses from demonstrations using task-parameterized dynamical systems. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 464–470. IEEE, 2015.
- [14] A.-L. Pais Ureche and A. Billard. Learning bimanual coordinated tasks from human demonstrations. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, pages 141–142, 2015.
- [15] Z. Fu, T. Z. Zhao, and C. Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. In *arXiv*, 2024.

- [16] L. X. Shi, Z. Hu, T. Z. Zhao, A. Sharma, K. Pertsch, J. Luo, S. Levine, and C. Finn. Yell at your robot: Improving on-the-fly from language corrections. *arXiv preprint arXiv:2403.12910*, 2024.
- [17] N. Becker, E. Gattung, K. Hansel, T. Schneider, Y. Zhu, Y. Hasegawa, and J. Peters. Integrating visuo-tactile sensing with haptic feedback for teleoperated robot manipulation. *arXiv preprint arXiv:2404.19585*, 2024.
- [18] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.
- [19] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic. Multimodal affective analysis using hierarchical attention strategy with word-level alignment. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 2225. NIH Public Access, 2018.
- [20] C. Wu, Y. Wei, X. Chu, S. Weichen, F. Su, and L. Wang. Hierarchical attention-based multimodal fusion for video captioning. *Neurocomputing*, 315:362–370, 2018.
- [21] I. Chalkidis, I. Androutsopoulos, and N. Aletras. Neural legal judgment prediction in english. *arXiv preprint arXiv:1906.02059*, 2019.
- [22] C. Wu, F. Wu, T. Qi, and Y. Huang. Hi-transformer: Hierarchical interactive transformer for efficient and effective long document modeling. *arXiv preprint arXiv:2106.01040*, 2021.
- [23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [25] J. Luo, C. Xu, X. Geng, G. Feng, K. Fang, L. Tan, S. Schaal, and S. Levine. Multi-stage cable routing through hierarchical imitation learning. *IEEE Transactions on Robotics*, 2024.
- [26] H. Ivison and M. E. Peters. Hyperdecoders: Instance-specific decoders for multi-task nlp. *arXiv preprint arXiv:2203.08304*, 2022.
- [27] X. Xu, H. Zhao, V. Vineet, S.-N. Lim, and A. Torralba. Mtformer: Multi-task learning via transformer and cross-task reasoning. In *European Conference on Computer Vision*, pages 304–321. Springer, 2022.
- [28] I. Lopes, T.-H. Vu, and R. de Charette. Cross-task attention mechanism for dense multi-task learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2329–2338, 2023.
- [29] ALOHA 2 Team. Aloha 2: An enhanced low-cost hardware for bimanual teleoperation, 2024. URL <https://aloha-2.github.io/>.
- [30] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, and P. Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 5628–5635. IEEE, 2018.
- [31] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, pages 991–1002. PMLR, 2022.

- 348 [32] N. M. Shafiullah, Z. Cui, A. A. Altanzaya, and L. Pinto. Behavior transformers: Cloning k
349 modes with one stone. *Advances in neural information processing systems*, 35:22955–22968,
350 2022.
- 351 [33] J. Pari, N. M. Shafiullah, S. P. Arunachalam, and L. Pinto. The surprising effectiveness of
352 representation learning for visual imitation. *arXiv preprint arXiv:2112.01511*, 2021.

353 Appendix A: New task definitions

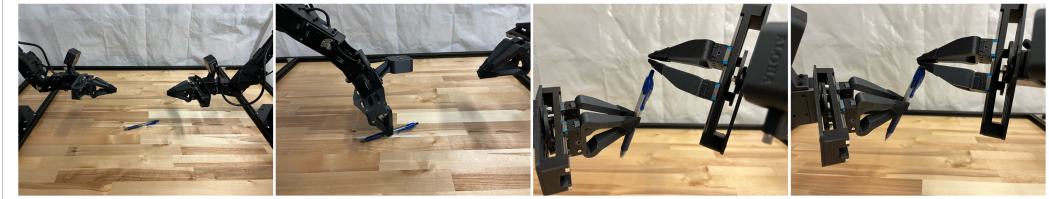
354 In this section, we define the new tasks we introduced in this work including one simulated task and
 355 four real-world tasks.



Slot Insertion (Sim): Slot insertion is a simulated task where both arms need to lift each side of a long peg together (*Lift*) and place it in a slot on the table (*Insert*).



Insert Plug (Real): Insert Plug is a real-world task where each arm grabs a male and a female electrical plug respectively (*Grasp*) and connects the two plugs above the table (*Insert*).



Click Pen (Real): Click Pen is a real-world task where each one (left) arm grabs a retractable pen in the middle (*Grasp*), and clicks the pen with the other (right) arm (*Click*).



Sweep (Real): Sweep is a real-world task where one arm grabs a brush and the other arm grabs a dustpan (*Grasp*). The arms then move towards a toy object lying on the table and sweep it into the dustpan (*Sweep*).



Unscrew Cap (Real): Unscrew Cap is a real-world task where one arm grabs a plastic water bottle while the other arm reaches and touches the bottle cap (*Touch*), then grabs the bottle cap and unscrews it (*Unscrew*).

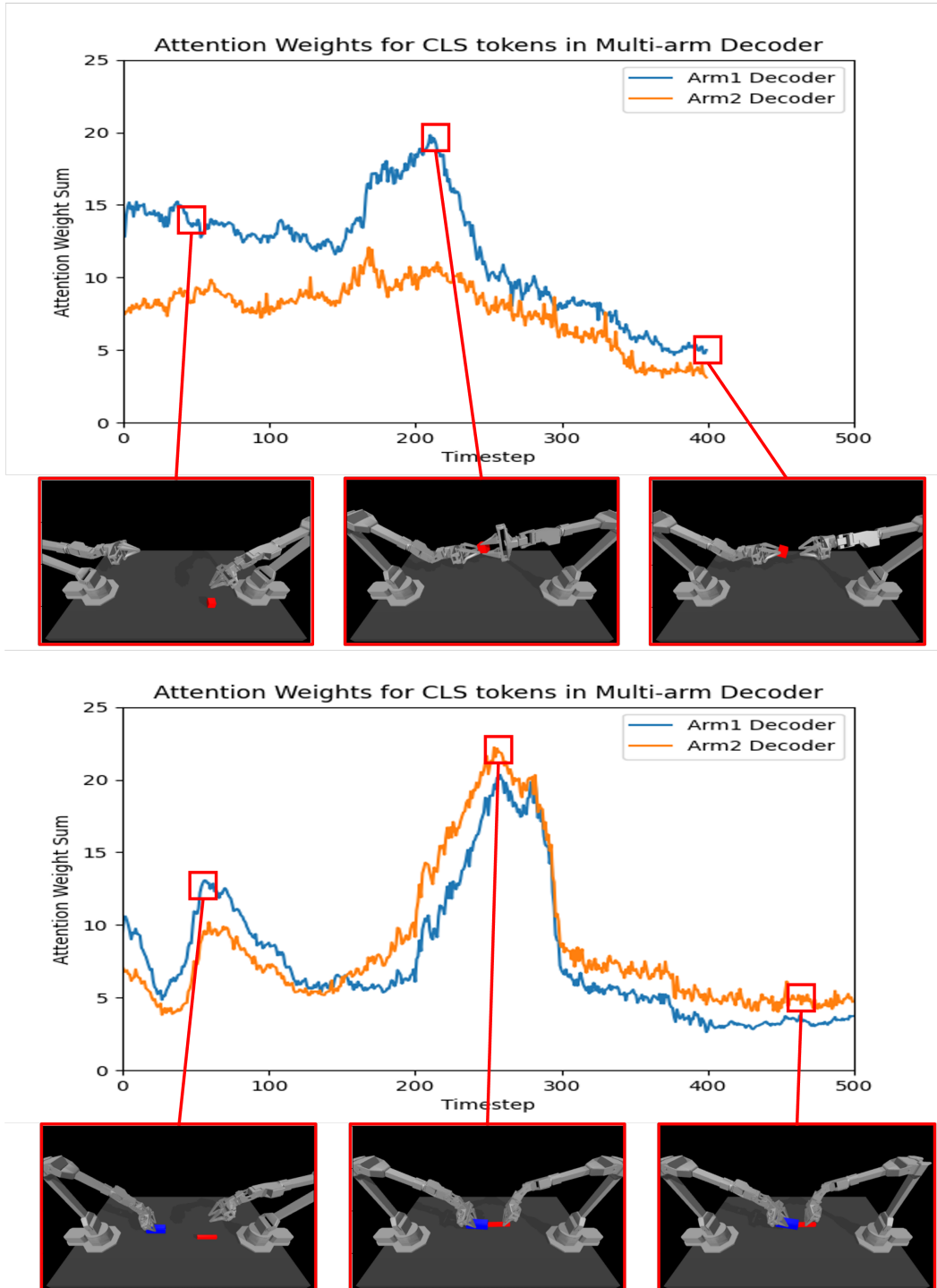


Figure 3: Attention weights for CLS tokens in the Multi-arm Decoder over time for *Transfer Cube* (top) and *Peg Insertion* (bottom). The blue line represents the sum of attention weights for the CLS tokens in the Arm1 Decoder, and the orange line represents the same for the Arm2 Decoder. The red highlighted sections correspond to specific timesteps in executing the task. Spikes in attention weights are observed during coordinated phase.

357 To gain deeper insights into the model’s behavior, we studied how attention weights to CLS tokens
358 change over the timesteps. This study is essential to understand the dynamics of information usage
359 and how the model prioritizes different parts of the input sequence during the decoder phase. Higher
360 attention weights for the CLS tokens indicate a stronger focus on the aggregated information they
361 represent.

362 The results illustrated in Figure 3 showed that during the phase of interaction between the two arms,
363 significant spikes in attention weights were observed. These spikes occurred at key moments where
364 coordinated actions between the arms were necessary. This indicates that the model heavily relies
365 on the CLS tokens to aggregate and process crucial information when coordinating actions between
366 the arms, highlighting their importance in facilitating precise bimanual manipulation.

367 **Appendix C: Real-robot Setup**

368 We utilize the ALOHA 2 setup [29] for our real-world experiments. Rather than cropping the top
369 camera frame, we lower the camera’s height to focus on the table environment, thereby maintaining
370 resolution and capturing the necessary details. Additionally, similar to the ALOHA setup [4], we
371 use a tarp around the setup to block unnecessary background distractions. These modifications help
372 enhance the quality of data collected by ensuring that the attention is solely on the manipulation
373 tasks. A photo of our setup is shown in Figure 4.

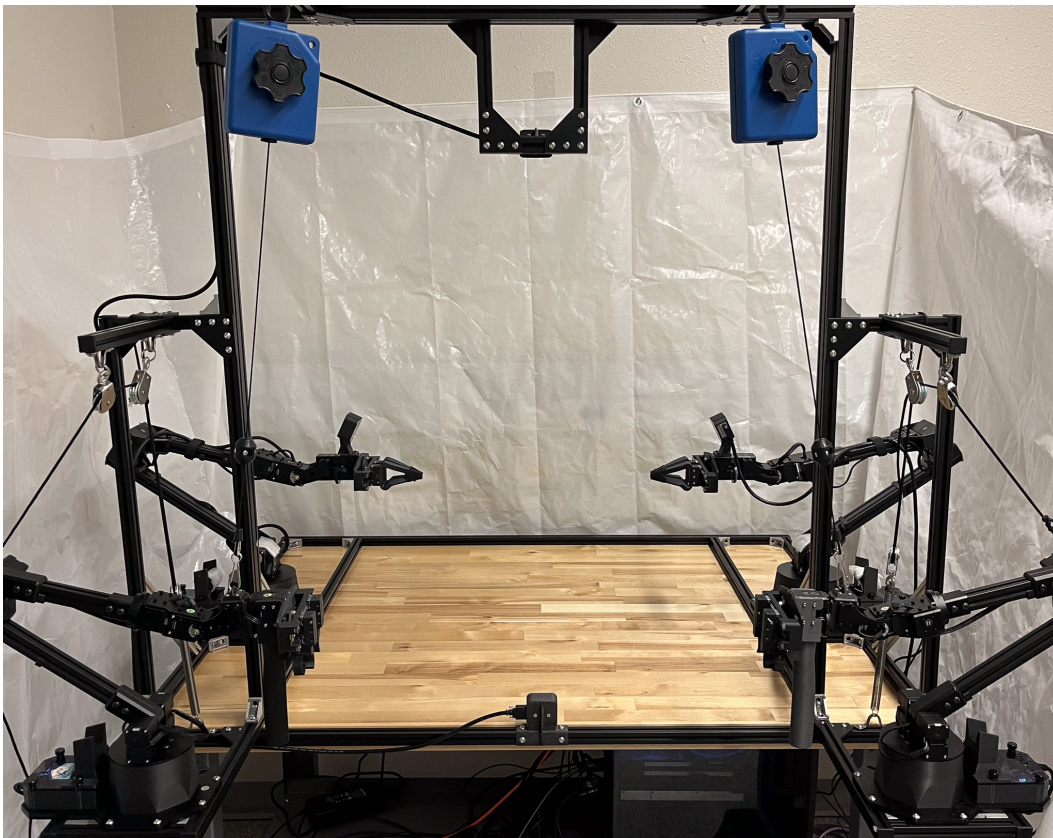


Figure 4: **Our Real-robot Setup.** We have modified the ALOHA 2 setup for our real-world experiments. Modifications include adjusting the camera height and using a tarp around the setup.

374 **Appendix D: Hyperparameters**

375 In this section, we summarize the hyperparameters of InterACT and ACT models used for training
 376 and evaluation in this paper.

Hyperparameters	
# Segment-Wise Encoder Layers	3
# Cross-Segment Encoder Layers	3
# Multi-arm Decoder Layers	4
# Synchronization Block Layers	1
# CLS tokens for Arm Joints	7
# CLS tokens for Visual Features	5

Table 4: Unique hyperparameters of InterACT

Hyperparameters	
# Encoder Layers	4
# Decoder Layers	7

Table 5: Unique hyperparameters of ACT

Hyperparameters	
Learning Rate	1e-5
Batch Size	8
Feedforward Dimension	3200
Hidden Dimension	512
# Heads	8
Chunk Size	50
Beta	10
Dropout	0.1

Table 6: Common hyperparameters of InterACT and ACT