DETACH: CROSS-DOMAIN LEARNING FOR LONG-HORIZON TASKS VIA MIXTURE OF DISENTANGLED EXPERTS

Yutong Shen[1] **Hangxu Liu**[2] **Lei Zhang**[3,†] **Penghui Liu**[1] **Ruizhe Xia**[1]
Tongtong Feng[4,‡]

[1]Beijing University of Technology [2]Fudan University
[3]University of Hamburg [4]Tsinghua University

fengtongtong@tsinghua.edu.cn, zhanglei.cn.de@gmail.com
*

ABSTRACT

Long-Horizon (LH) tasks in Human-Scene Interaction (HSI) are complex multi-step tasks that require continuous planning, sequential decision-making, and extended execution across domains to achieve the final goal. However, existing methods heavily rely on skill chaining by concatenating pre-trained subtasks, with environment observations and self-state tightly coupled, lacking the ability to generalize to new combinations of environments and skills, failing to complete various LH tasks across domains. To solve this problem, this paper presents DETACH, a cross-domain learning framework for LH tasks via biologically inspired dual-stream disentanglement. Inspired by the brain's "where-what" dual pathway mechanism, DETACH comprises two core modules: i) an environment learning module for spatial understanding, which captures object functions, spatial relationships, and scene semantics, achieving cross-domain transfer through complete environment-self disentanglement; ii) a skill learning module for task execution, which processes self-state information including joint degrees of freedom and motor patterns, enabling cross-skill transfer through independent motor pattern encoding. We conducted extensive experiments on various LH tasks in HSI scenes. Compared with existing methods, DETACH can achieve an average subtasks success rate improvement of 23% and average execution efficiency improvement of 29%. More details can be found at: https://sites.google.com/view/detach-learning-anonymous.

1 INTRODUCTION

Long-Horizon (LH) tasks in Human-Scene Interaction (HSI) require continuous planning and cross-domain execution, posing challenges due to their complexity and need for environmental adaptation. These tasks have broad applications in robotics Qiu et al. (2024), medical intervention Kim et al. (2024), and smart homes Kim et al. (2024), with canonical examples including dexterous hand manipulation Zhang et al. (2024) and humanoid whole-body control Sferrazza et al. (2024). However, recent benchmarks show that HSI methods achieve low success rates on cross-domain tasks and demand extensive retraining Zhang et al. (2025); Wang et al. (2024); Xu et al. (2024), severely limiting real-world deployment.

Recent large-scale vision-language-action (VLA) models Black et al. (2024); Team et al. (2025) and agent-based manipulation Ni et al. (2024) achieve strong results on long-horizon embodied tasks. However, both paradigms typically adopt monolithic or tightly coupled end-to-end designs, where perception and control remain entangled, thereby limiting cross-domain generalization and modular skill reuse.

*‡Primary Corresponding Author. †Secondary Corresponding Author.

To bridge these gaps, current approaches Pan et al. (2025); Li et al. (2024); Park et al. (2024) focus on processing self-state information in unified representation spaces, while other solutions Zhang et al. (2025); Xiao et al. (2023); Xu et al. (2024) xiao2023unified, xu2024interdreamer

further encode self-state information mixed with environmental information. The efficacy of the *decompose-reuse-compose* paradigm has been confirmed by various studies Huang et al. (2020); Lan et al. (2023); Xu et al. (2023); Hu et al. (2024), which also introduced a new modular learning paradigm for rapid adaptation to new skills by utilizing skill modules that have already been learned. lan2023contrastive, xu2023composite, hu2024disentangled which also introduced a new modular learning paradigm for rapid adaptation to new skills by utilizing skill modules that have already been learned. In particular, CML Lan et al. (2023) and TokenHSI Pan et al. (2025) have explicitly demonstrated that such modular decomposition significantly outperforms standard end-to-end approaches in multi-task reinforcement learning (RL) and long-horizon HSI, respectively.

Despite their promising performance, these methods suffer from the same architectural flaw: they adopt unified feature representation spaces that tightly couple environmental understanding with self-states. This flaw poses significant challenges in two main aspects: (1) Limited environmental transfer capability: When environmental changes occur (such as shifts from bright laboratory to dim factory settings), these systems cannot effectively separate the effects of environmental changes from self-state changes. This limitation necessitates relearning the entire perception-action mapping Li et al. (2025a), significantly constraining their cross-domain generalization capability. (2) Inefficient skill transfer capability: Current methods fail to achieve functional separation between perception and motor control. When encountering novel skills, even those involving similar motor patterns (such as grasping different objects), the system must retrain the entire perception-action network. This limitation makes it difficult to reuse, prevents effective reuse of learned motor skills, resulting in extremely low knowledge transfer efficiency due to a high risk of skill forgetting van de Ven et al. (2024). Even advanced modular approaches such as CML Lan et al. (2023) and TokenHSI Pan et al. (2025) still rely partly on unified feature spaces, thereby inheriting some of these limitations.

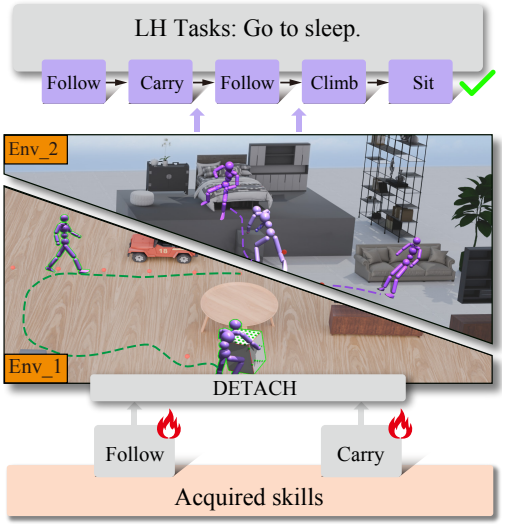

Figure 1: DETACH achieves generative generalization by learning fundamental subtasks in single environment (Env_1), enabling it to generalize to novel environments and accomplish Long-Horizon tasks that involve previously unseen subtasks.

To address these challenges, this paper introduces **DETACH**: a biologically inspired functional disentanglement architecture that draws from the dorsal-ventral stream hypothesis in neuroscience Ungerleider (1982). According to this hypothesis, the brain's ventral *what* pathway specializes in object recognition, while the dorsal *where-how* pathway handles spatial processing and motor control. Unlike existing dual-stream approaches Ibrayev et al. (2024) that separate visual modalities, DETACH introduces a functional disentanglement: the **Environmental Encoder** learns scene-invariant spatial relationships Arkhangelsky & Imbens (2024) while the **Self-Encoder** captures body-schema-specific motor primitives.

Proposed method is extensively evaluated on various self-designed LH-embodied AI tasks, including cross-scene adaptation, novel skill adaptation, and particularly LH control tasks in complex environments. The contributions of this paper can be summarized as follows.

- Proposing the **DETACH disentangled architecture**, the first Embodied AI control framework in HSI based on biologically inspired cognitive principles. This architecture separates traditional unified encoding into specialized parallel processing of environmental perception streams and self-state perception streams.

- Designing **specialized dual-stream encoders**, where the environmental encoder enhances **cross-domain transfer capability**, and the self-encoder achieves **cross-task skill reuse**. Both encoders are independently optimized and flexibly combined.
- Establishing comprehensive benchmark scenarios for LH tasks through designed progressive LH task benchmarks, and validating the effectiveness of DETACH on these benchmarks. Compared to existing methods, DETACH achieves a $2\times$ **improvement in cross-domain adaptation capability** and a $1.5\times$ **improvement in skill reuse efficiency**.

2 Related Works

2.1 Human-Scene Interaction

HSI focuses on enabling embodied agents to interact naturally and effectively with complex 3D environments. Existing approaches include unified representation learning (e.g., Chain of Contacts Xiao et al. (2023)), which integrates contact and object encoding with LLM-based planning but exhibits limited generalizability due to tight coupling between perception and action components; staged processing (e.g., Dynamic HSI Jiang et al. (2024)), which uses autoregressive diffusion for disentangled scene understanding and action generation, ensuring temporal coherence but at high computational cost; and end-to-end methods (e.g., TokenHSI Pan et al. (2025), and Zhang et al. (2025); Xu et al. (2024)), which synthesize motion from text using pre-trained models and object sensors but are limited to simple skill composition scenarios. A key limitation shared by these approaches is their tight coupling between perception and control modules, which hinders cross-domain transfer and skill reusability.

2.2 Long-Horizon Task

LH tasks in HSI require agents to perform multi-step reasoning and manage long-term dependencies Li et al. (2024). Current approaches include hierarchical planning (e.g., MLLM-based instruction parsing with visual encoders Li et al. (2025c); Zheng et al. (2023)), which decomposes tasks into subgoals but suffers from low skill prediction accuracy; memory augmentation (e.g., hierarchical memory and knowledge graphs Li et al. (2024)), which models long-term dependencies yet lack dynamic adaptation; and causal modeling Li et al. (2025b), which enhances policy learning through observation-action causality but requires high computational resources and relies on limited training data. These methods are limited by their reliance on static representations, which constrains cross-domain transfer, policy reuse, and adaptation to dynamic interaction scenarios.

2.3 Disentangled Learning

Disentangled representation learning addresses these limitations by decomposing complex systems into independent, interpretable modules, improving generalization and controllability Ada et al. (2024). Key approaches include mutual information-based disentanglement (e.g., Hu et al. (2024)), which minimizes mutual information between skill components but requires domain-specific prior knowledge; factorized representation learning (e.g., β-VAE framework Uppal et al. (2025)), which uses disentanglement regularization; and variational disentanglement (e.g., Bhowal et al. (2024)), which optimizes a variational lower bound. An alternative method Yang et al. (2025) employs Wasserstein distance for stable disentanglement, though it remains theoretical, while Yang et al. (2025) also identifies valuable factors at high computational cost. However, these methods focus on static factor separation, which are ill-suited for the dynamic, continuous interactions and generative adaptation required in LH embodied tasks.

3 Method

3.1 Observation Space Reconstruction Model

DETACH employs observation disentanglement, modeling unified observation space as a Dual-Stream Separation Process (DSP). The disentanglement objective minimizes mutual information be-

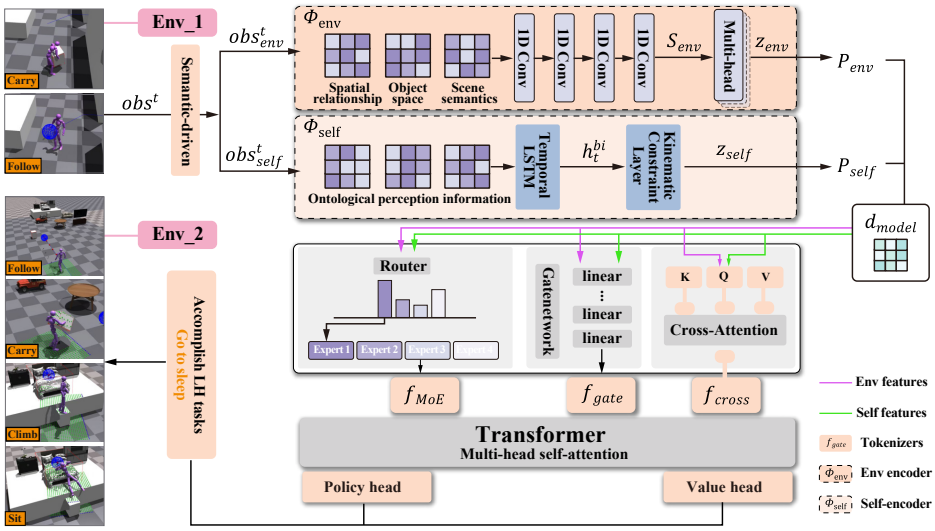

Figure 2: Illustrating the operational workflow of the DETACH, Raw observation obs^t is semantically disentangled into environmental $\text{obs}^t_{\text{env}}$ and self-state $\text{obs}^t_{\text{self}}$ components. Environmental encoder Φ_{env} and self-encoder Φ_{self} process respective inputs, with projection layers P_{env} and P_{self} mapping outputs to unified d_{model} space. Multi-strategy adaptive fusion integrates features via three components: MoE fusion, gated fusion network, and cross-attention fusion module, producing outputs $(f_{\text{MoE}}, f_{\text{gate}}, f_{\text{cross}})$. These fused representations undergo Transformer multi-head self-attention before feeding into policy and value heads.

tween environmental and self-state representations, quantified as $D = \sum_{t=0}^{T} \gamma^t I(obs^t_{env}, obs^t_{self})$, implemented using correlation-based mutual information estimators.

3.2 Disentangled Dual-Encoder

Environmental encoder Φ_{env}. The environmental encoder processes spatial information such as object positions and scene semantics. Given environmental observations $obs^t_{env} \in \mathbb{R}^{T \times d_{env}}$, we adopt parallel convolutional layers for feature extraction. Feature extraction is performed as:

$$S_{env} = \text{Concat}[\text{Conv1D}_k(obs^t_{env})] \tag{1}$$

Features are aggregated through multi-head self-attention for spatial feature aggregation:

$$\begin{aligned} z_{env} = \text{LayerNorm}(\text{MultiHeadAttn}(S_{env}, S_{env}, S_{env}) \\ + S_{env}) \end{aligned} \tag{2}$$

The environmental encoder is paired with decoder Decoder_{env} for reconstruction-based pre-training.

Self-encoder Φ_{self}. The self-encoder processes self-state information $obs^t_{self} \in \mathbb{R}^{T \times d_{self}}$ including joint angles and velocities. Since accurate self-state understanding requires bidirectional temporal context for accurate motion understanding, the self-encoder employs a recurrent neural architecture with bidirectional processing capabilities. The model is defined as:

$$h^{bi}_t = [h^f_t; h^b_t] \tag{3}$$

where h^f_t and h^b_t represent forward and backward temporal representations, respectively.

The kinematic constraint layer ensures outputs remain within physically feasible ranges through a element-wise soft gating mechanism:

$$z_{self} = h^{bi}_t \odot \sigma(W_k h^{bi}_t + b_k) \tag{4}$$

where σ is the sigmoid function, and W_k and b_k are learnable parameters. The kinematic constraint layer ensures the physical feasibility of generated actions.

The self-encoder is paired with a temporal prediction network f_{pred} for sequence prediction-based pre-training, which learns to predict future self-state representations from current ones. *Feature projection layers P_{env} and P_{self}.* Two independent linear layers map the encoder outputs to a unified d_{model} dimensional space:

$$f_{env} = P_{env}(z_{env}), \quad f_{self} = P_{self}(z_{self}) \tag{5}$$

where $f_{env}, f_{self} \in \mathbb{R}^{d_{model}}$ are the projected features used for fusion.

3.3 Multi-Strategy Adaptive Fusion Mechanism

Cross-attention fusion module. This module Vaswani et al. (2017) is selected for its capability to enable internal states to actively query key information from the environment, thereby achieving state-driven dynamic feature alignment. It uses self-state features $f_{self} \in \mathbb{R}^{d_{model}}$ as Query, and environment features $f_{env} \in \mathbb{R}^{d_{model}}$ as Key and Value, achieving dynamic weight allocation through a multi-head attention mechanism:

$$\begin{aligned} f_{cross} &= \text{MultiHead}(f_{self}, f_{env}, f_{env}) \\ &= \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \end{aligned} \tag{6}$$

where each attention head:

$$\text{head}_i = \text{Attention}(f_{self}W_i^Q, f_{env}W_i^K, f_{env}W_i^V) \tag{7}$$

Gated Fusion Network. This module dynamically modulates contribution weights between environmental perception and self-state features to prevent imbalance, using learnable gating units. It is implemented via a multi-layer MLP Tolstikhin et al. (2021) with decreasing hidden units, matching the fused feature dimension. The gated fusion strategy is defined as:

$$\begin{aligned} f_{gate} = \sigma(W_g[f_{env}; f_{self}] + b_g) \odot f_{env} \\ + (1 - \sigma(W_g[f_{env}; f_{self}] + b_g)) \odot f_{self} \end{aligned} \tag{8}$$

Mixture of Experts (MoE) fusion module . This module is adopted for its capacity to dynamically select optimal fusion experts based on task characteristics and environmental complexity, enabling adaptive feature integration. It designs multiple specialized fusion experts Zadouri et al. (2023), each modeled by a multi-layer MLP network with a hierarchical structure, dynamically selecting the most suitable expert for feature fusion through a routing network. The mixture of experts' fusion is represented as:

$$f_{moe} = \sum_{i=1}^{4} w_i \cdot E_i(f_{env}, f_{self}) \tag{9}$$

where the routing weights are

$$w_i = \text{Softmax}(W_r[f_{env}; f_{self}] + b_r)_i \tag{10}$$

The three fusion strategies are combined through a learnable weighted combination:

$$f_{fused} = \alpha \cdot f_{cross} + \beta \cdot f_{gate} + \gamma \cdot f_{moe} \tag{11}$$

where α, β, γ are learnable parameters that balance the contributions of different fusion strategies.

Shared Transformer Encoder. The fused features are processed by a shared transformer encoder ϕ to enhance perception-control collaboration:

$$h_{transformer} = \phi(f_{fused}) \tag{12}$$

where ϕ consists of multiple transformer layers with self-attention mechanisms to capture long-range dependencies and temporal relationships.

Policy and Value Heads. The transformer output is fed into separate policy and value heads for action prediction and value estimation:

$$\begin{aligned} \pi(a|s) &= \text{PolicyHead}(h_{transformer}) \\ V(s) &= \text{ValueHead}(h_{transformer}) \end{aligned} \tag{13}$$

where $\pi(a|s)$ represents the action probability distribution and $V(s)$ represents the state value function.

3.4 PROGRESSIVE TRAINING PROTOCOL

Independent Pre-training Stage. In this stage, the environmental encoder Φ_{env} and self-encoder Φ_{self} are trained independently to establish their respective feature representation capabilities. The environmental encoder is pre-trained through the scene reconstruction loss:

$$\mathcal{L}_{env} = \|\text{Decoder}_{env}(\Phi_{env}(obs^t_{env})) - obs^t_{env}\|_2^2 \tag{14}$$

The self-encoder is pre-trained through action sequence prediction tasks:

$$\mathcal{L}_{self} = \sum_{t=1}^{T-1} \|\Phi_{self}(obs^{t+1}_{self}) - f_{pred}(\Phi_{self}(obs^t_{self}))\|_2^2 \tag{15}$$

where f_{pred} is the temporal prediction network.This stage establishes domain-specific representation foundations.

Fusion Layer Optimization Stage. In this stage, the pre-trained encoder parameters $\theta_{env}, \theta_{self}$ are frozen to preserve learned representations, focusing on training the feature fusion layer and Transformer encoder ϕ:

$$\mathcal{L}_{fusion} = \mathcal{L}_{task} + \lambda_{quality}\mathcal{L}_{fusion\_quality} \tag{16}$$

where $\mathcal{L}_{task}$ represents the standard reinforcement learning objective (e.g., policy gradient loss for PPO), which guides the agent to maximize expected cumulative rewards.

where the fusion quality loss is defined as:

$$\begin{aligned}\mathcal{L}_{fusion\_quality} = \|f_{cross} - (f_{env} + f_{self})\|_2^2 \\ + \lambda_{disentangle} \cdot I(z_{env}, z_{self})\end{aligned} \tag{17}$$

where $I(z_{env}, z_{self})$ represents the mutual information between environmental and self-state features. The first term ensures fusion consistency, while the second term maintains disentanglement by minimizing mutual information between representations.

End-to-End Joint Optimization Stage. In the end-to-end joint optimization stage, all network parameters are unfrozen for end-to-end joint optimization, while introducing specialized preservation regularization:

$$\begin{aligned}\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda_{disentangle} \cdot I(z_{env}, z_{self}) \\ + \sum_i \lambda_i \mathcal{R}_i\end{aligned} \tag{18}$$

4 EXPERIMENT

Our experiments are conducted entirely on three LH tasks that we designed:

LH1: "Sit on Chair!" This LH task comprises a sequence of four fundamental skills: *Follow, Carry, Climb,* and *Sit*, where the target object for *Sit* is a Chair.

LH2: "Sit on Sofa!" This LH task similarly comprises a sequence of four fundamental skills: *Follow, Carry, Follow,* and *Sit*, where the target object for *Sit* is a Sofa.

LH3: "Go to Bed!" This LH task comprises a sequence of five fundamental skills: *Follow, Carry, Follow, Climb,* and *Sit*, where the target object for *Sit* is a Bed.

Our object assets are sourced from the 3D-FRONT dataset Fu et al. (2021), while the motion data is inherited from TokenHSI Pan et al. (2025).

4.1 EVALUATION ON FOUNDATIONAL SKILL LEARNING AND TASK COMPLETION

Experimental Setup. To evaluate the robustness and universality of our disentangled architecture, we employ a progressive learning protocol where foundational skills *Follow* and *Carry* are established through comprehensive training, while *Climb* and *Sit* skills are acquired through compositional learning. This approach enables systematic assessment of skill generalization capabilities and

Method	Follow	Carry	Climb	Sit	LH1
CML Xu et al. (2023)	0.95	0.25	0.00	0.00	0.30
TokenHSI Pan et al. (2025)	1.00	1.00	0.19	0.01	0.55
Ours	0.98	0.97	**0.51**	**0.42**	**0.72**

Table 1: Success rates for foundational skills and composite task completion.

Experiment	Method	Follow	Carry	Follow	Climb	Sit	Time(s)	LH.	SGR.	EGR.
LH2	TokenHSI	1.00	0.56	0.13	-	0.01	99.00	0.42	0.01	0.76
	Ours	**1.00**	**0.96**	**0.67**	**-**	**0.16**	**85.00**	**0.70**	**0.08**	**0.97**
LH3	TokenHSI	1.00	0.50	0.21	0.20	0.00	102.90	0.38	0.67	0.69
	Ours	**1.00**	**0.95**	**0.50**	**0.40**	**0.10**	**97.60**	**0.59**	**0.13**	**0.81**

Table 2: Comparison of generalization performance between TokenHSI and DETACH on LH tasks.

adaptation to diverse environments. The training procedure uses large-scale parallelization in 4,096 environments, employing PPO Schulman et al. (2017) with 10k iterative updates. We conducted 100 independent experimental trials to ensure statistical reliability, quantifying robustness and universality through the success rate means of all skills and L1 tasks. This rigorous evaluation framework provides a comprehensive assessment of the architecture's performance through systematic skill composition and environmental adaptation.

Baselines. We train TokenHSI from scratch using our custom dataset. TokenHSI is a state-of-the-art full-body humanoid controller that learns a set of foundational skills comparable to ours. We also include CML Xu et al. (2023), a composite motion learning baseline commonly used alongside TokenHSI, as an additional point of reference.

Follow and Carry. The success rate of *Follow* is defined as maintaining the pelvis within a 30cm distance threshold from the target path in the XY plane. For the *Carry* task, which can be decomposed into 'grasp' and 'transport' components, achieving only the grasp phase without successful transport to the designated target location is considered 0.5 task completion. *Follow* task training utilized procedurally generated trajectories, while *Carry* task training employed 9 boxes of varying dimensions. Subsequently, we trained on the compositional LH1 task combining these two primitives and evaluated performance on identical task compositions.

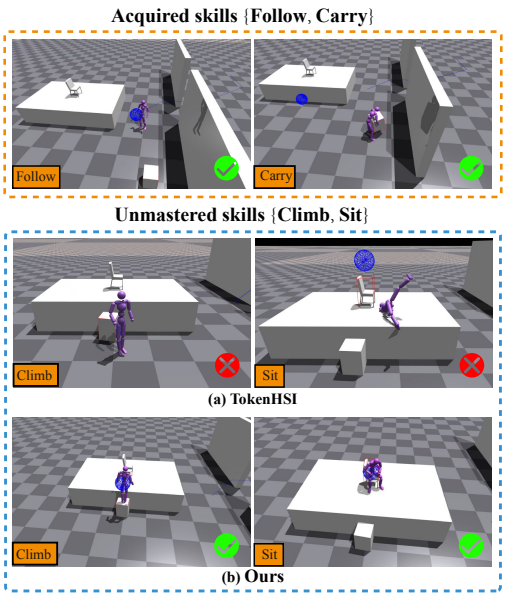

Figure 3: Skill acquisition performance comparison between Detach and TokenHSI. The orange box represents the skills learned in pre-training, and the blue boxes represent new generalized skills.

Climb and Sit. The success of *Climb* is defined as reaching the target object with the pelvis positioned at or above the target elevation. Success of *Sit* requires the pelvis to be positioned on the upper surface of the target object.

LH1 task. The Success rate for LH1 is the success rate of the sub-skill sequence. Due to the sequential nature of LH tasks, where skills must be executed in order, failure in a preceding task prevents the execution of subsequent tasks. Therefore, the skill sequence success rate serves as an excellent metric for evaluating the success rate of LH tasks.

Results. Table I presents the quantitative analysis results, where we evaluated the effectiveness of three methods: CML, TokenHSI, and DETACH. While all methods demonstrate comparable performance on pre-trained skills such as *Follow* and *Carry*, Figure 4 reveals that, compared to methods with limited generalization capabilities like CML and TokenHSI, our DETACH method maintains high success rates for pre-trained skills while achieving success rates of 51% and 42% on two additional tasks, *Climb* and *Sit*, respectively, significantly outperforming the other two approaches. In contrast, TokenHSI and CML exhibit limited generalization on these tasks, resulting in success rates approaching zero. Furthermore, in terms of overall task success rate, DETACH achieves 72%, surpassing CML and TokenHSI by 42% and 17%, respectively. These results highlight DETACH's stability in executing existing skills and its versatility in handling novel tasks, demonstrating its superior performance capabilities.

4.2 Long-Horizon Task Completion

This section evaluates the DETACH framework's performance on Long-Horizon (LH) tasks, designed to test generalization across skills and environments. We focus on **skill generalization** and **environment generalization**, using LH2 and LH3 for assessment, which target adaptation to novel environments and task compositions. Generalization is evaluated over 100 test runs per task, measuring subtask success rates in diverse, unseen scenes to assess robustness.

Task Execution Times. Task execution time refers to the duration from the start of the current LH task to the initiation of the next LH task. The criteria for determining the execution of the next task include the occurrence of errors (such as falling) or exceeding the threshold time for task execution.

Experiment setup. As described in Section 4.1, to validate the environment generalization capability in this section, we employ the same progressive learning protocol on the LH1 task and directly evaluate generalization on the LH2 and LH3 tasks. This approach allows us to observe the environment generalization capability of our DETACH framework more intuitively. Since we similarly establish foundational skills *Follow* and *Carry* through comprehensive training, we can also assess skill generalization rates through the completion performance on *Climb* and *Sit*.

Generalization Rate Definition. Based on our experimental data, we formally define the Environment Generalization Rate (EGR) and Skill Generalization Rate (SGR) as follows:

$$EGR = \frac{S_{Li}}{S_{L1}}, i \in 2, 3 \tag{19}$$

$$SGR = \frac{(S_{climb} + S_{sit})/2}{(S_{follow} + S_{carry})/2} \tag{20}$$

where $S_{Li}, i \in \{1, 2, 3\}$ represents the success rate of LH tasks, and the testing on LH2 and LH3 involves direct transfer from the LH1 environment training, demonstrating its rationality. Similarly, S_{climb} etc. represent the success rates of skills, where *Climb* and *Sit* are composed from foundational *Follow* and *Carry* skills; therefore, we define the skill generalization rate using this formula.

Results. Figure 5 visually highlights DETACH's superior skill composition over TokenHSI. Table II compares subtask success rates, LH task success rates, execution times, and environment/skill generalization rates, showing DETACH's consistent outperformance. Specifically, DETACH reduces average execution times by 14s (LH2) and 5s (LH3) compared to TokenHSI, enhancing efficiency in human body control. Task success rates improve by 28% (LH2) and 21% (LH3), balancing efficiency with success. Notably, environment generalization rates reach 0.08 (LH2) and 0.13 (LH3), with skill generalization rates of 0.97 (LH2) and 0.81 (LH3), significantly exceeding TokenHSI. These results underscore DETACH's enhanced composition capabilities for long-horizon HSI tasks.

4.3 Ablation Experiment

To evaluate the individual contributions of key modules in our DETACH framework, we perform a comprehensive ablation study. Each variant is constructed by disabling or removing a specific component from the full model while keeping all other settings fixed. The experiments are conducted on LH3 tasks composed of foundational skill primitives (e.g., *follow*, *carry*, *climb*, *sit*) in diverse environments.

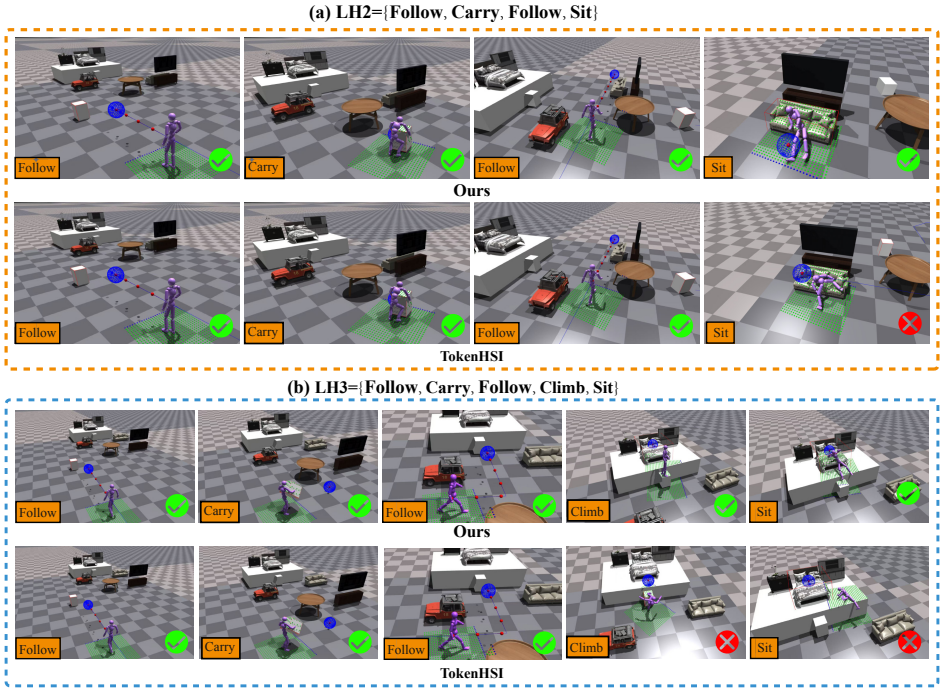

Figure 4: Generalization comparison between DETACH and TokenHSI on LH tasks, where (a) and (b) represent tasks composed of sequences of four and five foundational skills, respectively. We only pre-trained the first two actions, *Follow* and *carry* on LH1 tasks, and tested skill generalization and environmental generalization in new scenarios.

Experimental Setup. We recorded three key metrics: environment generalization success rate, skill generalization success rate, and overall LH task success rate. Each model was trained under the same progressive learning protocol as described in Section 4.1 and evaluated by executing LH3, with results shown in Table III below.

Results. As shown in Table III, removing the environmental encoder (**A1**) causes the environment generalization rate to drop from 0.81 to 0.64, while removing the self encoder (**A2**) leads to a decrease in skill generalization rate from 0.127 to 0.045. Removing any component results in either substantial or moderate degradation across all metrics. This confirms that each encoder in our framework serves a distinct function and is indispensable to the overall architecture.

ID	Configuration	EGR.	SGR.	LH.
Full	All modules enabled	**0.81**	**0.13**	**0.58**
A1	w/o Env Encoder Φ_{env}	0.64	0.12	0.41
A2	w/o Self Encoder Φ_{self}	0.74	0.05	0.38

Table 3: Ablation study results. Each variant disables one key module. Bold indicates best performance.

5 Conclusion

In this work, we presented **DETACH**, a biologically inspired dual-stream disentanglement framework that explicitly separates environment understanding from self-state encoding. This design enables **cross-domain transfer, modular skill reuse, and efficient long-horizon task composition**. Extensive experiments on diverse HSI scenarios demonstrate that DETACH achieves substantial improvements of 23% in subtask success rate and 29% in execution efficiency, along with stronger generalization over state-of-the-art modular baselines. While our current implementation relies on a pre-defined skill set, future work will explore open-ended skill discovery from unlabeled data and real-world deployment under dynamic environments. We believe DETACH provides a promising step toward scalable, generalizable embodied intelligence in complex human-scene interactions.

REFERENCES

Suzan Ece Ada, Erhan Oztop, and Emre Ugur. Diffusion policies for out-of-distribution generalization in offline reinforcement learning. *IEEE Robotics and Automation Letters*, 9(4):3116–3123, 2024.

Dmitry Arkhangelsky and Guido Imbens. Causal models for longitudinal and panel data: A survey. *The Econometrics Journal*, 27(3):C1–C61, 2024.

Pratik Bhowal, Achint Soni, and Sirisha Rambhatla. Why do variational autoencoders really promote disentanglement? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 3817–3849, 2024.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2024. URL `https://arxiv.org/abs/2410.24164`.

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10933–10942, 2021.

Jiaheng Hu, Zizhao Wang, Peter Stone, and Roberto Martín-Martín. Disentangled unsupervised skill discovery for efficient hierarchical reinforcement learning. *Advances in Neural Information Processing Systems*, 37:76529–76552, 2024.

Wenlong Huang, Igor Mordatch, and Deepak Pathak. One policy to control them all: Shared modular policies for agent-agnostic control. In *International Conference on Machine Learning*, pp. 4455–4464. PMLR, 2020.

Timur Ibrayev, Amitangshu Mukherjee, Sai Aparna Aketi, and Kaushik Roy. Toward two-stream foveation-based active vision learning. *IEEE Transactions on Cognitive and Developmental Systems*, 16(5):1843–1860, 2024.

Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1737–1747, 2024.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Siming Lan, Rui Zhang, Qi Yi, Jiaming Guo, Shaohui Peng, Yunkai Gao, Fan Wu, Ruizhi Chen, Zidong Du, Xing Hu, et al. Contrastive modules with temporal attention for multi-task reinforcement learning. *Advances in Neural Information Processing Systems*, 36:36507–36523, 2023.

Sizhe Lester Li, Annan Zhang, Boyuan Chen, Hanna Matusik, Chao Liu, Daniela Rus, and Vincent Sitzmann. Controlling diverse robots by inferring jacobian fields with deep networks. *Nature*, pp. 1–7, 2025a.

Zaijing Li, Yuquan Xie, Rui Shao, Gongwei Chen, Dongmei Jiang, and Liqiang Nie. Optimus-1: Hybrid multimodal memory empowered agents excel in long-horizon tasks. *Advances in neural information processing systems*, 37:49881–49913, 2024.

Zaijing Li, Yuquan Xie, Rui Shao, Gongwei Chen, Weili Guan, Dongmei Jiang, and Liqiang Nie. Optimus-3: Towards generalist multimodal minecraft agents with scalable task experts. *arXiv preprint arXiv:2506.10357*, 2025b.

Zaijing Li, Yuquan Xie, Rui Shao, Gongwei Chen, Dongmei Jiang, and Liqiang Nie. Optimus-2: Multimodal minecraft agent with goal-observation-action conditioned policy. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9039–9049, 2025c.

Minheng Ni, Lei Zhang, Zihan Chen, Kaixin Bai, Zhaopeng Chen, Jianwei Zhang, and Wangmeng Zuo. Don't let your robot be harmful: Responsible robotic manipulation via safety-as-policy. *arXiv preprint arXiv:2411.18289*, 2024.

Liang Pan, Zeshi Yang, Zhiyang Dou, Wenjia Wang, Buzhen Huang, Bo Dai, Taku Komura, and Jingbo Wang. Tokenhsi: Unified synthesis of physical human-scene interactions through task tokenization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5379–5391, 2025.

Core Francisco Park, Andrew Lee, Ekdeep Singh Lubana, Yongyi Yang, Maya Okawa, Kento Nishi, Martin Wattenberg, and Hidenori Tanaka. Iclr: In-context learning of representations. *arXiv preprint arXiv:2501.00070*, 2024.

Ri-Zhao Qiu, Yafei Hu, Yuchen Song, Ge Yang, Yang Fu, Jianglong Ye, Jiteng Mu, Ruihan Yang, Nikolay Atanasov, Sebastian Scherer, et al. Learning generalizable feature fields for mobile manipulation. *arXiv preprint arXiv:2403.07563*, 2024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Carmelo Sferrazza, Dun-Ming Huang, Xingyu Lin, Youngwoon Lee, and Pieter Abbeel. Humanoidbench: Simulated humanoid benchmark for whole-body locomotion and manipulation. *arXiv preprint arXiv:2403.10506*, 2024.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.

Leslie G Ungerleider. Two cortical visual systems. *Analysis of visual behavior*, 549:chapter–18, 1982.

Anshuk Uppal, Yuhta Takida, Chieh-Hsin Lai, and Yuki Mitsufuji. Denoising multi-beta vae: Representation learning for disentanglement and generation. *arXiv preprint arXiv:2507.06613*, 2025.

Gido M van de Ven, Nicholas Soures, and Dhireesha Kudithipudi. Continual learning and catastrophic forgetting. *arXiv preprint arXiv:2403.05175*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19757–19767, 2024.

Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts. *arXiv preprint arXiv:2309.07918*, 2023.

Pei Xu, Xiumin Shang, Victor Zordan, and Ioannis Karamouzas. Composite motion learning with task control. *ACM Transactions on Graphics (TOG)*, 42(4):1–16, 2023.

Sirui Xu, Yu-Xiong Wang, Liangyan Gui, et al. Interdreamer: Zero-shot text to 3d dynamic human-object interaction. *Advances in Neural Information Processing Systems*, 37:52858–52890, 2024.

Yucheng Yang, Tianyi Zhou, Qiang He, Lei Han, Mykola Pechenizkiy, and Meng Fang. Task adaptation from skills: Information geometry, disentanglement, and new objectives for unsupervised reinforcement learning. *arXiv preprint arXiv:2506.10629*, 2025.

Ted Zadouri, Ahmet Üstün, Arash Ahmadian, Beyza Ermiş, Acyr Locatelli, and Sara Hooker. Pushing mixture of experts to the limit: Extremely parameter efficient moe for instruction tuning. *arXiv preprint arXiv:2309.05444*, 2023.

Jinlu Zhang, Yixin Chen, Zan Wang, Jie Yang, Yizhou Wang, and Siyuan Huang. Interactanything: Zero-shot human object interaction synthesis via llm feedback and object affordance parsing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7015–7025, 2025.

Lei Zhang, Kaixin Bai, Guowen Huang, Zhenshan Bing, Zhaopeng Chen, Alois Knoll, and Jianwei Zhang. Contactdexnet: Multi-fingered robotic hand grasping in cluttered environments through hand-object contact semantic mapping. *arXiv preprint arXiv:2404.08844*, 2024.

Sipeng Zheng, Jiazheng Liu, Yicheng Feng, and Zongqing Lu. Steve-eye: Equipping llm-based embodied agents with visual perception in open worlds. *arXiv preprint arXiv:2310.13255*, 2023.