

---

# Supplementary Material for Reinventing Multi-Agent Collaboration through Gaussian-Image Synergy in Diffusion Policies

---

Ziye Wang<sup>1,2†\*</sup> Li Kang<sup>3†</sup> Yiran Qin<sup>1,4†</sup> Jiahua Ma<sup>1</sup> Zhanglin Peng<sup>2</sup>  
Lei Bai<sup>5</sup> Ruimao Zhang<sup>1‡</sup>

<sup>1</sup>Sun Yat-sen University   <sup>2</sup>The University of Hong Kong   <sup>3</sup>Shanghai Jiao Tong University  
<sup>4</sup>The Chinese University of Hong Kong, Shenzhen   <sup>5</sup>Shanghai AI Laboratory

## A Task Description

We use 6 task in RoboFactory benchmark dataset. Table 1 presents the number of agents for each task, the task description, and the corresponding target condition. Our primary focus is on multi-agent tasks, which evaluate the coordination and cooperation abilities between agents.

## B Global Policy vs. Local Policy

In our main experiments, we employ a Global Diffusion Policy to jointly predict the actions for all agents. To assess the effectiveness of this approach, we compare it against a baseline where each robotic arm is controlled independently by its own Local Diffusion Policy. As shown in Table 2, GauDP with a global policy outperforms its local-policy counterpart in terms of task success rate, highlighting the advantage of modeling inter-agent dependencies through a unified representation to reduce redundant or conflicting actions across agents.

## C Model Size Normalization and Parameter Analysis

To ensure that our observed performance improvement is not simply due to an increase in model size, we conducted model size normalization experiments, summarized in Table 3 and 4. Here, **GauDP-full** includes both the Gaussian estimation and policy learning modules, while **GauDP-policy** only involves the policy component. To ensure a fair comparison, we also introduce **LargeDP**, a scaled-up version of the Diffusion Policy (DP) that matches the parameter count of GauDP-full.

It is worth noting that most parameters in GauDP are **frozen** during policy learning, whereas all parameters in LargeDP are learnable. This comparison reveals that GauDP introduces negligible additional overhead in the policy module while achieving a substantial improvement in task success rate, underscoring the efficiency and effectiveness of our framework.

Even after scaling up DP into LargeDP to match or exceed GauDP’s parameter size, GauDP still achieves the highest success rate across all tasks. This confirms that the performance gain stems from our Gaussian-based global perception framework, not merely from an increase in model capacity. LargeDP continues to lack an effective representation for global perception and fine-grained 3D spatial reasoning, whereas GauDP explicitly encodes these aspects to facilitate cooperative behavior among multiple agents.

---

\*Work completed by Ziye Wang as a visiting research student at Sun Yat-sen University.

†Equal contribution.

‡Corresponding author: Ruimao Zhang ruimao.zhang@ieee.org.

Table 1: Task Descriptions for the RoboFactory Tasks

Task	Agent Number	Description	Target Condition
Lift Barrier	2	A long barrier is placed on the table. Two robotic arms simultaneously grasp both ends of the barrier and lift it to a specified height.	The barrier is elevated to the specified height while maintaining stability.
Place Food	2	A pot and a kind of food are placed on the table. One robotic arm lifts the pot's lid, while the other picks up the food and places it inside the pot.	The food is placed inside the pot, with the distance between the food and the center of the pot being within a predefined threshold.
Two Robots Stack Cube	2	A blue cube and a red cube are placed on the table. A robotic arm picks up the blue cube to a specified position, while the other places the red cube on top of it.	The blue cube is within the specified threshold distance from the target position. The distance between the blue and red cubes remains within a defined threshold, with the red cube positioned at a greater height than the blue cube.
Camera Alignment	3	A camera and an object are placed on the table. One robotic arm picks up the object to a specified position. The other two robotic arms grasp both sides of the camera and align it to the object.	The camera reaches a specified height, and the object is placed at the designated position that aligns with the camera.
Three Robots Stack Cube	3	A blue cube, a red cube and a green cube are placed on the table. One robotic arm picks up the blue cube to a specified position. Another arm places the red cube on top of the blue one. The last arm places the green cube on top of the red one.	The blue cube is positioned within the specified target range. Additionally, the red cube is successfully placed on top of the blue cube, and the green cube is positioned atop the red cube.
Take Photo	4	A camera and an object are placed on the table. One robotic arm picks up the object and places it to a specified position. Another two robotic arms grasp both sides of the camera and align it to the object. The last robotic arm clicks the shutter.	The camera reaches a specified height, and the object is placed at the designated position that aligns with the camera. Additionally, the distance between the end effector of the last robotic arm and the camera's shutter is within a certain threshold.

Table 2: Comparison result on Global Policy and Local Policy. Our GauDP with shared policy achieves the highest average performance across all settings.

Method	2 Arms			3 Arms		4 Arms	Avg.
	Lift Barrier	Place Food	Stack Cube	Align Camera	Stack Cube	Take Photo	
Local GauDP	3%	12%	0%	15%	0%	2%	5.33%
<b>Global GauDP</b>	<b>72%</b>	<b>15%</b>	<b>2%</b>	<b>26%</b>	0%	<b>3%</b>	<b>19.67%</b>

Table 3: Parameter comparison across different policy variants.

Setting	DP	<b>GauDP-policy</b>	LargeDP	GauDP-full
2 Agents	129.949M	129.959M	721.286M	750.505M
3 Agents	163.584M	163.594M	956.335M	784.140M
4 Agents	197.130M	197.140M	1.191G	817.686M

## D Robustness Evaluation

Our Gaussian estimation module is pre-trained on large-scale datasets covering diverse viewpoints and illumination conditions. Furthermore, local depth constraints from multiple viewpoints are incorporated to enhance both accuracy and robustness in Gaussian estimation. This design naturally improves the model's resilience to visual disturbances.

To further assess robustness, we evaluate the models under challenging conditions that include random lighting variations (in color, direction, and intensity) and random distractor objects (10 task-irrelevant

Table 4: Performance comparison after model size normalization. GauDP consistently outperforms baselines of similar or larger size.

Method	2 Arms			3 Arms		4 Arms	Avg.
	Lift Barrier	Place Food	Stack Cube	Align Camera	Stack Cube	Take Photo	
DP	9%	12%	<b>6%</b>	3%	0%	0%	5.00%
LargeDP	60%	12%	4%	<b>29%</b>	0%	0%	17.50%
<b>GauDP</b>	<b>72%</b>	<b>15%</b>	2%	26%	0	<b>3%</b>	<b>19.67%</b>

Table 5: Robustness evaluation under random lighting and distractor objects.

Model	DP	DP3	<b>GauDP</b>
Grab Roller	20%	46%	<b>50%</b>

items placed near the target objects), following the RoboTwin 2.0 benchmark. As shown in Table 5, GauDP maintains the highest success rate even under these challenging scenarios. These results demonstrate that the Gaussian-based scene reconstruction provides strong robustness and consistent 3D perception even under unseen conditions.

## E Limitations

Despite the promising results, our current approach has several limitations. First, due to the limited number and diversity of available embodied tasks and configurations, it is challenging to fine-tune a more generalizable Gaussian-based network capable of adapting to a broad range of embodied scenarios. The scarcity of large-scale, diverse benchmarks hinders the development of a universal 3D representation model for embodied scenarios. Second, our framework currently does not incorporate multi-modal inputs or semantic features. Its potential for scene understanding and interaction could be further enhanced by integrating additional sensory modalities and high-level semantic cues.

## F Broader Impacts

Existing methods for action prediction often emphasize either architectural design or the inclusion of multi-modal observations. In contrast, our work revisits a foundational perspective: that accurate action prediction fundamentally depends on robust 3D scene understanding and localization. We argue that 2D observations alone are insufficient for precise spatial reasoning. Instead, we advocate for explicit 3D reconstruction using Gaussian-based representations as a prerequisite for grounding and decision-making.

In the context of multi-agent collaboration, current approaches often resort to implicitly aggregating information from different viewpoints through concatenation or cross-attention mechanisms, without incorporating any 3D geometric priors. Such strategies are not only inefficient but also prone to spatial inconsistencies. Our framework explicitly reconstructs a coherent global 3D representation from multi-view images and redistributes the resulting global context back to each agent. This facilitates consistent spatial awareness and improves coordination among agents.

Moreover, our framework is not limited to multi-arm manipulation. It is broadly applicable to a variety of embodied scenarios. For example, in complex manipulation tasks, our method can leverage 2D observations to reconstruct a detailed 3D scene. The collaboration between 3D Gaussian representations and 2D images enables more accurate interaction planning and effective collision avoidance in cluttered environments.

Importantly, our method does not require additional sensory inputs and can be seamlessly integrated into existing policy learning pipelines that rely on 2D inputs. This allows for plug-and-play enhancement of existing models by providing stronger 3D understanding and localization capabilities.

Finally, the proposed 3D Gaussian representation is flexible and can be extended to encode semantic features. When tokenized, it can also be incorporated into broader vision-language-action (VLA) models and world models, further bridging the gap between perception and high-level decision making.