H-GAR: A Hierarchical Interaction Framework via Goal-Driven Observation-Action Refinement for Robotic Manipulation

Yijie Zhu[1,2], **Rui Shao**[1,3,4*], **Ziyang Liu**[1], **Jie He**[1], **Jizhihui Liu**[1], **Jiuru Wang**[5], **Zitong Yu**[2,6*]
[1] Harbin Institute of Technology, Shenzhen
[2] Great Bay University
[3] Shenzhen Loop Area Institute
[4] Shenzhen Ruoyu Technology Co., Ltd.
[5] Linyi University
[6] Dongguan Key Laboratory for Intelligence and Information Technology
shaorui@hit.edu.cn, yuzitong@gbu.edu.cn

Abstract

Unified video and action prediction models hold great potential for robotic manipulation, as future observations offer contextual cues for planning, while actions reveal how interactions shape the environment. However, most existing approaches treat observation and action generation in a monolithic and goal-agnostic manner, often leading to semantically misaligned predictions and incoherent behaviors. To this end, we propose **H-GAR**, a **H**ierarchical interaction framework via **G**oal-driven observation-**A**ction **R**efinement. To anchor prediction to the task objective, H-GAR first produces a goal observation and a coarse action sketch that outline a high-level route toward the goal. To enable explicit interaction between observation and action under the guidance of the goal observation for more coherent decision-making, we devise two synergistic modules. **(1) Goal-Conditioned Observation Synthesizer (GOS)** synthesizes intermediate observations based on the coarse-grained actions and the predicted goal observation. **(2) Interaction-Aware Action Refiner (IAAR)** refines coarse actions into fine-grained, goal-consistent actions by leveraging feedback from the intermediate observations and a **Historical Action Memory Bank** that encodes prior actions to ensure temporal consistency. By integrating goal grounding with explicit action-observation interaction in a coarse-to-fine manner, H-GAR enables more accurate manipulation. Extensive experiments on both simulation and real-world robotic manipulation tasks demonstrate that H-GAR achieves state-of-the-art performance.

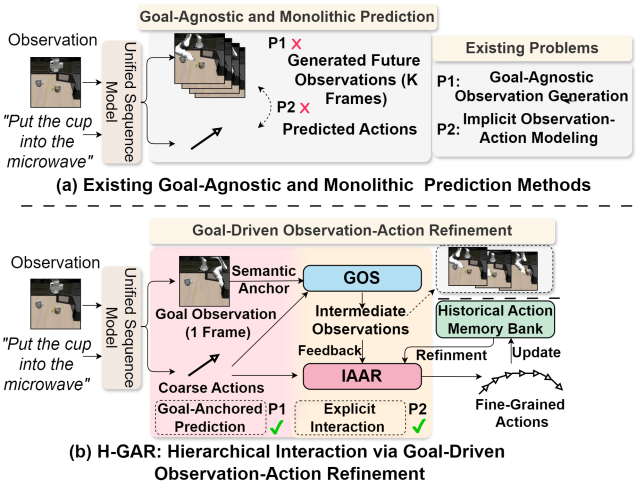

Figure 1: Comparison between existing methods and our proposed H-GAR. (a) Existing approaches follow a goal-agnostic and monolithic prediction paradigm, jointly generating observations and actions without explicit goal grounding or structured interaction. (b) H-GAR introduces a goal-conditioned observation synthesizer and an interaction-aware action refiner, enabling goal-anchored prediction and explicit observation-action interaction.

1 Introduction

Effective planning and manipulation in robotics (Zhou et al. 2025b; Li et al. 2024a; Zhong et al. 2025; Zhou et al. 2025a; Wu et al. 2023; Cheang et al. 2024; Bharadhwaj et al. 2024; Zhu et al. 2025b,c; Zhang et al. 2025a; Li et al. 2025b; Ye et al. 2023), require the ability to anticipate both how the environment evolves and how actions unfold over time. To this end, a growing body of research (Zhao et al. 2025a; Liang et al. 2024; Li et al. 2025a) has shown that jointly predicting future visual observations and action sequences brings substantial benefits to robotic manipulation. This owes to the fact that anticipated visual observations provide rich contextual cues, including spatial layouts, object affordances, and interaction dynamics, which help guide action planning. In parallel, predicted actions expose the causal structure of interactions, enabling more accurate forecasting of how scenes evolve over time. This bidirectional coupling between vision and action has proven critical for modeling real-world dynamics and achieving coherent manipulation.

However, most existing approaches (Li et al. 2025a; Du et al. 2023; Zhu et al. 2025a; Zhang et al. 2025b; Zhao et al. 2025a) adopt a monolithic and goal-agnostic generation strategy. As illustrated in Fig. 1(a), they directly predict

*Corresponding authors.

the entire rollout of future observations and actions from the current observation and instruction, without explicitly modeling how actions influence observations, or how the target goal should guide the trajectory. This design leads to two fundamental limitations: **(1) Goal-Agnostic Observation Generation.** Without an explicit goal, models lack semantic guidance during rollout prediction. This often results in visually plausible but task-irrelevant sequences, undermining interpretability and degrading downstream planning accuracy (Liu et al. 2025; Hu et al. 2023; Brohan et al. 2023). **(2) Implicit Observation-Action Modeling.** Observations and actions are often generated in parallel or via loosely coupled pathways, lacking explicit modeling of their causal interplay. This weakens temporal coherence, limits adaptability, and hinders the mutual reinforcement between perception and manipulation essential for decision-making.

To address these limitations, we propose **H-GAR**, a **H**ierarchical interaction framework via **G**oal-driven observation-**A**ction **R**efinement. It introduces explicit goal grounding and structured bidirectional interaction between observation and action. Concretely, as shown in Fig. 1(b), to provide a semantically grounded plan, H-GAR first predicts a goal observation that represents the final visual state to be achieved, alongside a sequence of coarse-grained actions leading toward the goal. Subsequently, to enable explicit bidirectional interaction between observation and action under the guidance of the goal observation for more coherent decision-making, H-GAR employs two key modules: **(1) Goal-Conditioned Observation Synthesizer (GOS).** To provide semantically meaningful visual guidance aligned with the task goal, GOS synthesizes intermediate observations conditioned on the predicted goal observation and the initially generated coarse-grained actions. These synthesized observations serve as goal-consistent visual context that bridges the gap between high-level task intent and low-level execution. **(2) Interaction-Aware Action Refiner (IAAR).** To refine the initial coarse-grained actions into fine-grained, temporally coherent commands, IAAR first leverages a **Historical Action Memory Bank**, which encodes prior fine-grained actions to guide behaviorally consistent and temporally grounded refinements. It then integrates goal-aligned visual feedback from the intermediate observations synthesized by GOS to further adjust and refine each coarse action. Through this hierarchical coarse-to-fine and goal-conditioned design, H-GAR effectively addresses the limitations of existing approaches. The predicted goal observation anchors generation to the task objective, mitigating semantic drift. while the explicit bidirectional interaction between GOS and IAAR overcomes the implicit modeling of observation and action. By refining coarse action plans into temporally coherent and physically plausible actions, H-GAR enables adaptive, semantically aligned planning, surpassing prior monolithic and goal-agnostic methods in robotic manipulation tasks.

We conduct comprehensive evaluations of H-GAR on both simulation benchmark and real-world robotic manipulation tasks. H-GAR consistently outperforms state-of-the-art approaches, demonstrating particularly strong performance. Ablation studies further confirm the complementary roles and synergistic effects of the GOS and IAAR modules. Our main contributions are as follows:

- We propose **H-GAR**, a goal-driven observation-action refinement framework for robotics that adopts a coarse-to-fine planning paradigm with integrated goal grounding and observation-action interaction.
- We propose **GOS** and **IAAR** to explicitly model the interaction between observation and action under the guidance of goal grounding: GOS synthesizes goal-aligned intermediate observations, and IAAR refines coarse actions using both historical actions and intermediate observations from GOS to produce coherent and task-consistent actions.
- We conduct comprehensive evaluations on both simulation and real-world robotic manipulation tasks, demonstrating the superior performance of H-GAR.

2 Related Work

Goal-Conditioned Planning for Robotics. Recent works have shown that conditioning for goal states can significantly improve planning (Gong et al. 2024; Rens 2025; Shao et al. 2024; Chen et al. 2025; Xie et al. 2025; Zhang et al. 2025c). LBP (Liu et al. 2025) learns the dynamics conditioned by latent goals and performs backward planning from the goal, improving the manipulation in the simulation. In the real world, SayCan (Brohan et al. 2023) combines language instructions with value grounding to guide robotic behavior, demonstrating strong performance in instruction-following tasks. Hu et al.(Hu et al. 2023) propose Planning Exploratory Goals (PEG) to encourage intrinsic goal-driven exploration, while Li et al.(Li et al. 2022) decompose long-horizon tasks into goal-conditioned subtasks using hierarchical policies. Although these approaches leverage goal conditioning to improve efficiency or task decomposition, they often treat goal information as a constraint or policy input, without explicitly incorporating it into the perception-action generation process. In contrast, our method introduces a goal observation as a semantic anchor and explicitly integrates it into both observation synthesis and action refinement.

Future Observation Prediction for Robotic Manipulation. Anticipating future visual observations has proven to be a powerful mechanism for enhancing policy learning in robotics (Hu et al. 2024; Lu et al. 2025; Huang et al. 2025; Zhao et al. 2024; Xu, Qiu, and She 2025; Zhu et al. 2025a). UWM (Zhu et al. 2025a) integrates action and video diffusion within a unified transformer, enhancing robustness and generalization in imitation learning. UVA (Li et al. 2025a) introduces the unified video action model, which jointly optimizes video and action predictions to achieve high accuracy. UP-VLA (Zhang et al. 2025b) trains a unified vision-language-action model with joint multimodal understanding and future prediction objectives, enhancing both high-level semantic reasoning and low-level spatial understanding. PAD (Guo et al. 2024) unifies image prediction and robot action within a joint denoising process. However, these approaches often treat observation and action generation as

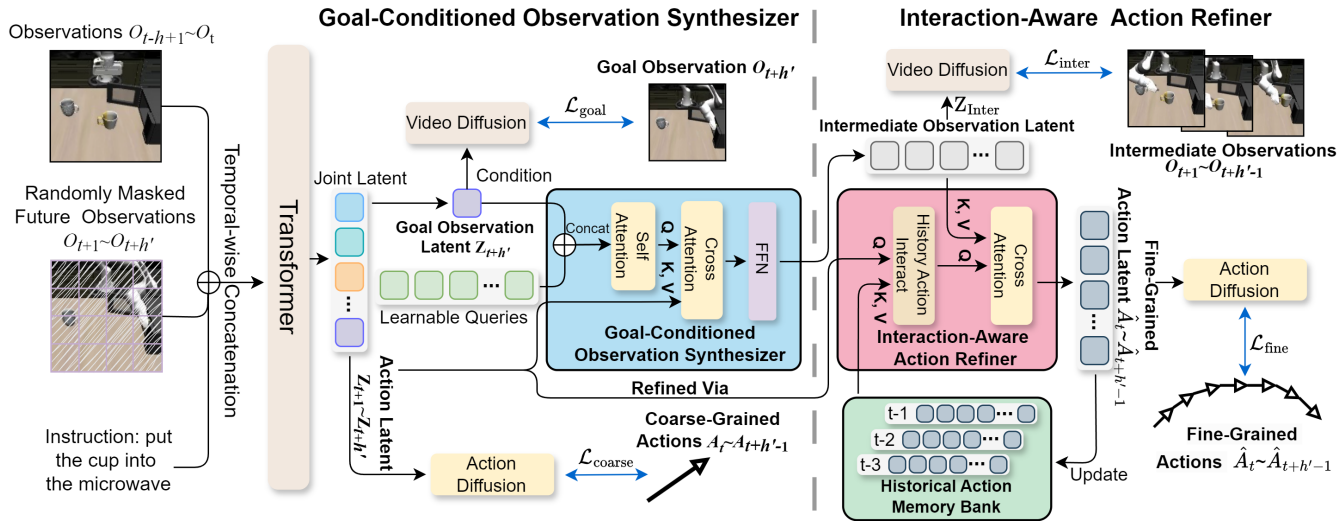

Figure 2: Overview of the H-GAR framework. H-GAR takes a task instruction and past observations $\{O_{t-h+1}, \ldots, O_t\}$, along with masked future observations $\{O_{t+1}, \ldots, O_{t+h'}\}$. The goal observation $O_{t+h'}$, which represents the final visual state to be achieved, is encoded and supervised to produce a semantic latent anchor that conditions the Goal-Conditioned Observation Synthesizer (GOS) to generate intermediate visual features. These features, together with a Historical Action Memory Bank, guide Interaction-Aware Action Refiner (IAAR) in denoising and refining latent actions. Both branches are trained via diffusion objectives to match ground-truth observations and actions. During training, future observations are randomly masked at the same positions across frames to avoid leakage, while inference starts from an empty image.

parallel or loosely coupled processes, lacking structured interaction and goal conditioning. In contrast, our hierarchical coarse-to-fine framework couples observation and action bidirectionally, using goal-grounded prediction and intermediate visual states to enable temporally coherent, semantically aligned manipulation.

3 Methods

3.1 Preliminaries and Problem Formulation

In robotic manipulation, recent studies (Zhao et al. 2025a; Du et al. 2023; Liang et al. 2024; Li et al. 2025a) have demonstrated that predicting future visual observations can significantly enhance policy learning. A common paradigm in such frameworks is defined as follows: Given a task instruction I and a sequence of past visual observations $\{O_{t-h+1}, \ldots, O_t\}$ over a history horizon h, the objective is to jointly predict a sequence of future action chunks $\{A_t, \ldots, A_{t+h'-1}\}$ and the corresponding future observations $\{O_{t+1}, \ldots, O_{t+h'}\}$ over a future horizon h'. Each action chunk $A_t \in \mathbb{R}^{K \times D}$ encodes K consecutive low-level manipulation steps, where each step is represented as a D-dimensional action vector.

While effective, this paradigm presents two key limitations: **(1) Goal-Agnostic Observation Generation.** Without conditioning on a concrete goal observation, the model often generates future observations $\{O_{t+1}, \ldots, O_{t+h'}\}$ that are visually plausible but semantically misaligned with the task objective implied by instruction I. **(2) Implicit Observation-Action Modeling.** Actions $\{A_t, \ldots, A_{t+h'-1}\}$ and observations $\{O_{t+1}, \ldots, O_{t+h'}\}$ are typically predicted independently or with weak interaction, limiting the model's ability to reason about their causal relationship and adapt to evolving visual feedback.

3.2 H-GAR: Overview

To address these limitations, we propose **H-GAR**, which introduces explicit goal grounding and structured bidirectional interaction between observation and action. As illustrated in Fig. 2, instead of predicting the full observation sequence $\{O_{t+1}, \ldots, O_{t+h'}\}$ at once, H-GAR decomposes it into a *goal observation* $O_{t+h'}$—the final state after executing the action sequence—and *intermediate observations* $\{O_{t+1}, \ldots, O_{t+h'-1}\}$ capturing transitional visual feedback. To model their interaction with actions, we introduce the **Goal-Conditioned Observation Synthesizer (GOS)** and **Interaction-Aware Action Refiner (IAAR)**, which jointly enable coarse-to-fine action refinement.

Following prior works (Chang et al. 2022; Li et al. 2024b, 2025a), we encode each image observation into a sequence of N latent tokens using a pretrained VAE (Rombach et al. 2022), followed by a linear projection. Let $\mathbf{V}_{\mathcal{H}}$ and $\mathbf{V}_{\mathcal{F}}$ denote the tokenized representations from past and masked future observations, respectively. The task instruction I is encoded into language embedding T_I using a pretrained CLIP text encoder (Radford et al. 2021). All tokens are concatenated channel-wise to form a multimodal sequence, which is then processed by a Transformer encoder to produce joint latent representations for future steps:

$$\mathbf{Z}_{t+1:t+h'} = \text{Transformer}([\mathbf{V}_{\mathcal{H}},\ \mathbf{V}_{\mathcal{F}},\ T_I]). \quad (1)$$

Here, $\mathbf{Z}_{t+1:t+h'} \in \mathbb{R}^{h' \times N \times D}$ represents the latent features for h' future steps from $t+1$ to $t+h'$, where each step

contains N tokens with D-dimensional embeddings.

To generate a semantic goal anchor, we adopt a lightweight video diffusion decoder (Li et al. 2024b) conditioned on the final latent $\mathbf{Z}_{t+h'}$. Each token $\mathbf{z}_i \in \mathbf{Z}_{t+h'} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ predicts a visual patch through a denoising process, which is decoded by a pretrained VAE to reconstruct the goal frame $O_{t+h'}$. The training objective minimizes the denoising error:

$$\mathcal{L}_{\text{goal}} = \mathbb{E}_{\epsilon,m}\left[\frac{1}{N}\sum_{i=1}^{N}\left\|\epsilon_i - \epsilon_\phi(O_i^{(m)} \mid m, \mathbf{z}_i)\right\|_2^2\right], \quad (2)$$

where $O_i^{(m)}$ is the i-th noisy token of goal observation $O_{t+h'}$ at diffusion step m, and $\epsilon_\phi(\cdot)$ denotes the noise prediction network. Similarly, we aggregate the joint latents $\mathbf{Z}_{t+1:t+h'}$ to generate a coarse action sequence aligned with the goal observation. The training objective is defined as follows:

$$\mathcal{L}_{\text{coarse}} = \mathbb{E}_{\eta,\, m}\left[\left\|\eta - \eta_\theta(A^{(m)} \mid m, \mathbf{Z}_{t+1:t+h'})\right\|_2^2\right], \quad (3)$$

where $A^{(m)}$ is the ground-truth action chunk corrupted with noise at step m, and $\eta_\theta(\cdot)$ is the noise prediction network.

Next, given the above goal observation latent $\mathbf{Z}_{t+h'}$ and the coarse action latent $\mathbf{Z}_{t+1:t+h'}$, the **GOS** generates intermediate observation latent $\mathbf{Z}_{\text{Inter}}$ that bridges the high-level goal and low-level execution. Formally, we define it as:

$$\mathbf{Z}_{\text{Inter}} = \mathbf{GOS}(\mathbf{Z}_{t+h'}, \mathbf{Z}_{t+1:t+h'}). \quad (4)$$

We adopt a similar denoising objective as in goal prediction, using $\mathbf{Z}_{\text{Inter}}$ to reconstruct ground-truth intermediate observations corrupted with noise, resulting in the loss $\mathcal{L}_{\textbf{inter}}$. Finally, the **IAAR** updates the initial coarse action latent $\mathbf{Z}_{t+1:t+h'}$ by integrating intermediate observation features $\mathbf{Z}_{\text{Inter}}$ and the **Historical Action Memory Bank** $\mathcal{H}_t = [\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_{t-1}]$, producing a refined sequence $\hat{A}_{t:t+h'-1}$ that is temporally consistent and semantically aligned with the task goal:

$$\hat{A}_{t:t+h'-1} = \mathbf{IAAR}(\mathbf{Z}_{t+1:t+h'},\ \mathbf{Z}_{\text{Inter}},\ \mathcal{H}_t). \quad (5)$$

Similarly, the refined action sequence $\hat{A}_{t+1:t+h'-1}$ is used to reconstruct the ground-truth actions through a diffusion-based denoising objective, yielding the loss $\mathcal{L}_{\textbf{fine}}$. The overall training objective sums the losses introduced above:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{goal}} + \mathcal{L}_{\text{coarse}} + \mathcal{L}_{\text{inter}} + \mathcal{L}_{\text{fine}}. \quad (6)$$

3.3 H-GAR: Architectural Design of GOS and IAAR

As outlined in Section 3.2, the hierarchical refinement process in H-GAR is implemented by the **GOS** and **IAAR** modules. This section details their architectural design and how they enable structured observation-action interaction.

Goal-Conditioned Observation Synthesis (GOS). Given the coarse action latent $\mathbf{Z}_{t+1:t+h'}$ and the goal observation latent $\mathbf{Z}_{t+h'}$, the GOS module synthesizes intermediate observation features that bridge high-level goal semantics and low-level action dynamics. We introduce a set of learnable queries $\mathbf{Q}_{\text{Inter}} \in \mathbb{R}^{(h'-1)\times N\times D}$, which are designed to represent latent intermediate frames. To incorporate goal observation information, the queries are first updated via a self-attention module where the goal latent $\mathbf{Z}_{t+h'}$ is concatenated to the input:

$$\mathbf{Q}'_{\text{Inter}} = \textit{SelfAttn}([\mathbf{Q}_{\text{Inter}};\ \mathbf{Z}_{t+h'}]), \quad (7)$$

where $\mathbf{Q}'_{\text{Inter}}$ denotes the updated queries obtained from the self-attention module $\textit{SelfAttn}(\cdot)$, and $[\ ;\ ]$ represents the concatenation operation. Through this process, goal observation information is effectively aggregated into the queries. Subsequently, the updated queries $\mathbf{Q}'_{\text{Inter}}$ attend to the coarse action latent representations $\mathbf{Z}_{t+1:t+h'}$, which serve as keys and values, allowing the model to inject action-aware context into the intermediate observation synthesis. The resulting features are then refined through a feed-forward layer:

$$\mathbf{Z}_{\text{Inter}} = \text{FFN}\left(\textit{CrossAttn}(\mathbf{Q}'_{\text{Inter}},\ \mathbf{Z}_{t+1:t+h'})\right), \quad (8)$$

where $\textit{CrossAttn}(\cdot)$ and $\text{FFN}(\cdot)$ denote the cross-attention and feed-forward layers, respectively. The output $\mathbf{Z}_{\text{Inter}}$ aligns with the definition in Eq. 4. This enables GOS to explicitly inject goal semantics and action context into intermediate observation synthesis, establishing a structured interaction between vision and action.

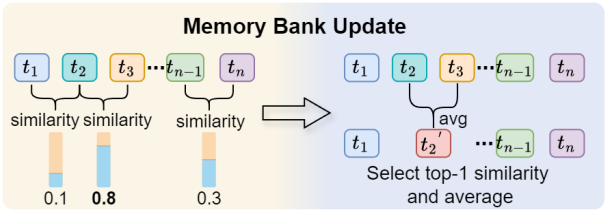

Figure 3: Update strategy of the historical action memory bank. When the memory exceeds a threshold, the most similar adjacent latents are merged via averaging.

Interaction-Aware Action Refiner (IAAR). Building on the intermediate visual representations $\mathbf{Z}_{\text{Inter}}$ synthesized by GOS, we introduce the IAAR to refine coarse action latent $\mathbf{Z}_{t+1:t+h'}$. It integrates two sources of feedback: (1) temporal behavior priors from the **Historical Action Memory Bank** $\mathcal{H}_t = [\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_{t-1}]$, and (2) semantic visual feedback from the intermediate observation features $\mathbf{Z}_{\text{Inter}}$.

Firstly, we apply a history-interact layer $\textit{HisInter}(\cdot)$ to enhance temporal consistency and correct artifacts in the coarse action sequence. In this process, the **Historical Action Memory Bank** $\mathcal{H}_t$—which encodes history fine-grained action latent—serves as both *key* and *value*, while the coarse action latent $Z_{t+1:t+h'}$ acts as the *query*.

$$\widetilde{A} = \textit{HisInter}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V. \quad (9)$$

Secondly, to inject observation-level context and improve semantic alignment with the synthesized intermediate observations, we refine $\widetilde{A}$ via a cross-attention layer $\textit{CrossAttn}(\cdot)$,

Method	PushT		PushT-M		Libero-10	
	SR ↑	RK ↓	SR ↑	RK ↓	SR ↑	RK ↓
Diffusion Policy-C* *[RSS'23]* (Chi et al. 2023a)	0.91	3	0.68	4	0.53	9
Diffusion Policy-T* *[RSS'23]* (Chi et al. 2023a)	0.78	5	0.63	5	0.58	6
UniPi* *[NeurIPS'23]* (Du et al. 2023)	0.42	6	0.19	7	-	-
OpenVLA* *[CoRL'24]* (Kim et al. 2024)	0.35	7	0.22	6	0.54	8
STAR* *[ICML'25]* (Hao et al. 2025)	-	-	-	-	0.89	3
CoT-VLA* *[CVPR'25]* (Zhao et al. 2025b)	-	-	-	-	0.69	5
SpatialVLA* *[RSS'25]* (Qu et al. 2025)	-	-	-	-	0.56	7
PD-VLA† *[arXiv'25]* (Song et al. 2025)	0.82	4	0.71	3	0.92	2
UVA† *[RSS'25]* (Li et al. 2025a)	0.96	2	0.85	2	0.89	3
H-GAR (Ours)	**0.99**	**1**	**0.90**	**1**	**0.94**	**1**

Table 1: Simulation experimental results. Comparison of task success rates (SR) and ranks (RK) across both single-task and multi-task simulation settings. "*" denotes results reported in the original paper, while "†" denotes our reproduced results.

Method	Object Placement		Drawer Manipulation			Towel Folding	Mouse Arrangement
	Cube→Plate	+Toy→Bowl	Open	+Place	+Close		
VQ-BeT* (Lee et al. 2024)	5/10	3/10	4/10	3/10	1/10	-	-
QueST* (Mete et al. 2024)	6/10	4/10	3/10	1/10	0/10	-	-
STAR* (Hao et al. 2025)	8/10	6/10	6/10	4/10	3/10	-	-
PD-VLA† (Song et al. 2025)	8/10	7/10	6/10	**6/10**	4/10	6/10	4/10
UVA† (Li et al. 2025a)	7/10	6/10	6/10	5/10	3/10	5/10	3/10
H-GAR (Ours)	**9/10**	**8/10**	**7/10**	**6/10**	**6/10**	**8/10**	**6/10**

Table 2: Real-World experimental results. "*" denotes results reported in the original paper, while "†" denotes our reproduced results. For long-horizon tasks (Object Placement, Drawer Manipulation), we report stage-wise completion rates.

Method	Libero-10 ↓	Mouse Arrangement ↓
UniPi	56.55	72.56
UVA (1 step)	89.36	59.32
UVA (8 steps)	51.10	32.78
H-GAR (1 step)	**86.76**	**55.17**
H-GAR (8 steps)	**49.01**	**28.43**

Table 3: Observation generation results. Evaluation on simulated (Libero-10) and real-world (Mouse Arrangement) environments across different autoregressive steps.

where the intermediate observation latent $\mathbf{Z}_{\text{Inter}}$ serves as the *key* and *value* and $\widetilde{A}$ acts as the *query*.

$$\hat{A} = \mathit{CrossAttn}(\widetilde{A},\ \mathbf{Z}_{\text{Inter}}). \tag{10}$$

The resulting $\hat{A}$ represents the refined fine-grained action latent $\hat{A}_{t:t+h'-1}$ (as defined in Eq. 5). The historical action memory bank then incorporates this latent and updates its contents following the strategy illustrated in Fig. 3. Specifically, to maintain a compact yet informative memory $\mathcal{H}_t = [\hat{A}_1, \hat{A}_2, \ldots, \hat{A}_{t+h'-1}]$, we apply a redundancy-aware compression when the memory exceeds a threshold. We first compute the cosine similarity between temporally adjacent action latent:

$$s_j = \cos(\hat{A}_j,\ \hat{A}_{j+1}), \quad j \in [1, t+h'-2]. \tag{11}$$

Then we select the highest similarity pair across time and then average them to reduce memory length.

4 Experiments

4.1 Experiment Settings

Simulation Benchmarks. Our simulation experiments include two settings: **Single-Task Evaluation** and **Multi-Task Evaluation**. In the single-task, we train and evaluate separate policies for each task, using **PushT** (Chi et al. 2023b; Florence et al. 2022) as a representative example. In the multi-task setting, a single policy is trained to handle multiple task goals. Following standard protocols, we evaluate on **PushT-M** (Li et al. 2025a) and **Libero-10** (Liu et al. 2023).

Real-World Setup. To validate our approach in the real world, we deploy it on the **Cobot Agilex ALOHA** platform across four manipulation tasks: **Object Placement, Drawer Manipulation, Towel Folding, and Mouse Arrangement**. For the long-horizon tasks Object Placement and Drawer Manipulation, we report stage-wise completion rates.

Baselines. For robotic manipulation, we compare H-GAR with a range of state-of-the-art methods, such as Diffusion Policy-C/T, OpenVLA, SpatialVLA, and CoT-VLA. Diffusion Policy-C and T denote the CNN-based and Transformer-based variants of Diffusion Policy, respectively. For observation generation, we further compare H-GAR with UniPi and UVA, two state-of-the-art approaches in generative modeling for observation.

4.2 Overall Performance

Simulation Experimental Results. As shown in Table 1, H-GAR consistently achieves state-of-the-art performance

GOS	IAAR w/o Bank	IAAR w/ Bank	PushT	PushT-M	Libero-10
			0.90	0.78	0.85
✓			0.92	0.82	0.89
✓	✓		0.96	0.87	0.91
✓		✓	**0.99**	**0.90**	**0.94**

Table 4: Ablation study on model components. IAAR is evaluated with and without historical action memory bank.

Selection Strategy	PushT	PushT-M	Libero-10
Uniform Multi-Frame	0.96	0.83	0.91
Single Random Frame	0.95	0.86	0.88
Goal Frame Only	**0.99**	**0.90**	**0.94**

Table 5: Ablation study on goal-conditioned observation strategy. We compare different input strategies for GOS.

across both single-task and multi-task scenarios, highlighting the robustness of our hierarchical, goal-driven framework in addressing both focused and diverse task distributions. Furthermore, Table 3 shows that H-GAR achieves the lowest FVD scores under both 1-step and 8-step generation in both simulated and real-world settings, indicating superior visual fidelity. This demonstrates the advantage of our hierarchical generation strategy—first producing the goal observation and then synthesizing temporally coherent intermediate frames—over prior one-shot generation methods.

Real-World Experimental Results. We conduct real-world training and evaluation on four diverse manipulation tasks across robotic platform—Cobot Agilex ALOHA—to thoroughly assess the effectiveness and generalizability of our approach. As shown in Table 2, H-GAR achieves the highest stage-wise success rates across all tasks on the ALOHA platform, significantly outperforming prior methods in long-horizon settings such as Object Placement and Drawer Manipulation. Our method also excels in more dynamic and fine-grained tasks like Towel Folding and Mouse Arrangement, demonstrating robust visual grounding and precise control capabilities.

4.3 Ablation Studies

Ablation on Model Components. As shown in Table 4, both the Goal-Conditioned Observation Selection (GOS) and the Interaction-Aware Action Refinement (IAAR) modules contribute significantly to overall performance. In particular, the historical action memory in IAAR consistently improves performance, highlighting the benefit of using temporal action history for policy refinement.

Ablation on Goal-Guided Observation Generation strategy. Table 5 presents a comparison of different input strategies for the GOS module. Across all tasks, conditioning on the predicted goal observation consistently yields the highest success rates, surpassing both random single-frame and uniform multi-frame baselines. This highlights the critical role of explicitly generating a goal observation, which

Settings	PushT	PushT-M	Libero-10
Action Memory Bank Size			
8	0.96	0.88	0.90
16	0.97	0.88	0.91
32	0.98	**0.90**	**0.94**
64	**0.99**	0.89	0.92
Update Strategy			
Random	0.95	0.86	0.89
FIFO	0.97	0.88	0.91
Similarity	**0.99**	**0.90**	**0.94**

Table 6: Ablation study on historical action memory bank. We evaluate the impact of memory bank size and update strategies. Compared strategies include Random, FIFO (First-In, First-Out), and Similarity (Ours), which discards the most similar entry to promote diversity.

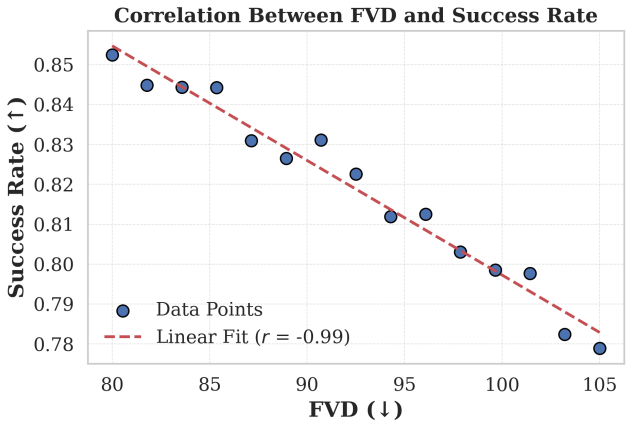

Figure 4: FVD vs. Success Rate on Libero-10. A strong negative correlation is observed, indicating that better observation generation quality leads to higher task success rates.

serves as a strong semantic anchor to align synthesized intermediate observations with the final task objective. In contrast, while other strategies provide additional local visual context, they lack explicit task-driven guidance, often leading to semantically plausible but goal-irrelevant rollouts. By grounding the generation process in a clear, high-level visual target, our *goal-first* design ensures temporally coherent transitions aligned with the intended outcome, highlighting the necessity of this architectural choice.

Ablation on Historical Action Memory Bank. Table 6 explores the design choices of the memory bank in IAAR. Increasing the memory size generally improves performance. However, overly long memory may introduce distracting or outdated historical actions, which can negatively impact policy execution. In terms of update strategies, our proposed similarity-based eviction method outperforms both FIFO and random baselines. The FIFO strategy (first-in, first-out) employs a queue-based mechanism that removes the oldest entry upon insertion. In contrast, the similarity-

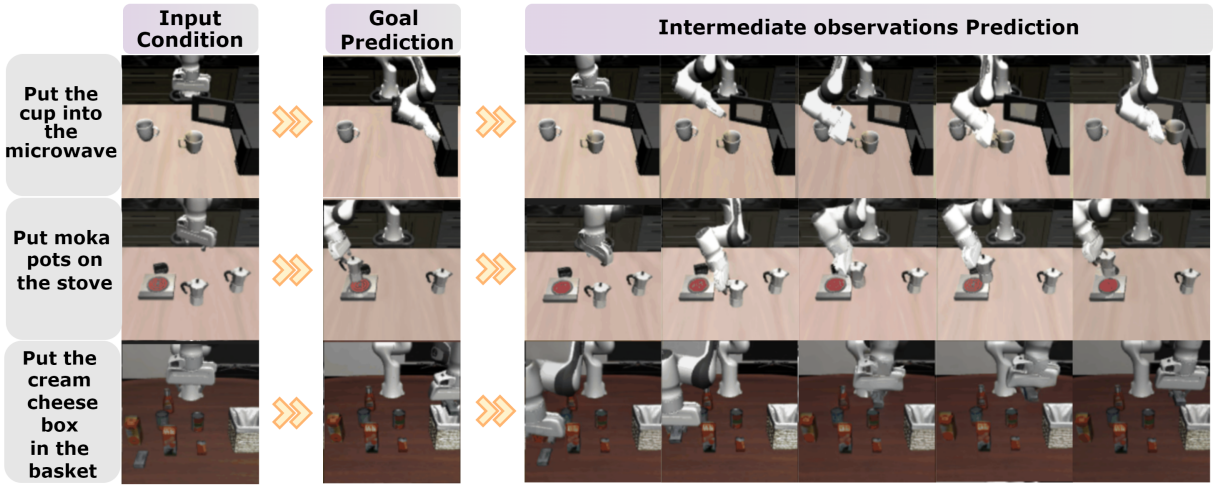

Figure 5: Visualization of goal and intermediate observation prediction. Given a task instruction and initial scene, H-GAR first predicts the goal observation, followed by intermediate frames capturing temporally coherent transitions.

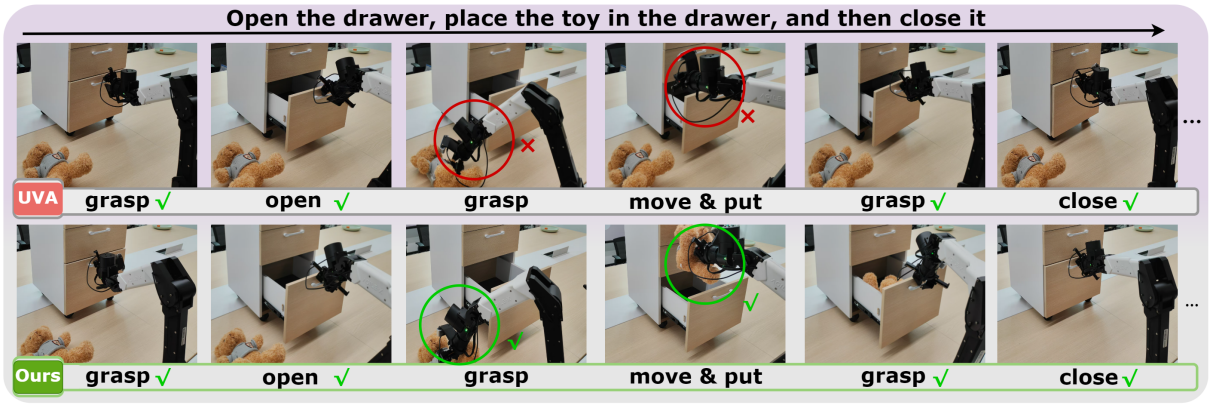

Figure 6: Visualization of real-world execution comparison. Given the long-horizon instruction, H-GAR successfully completes all stages of the manipulation task, achieving coherent transitions, while UVA fails at multiple critical steps.

based strategy enhances diversity by removing the most similar entry, preserving more distinct historical contexts.

The Correlation between Observation Generation and Task Success. Figure 4 shows a strong negative correlation between Fréchet Video Distance (FVD) and task success rate on Libero-10, indicating that higher-quality observation generation leads to better downstream execution. These results validate the effectiveness of our hierarchical, goal-driven design and emphasize the importance of aligning observation generation with action-level reasoning.

4.4 Qualitative Analysis

Figure 5 illustrates the hierarchical prediction process of H-GAR. Given a task instruction and an initial observation, the model first generates a goal observation that anchors the high-level intent, followed by a sequence of intermediate frames that depict temporally coherent transitions toward task completion. Figure 6 shows a real-world comparison between H-GAR and UVA on a complex multi-stage manipulation task. While UVA fails at key stages such as object placement and grasping, H-GAR successfully completes all subtasks—with high accuracy.

5 Conclusion

In this work, we presented H-GAR, a hierarchical goal-driven framework for robotic manipulation that integrates goal-conditioned observation generation with interaction-aware action refinement. By first predicting the goal observation and subsequently generating temporally coherent intermediate transitions, H-GAR facilitates structured long-horizon planning and execution. Extensive experiments across both simulation and real-world environments demonstrate that our method achieves state-of-the-art performance. Comprehensive ablation studies further highlight the critical role of both the Goal Observation Synthesizer (GOS) in providing goal-aligned visual guidance and the Interaction-Aware Action Refiner (IAAR) in facilitating precise and consistent action generation.

Acknowledgments

This study is supported by National Natural Science Foundation of China (Grant No. 62306090), Natural Science Foundation of Beijing, China(Grant No. L254018), Jiangsu Science and Technology Major Program (Grant No. BG2024041), Natural Science Foundation of Guangdong Province of China (Grant No. 2024A1515010147), Natural Science Foundation of Shenzhen City of China (Grant No. JCYJ20250604145700001), Shenzhen Science and Technology Program (KQTD20240729102207002), CCF-Tencent Rhino-Bird Open Research Fund, and Guangdong Basic and Applied Basic Research Foundation (Grant No. 2023A1515140037). The computational resources are supported by SongShan Lake HPC Center (SSL-HPC) in Great Bay University.

References

Bharadhwaj, H.; Vakil, J.; Sharma, M.; Gupta, A.; Tulsiani, S.; and Kumar, V. 2024. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 4788–4795. IEEE.

Brohan, A.; Chebotar, Y.; Finn, C.; Hausman, K.; Herzog, A.; Ho, D.; Ibarz, J.; Irpan, A.; Jang, E.; Julian, R.; et al. 2023. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on robot learning*, 287–318. PMLR.

Chang, H.; Zhang, H.; Jiang, L.; Liu, C.; and Freeman, W. T. 2022. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11315–11325.

Cheang, C.-L.; Chen, G.; Jing, Y.; Kong, T.; Li, H.; Li, Y.; Liu, Y.; Wu, H.; Xu, J.; Yang, Y.; et al. 2024. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*.

Chen, G.; Zhou, X.; Shao, R.; Lyu, Y.; Zhou, K.; Wang, S.; Li, W.; Li, Y.; Qi, Z.; and Nie, L. 2025. Less is More: Empowering GUI Agent with Context-Aware Simplification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Chi, C.; Xu, Z.; Feng, S.; Cousineau, E.; Du, Y.; Burchfiel, B.; Tedrake, R.; and Song, S. 2023a. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 02783649241273668.

Chi, C.; Xu, Z.; Feng, S.; Cousineau, E.; Du, Y.; Burchfiel, B.; Tedrake, R.; and Song, S. 2023b. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 02783649241273668.

Du, Y.; Yang, S.; Dai, B.; Dai, H.; Nachum, O.; Tenenbaum, J.; Schuurmans, D.; and Abbeel, P. 2023. Learning universal policies via text-guided video generation. *Advances in neural information processing systems*, 36: 9156–9172.

Florence, P.; Lynch, C.; Zeng, A.; Ramirez, O. A.; Wahid, A.; Downs, L.; Wong, A.; Lee, J.; Mordatch, I.; and Tompson, J. 2022. Implicit behavioral cloning. In *Conference on robot learning*, 158–168. PMLR.

Gong, X.; Dawei, F.; Xu, K.; Ding, B.; and Wang, H. 2024. Goal-conditioned on-policy reinforcement learning. *Advances in neural information processing systems*, 37: 45975–46001.

Guo, Y.; Hu, Y.; Zhang, J.; Wang, Y.-J.; Chen, X.; Lu, C.; and Chen, J. 2024. Prediction with action: Visual policy learning via joint denoising process. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Hao, L.; Qi, L.; Rui, S.; Xiang, D.; Yinchuan, L.; Jianye, H.; and Liqiang, N. 2025. STAR: Learning Diverse Robot Skill Abstractions through Rotation-Augmented Vector Quantization. *International Conference on Machine Learning (ICML)*.

Hu, E. S.; Chang, R.; Rybkin, O.; and Jayaraman, D. 2023. Planning goals for exploration. *arXiv preprint arXiv:2303.13002*.

Hu, Y.; Guo, Y.; Wang, P.; Chen, X.; Wang, Y.-J.; Zhang, J.; Sreenath, K.; Lu, C.; and Chen, J. 2024. Video prediction policy: A generalist robot policy with predictive visual representations. *arXiv preprint arXiv:2412.14803*.

Huang, S.; Chen, L.; Zhou, P.; Chen, S.; Jiang, Z.; Hu, Y.; Liao, Y.; Gao, P.; Li, H.; Yao, M.; et al. 2025. Enerverse: Envisioning embodied future space for robotics manipulation. *arXiv preprint arXiv:2501.01895*.

Kim, M. J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E.; Lam, G.; Sanketi, P.; et al. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.

Lee, S.; Wang, Y.; Etukuru, H.; Kim, H. J.; Shafiullah, N. M. M.; and Pinto, L. 2024. Behavior generation with latent actions. *arXiv preprint arXiv:2403.03181*.

Li, J.; Tang, C.; Tomizuka, M.; and Zhan, W. 2022. Hierarchical planning through goal-conditioned offline reinforcement learning. *IEEE Robotics and Automation Letters*, 7(4): 10216–10223.

Li, Q.; Liang, Y.; Wang, Z.; Luo, L.; Chen, X.; Liao, M.; Wei, F.; Deng, Y.; Xu, S.; Zhang, Y.; et al. 2024a. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*.

Li, S.; Gao, Y.; Sadigh, D.; and Song, S. 2025a. Unified video action model. *arXiv preprint arXiv:2503.00200*.

Li, T.; Tian, Y.; Li, H.; Deng, M.; and He, K. 2024b. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445.

Li, Z.; Xie, Y.; Shao, R.; Chen, G.; Jiang, D.; and Nie, L. 2025b. Optimus-2: Multimodal minecraft agent with goal-observation-action conditioned policy. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 9039–9049.

Liang, J.; Liu, R.; Ozguroglu, E.; Sudhakar, S.; Dave, A.; Tokmakov, P.; Song, S.; and Vondrick, C. 2024. Dreamitate:

Real-world visuomotor policy learning via video generation. *arXiv preprint arXiv:2406.16862*.

Liu, B.; Zhu, Y.; Gao, C.; Feng, Y.; Liu, Q.; Zhu, Y.; and Stone, P. 2023. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36: 44776–44791.

Liu, D.; Niu, H.; Wang, Z.; Zheng, J.; Zheng, Y.; Ou, Z.; Hu, J.; Li, J.; and Zhan, X. 2025. Efficient Robotic Policy Learning via Latent Space Backward Planning. *arXiv preprint arXiv:2505.06861*.

Lu, Y.; Tian, Y.; Yuan, Z.; Wang, X.; Hua, P.; Xue, Z.; and Xu, H. 2025. H^3DP: Triply-Hierarchical Diffusion Policy for Visuomotor Learning. *arXiv preprint arXiv:2505.07819*.

Mete, A.; Xue, H.; Wilcox, A.; Chen, Y.; and Garg, A. 2024. QueST: Self-Supervised Skill Abstractions for Learning Continuous Control. arXiv:2407.15840.

Qu, D.; Song, H.; Chen, Q.; Yao, Y.; Ye, X.; Ding, Y.; Wang, Z.; Gu, J.; Zhao, B.; Wang, D.; et al. 2025. SpatialVLA: Exploring Spatial Representations for Visual-Language-Action Model. *arXiv preprint arXiv:2501.15830*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rens, G. B. 2025. Proposing Hierarchical Goal-Conditioned Policy Planning in Multi-Goal Reinforcement Learning. *arXiv preprint arXiv:2501.01727*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Shao, R.; Wu, T.; Wu, J.; Nie, L.; and Liu, Z. 2024. Detecting and grounding multi-modal media manipulation and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Song, W.; Chen, J.; Ding, P.; Zhao, H.; Zhao, W.; Zhong, Z.; Ge, Z.; Ma, J.; and Li, H. 2025. Accelerating Vision-Language-Action Model Integrated with Action Chunking via Parallel Decoding. *arXiv preprint arXiv:2503.02310*.

Wu, H.; Jing, Y.; Cheang, C.; Chen, G.; Xu, J.; Li, X.; Liu, M.; Li, H.; and Kong, T. 2023. Unleashing large-scale video generative pre-training for visual robot manipulation. *arXiv preprint arXiv:2312.13139*.

Xie, B.; Shao, R.; Chen, G.; Zhou, K.; Li, Y.; Liu, J.; Zhang, M.; and Nie, L. 2025. GUI-explorer: Autonomous Exploration and Mining of Transition-aware Knowledge for GUI Agent. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Xu, Z.; Qiu, Q.; and She, Y. 2025. VILP: Imitation Learning with Latent Video Planning. *IEEE Robotics and Automation Letters*.

Ye, S.; Wang, Y.; Peng, Q.; You, X.; and Chen, C. P. 2023. The Image Data and Backbone in Weakly Supervised Fine-Grained Visual Categorization: A Revisit and Further Thinking. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhang, H.; Zhang, W.; Qu, H.; and Liu, J. 2025a. Enhancing human-centered dynamic scene understanding via multiple llms collaborated reasoning. *Visual Intelligence*, 3(1): 3.

Zhang, J.; Guo, Y.; Hu, Y.; Chen, X.; Zhu, X.; and Chen, J. 2025b. UP-VLA: A Unified Understanding and Prediction Model for Embodied Agent. *arXiv preprint arXiv:2501.18867*.

Zhang, J.; Tang, P.; Tan, Y.; and Wang, H. 2025c. MGTR-MISS: More Ground Truth Retrieving based Multimodal Interaction and Semantic Supervision for video description. *Neural Networks*, 107817.

Zhao, Q.; Lu, Y.; Kim, M. J.; Fu, Z.; Zhang, Z.; Wu, Y.; Li, Z.; Ma, Q.; Han, S.; Finn, C.; et al. 2025a. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. *arXiv preprint arXiv:2503.22020*.

Zhao, Q.; Lu, Y.; Kim, M. J.; Fu, Z.; Zhang, Z.; Wu, Y.; Li, Z.; Ma, Q.; Han, S.; Finn, C.; et al. 2025b. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1702–1713.

Zhao, W.; Chen, J.; Meng, Z.; Mao, D.; Song, R.; and Zhang, W. 2024. Vlmpc: Vision-language model predictive control for robotic manipulation. *arXiv preprint arXiv:2407.09829*.

Zhong, Y.; Huang, X.; Li, R.; Zhang, C.; Liang, Y.; Yang, Y.; and Chen, Y. 2025. DexGraspVLA: A Vision-Language-Action Framework Towards General Dexterous Grasping. arXiv:2502.20900.

Zhou, Z.; Atreya, P.; Tan, Y. L.; Pertsch, K.; and Levine, S. 2025a. Autoeval: Autonomous evaluation of generalist robot manipulation policies in the real world. *arXiv preprint arXiv:2503.24278*.

Zhou, Z.; Zhu, Y.; Zhu, M.; Wen, J.; Liu, N.; Xu, Z.; Meng, W.; Cheng, R.; Peng, Y.; Shen, C.; and Feng, F. 2025b. ChatVLA: Unified Multimodal Understanding and Robot Control with Vision-Language-Action Model. arXiv:2502.14420.

Zhu, C.; Yu, R.; Feng, S.; Burchfiel, B.; Shah, P.; and Gupta, A. 2025a. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint arXiv:2504.02792*.

Zhu, Y.; Lyu, Y.; Yu, Z.; Shao, R.; Zhou, K.; and Nie, L. 2025b. EmoSym: A Symbiotic Framework for Unified Emotional Understanding and Generation via Latent Reasoning. In *Proceedings of the 33nd ACM International Conference on Multimedia*.

Zhu, Y.; Zhang, L.; Yu, Z.; Shao, R.; Tan, T.; and Nie, L. 2025c. UniEmo: Unifying Emotional Understanding and Generation with Learnable Expert Queries. *arXiv preprint arXiv:2507.23372*.