GauDP: Reinventing Multi-Agent Collaboration through Gaussian-Image Synergy in Diffusion Policies

Ziye Wang[1,2†*] **Li Kang**[3†] **Yiran Qin**[4†] **Jiahua Ma**[1] **Zhanglin Peng**[2]
Lei Bai[5] **Ruimao Zhang**[1‡]

[1]Sun Yat-sen University [2]The University of Hong Kong [3]Shanghai Jiao Tong University
[4]The Chinese University of Hong Kong, Shenzhen [5]Shanghai AI Laboratory

Abstract

Recently, effective coordination in embodied multi-agent systems remains a fundamental challenge—particularly in scenarios where agents must balance individual perspectives with global environmental awareness. Existing approaches often struggle to balance fine-grained local control with comprehensive scene understanding, resulting in limited scalability and compromised collaboration quality. In this paper, we present *GauDP*, a novel Gaussian-image synergistic representation that facilitates scalable, perception-aware imitation learning in multi-agent collaborative systems. Specifically, *GauDP* constructs a globally consistent 3D Gaussian field from decentralized RGB observations, then dynamically redistributes 3D Gaussian attributes to each agent's local perspective. This enables all agents to adaptively query task-critical features from the shared scene representation while maintaining their individual viewpoints. This design facilitates both fine-grained control and globally coherent behavior without requiring additional sensing modalities. We evaluate *GauDP* on the RoboFactory benchmark, which includes diverse multi-arm manipulation tasks. Our method achieves superior performance over existing image-based methods and approaches the effectiveness of point-cloud-driven methods, while maintaining strong scalability as the number of agents increases. Codes are available at `https://ziyeeee.github.io/gaudp.io/`.

1 Introduction

Multi-agent embodied collaboration [1, 2, 3, 4] is emerging as a key enabler in a wide range of real-world domains, including industrial assembly [5], surgical robotics [6], and assistive household [7] tasks. Unlike single-agent settings, multi-agent collaboration introduces a unique challenge: each agent must complete its assigned task while remaining synchronized with others to avoid catastrophic failures such as collisions or task disruptions.

Existing approaches [8, 9] for multi-agent control typically rely on two paradigms of observation. The first aggregates local observations from all agents and feeds them into a single shared policy (Fig. 1a). While local views offer fine-grained details necessary for precise manipulation, simply concatenating these observations fails to capture the joint collaborative state, often leading to misaligned execution. For instance, one arm may attempt to place food into a pot before the other has finished lifting the lid—resulting in failed coordination. The second paradigm employs a global observation of the entire environment (Fig. 1b), which provides a consistent representation for joint decision-making.

[*]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`.

Preprint. Under review.

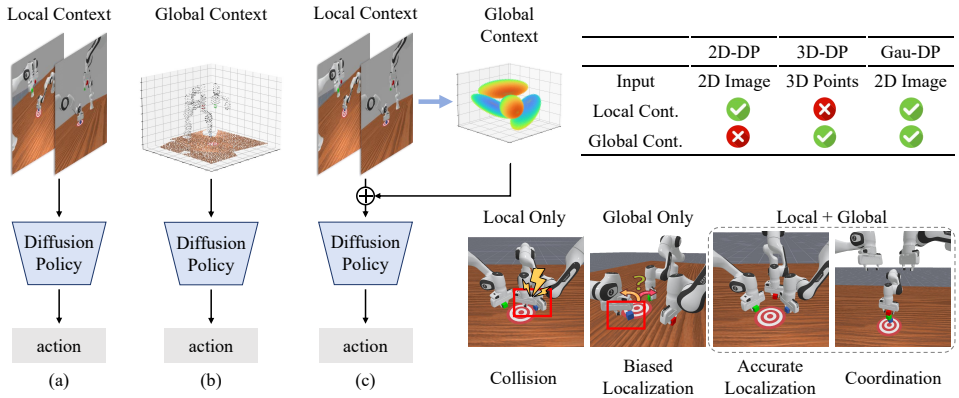

Figure 1: Both local and global context are essential in multi-agent collaboration. Comparison of multi-agent decision-making using different types of contextual information. (a) Using only local context leads to miscoordination such as collisions due to lack of global awareness. (b) Using only global context provides a holistic scene view but lacks detailed local features, resulting in inaccurate control, such as biased localization. (c) Our proposed method, ***GauDP***, fuses global context, which is reconstructed from 2D local images via a shared 3D Gaussian representation, on top of local observations. This integration enables both accurate localization and coordinated execution. Our proposed method, based solely on 2D observations, effectively aggregates global context on top of the local context.

However, this approach often lacks the high-resolution, agent-specific information required for reliable low-level control, such as grasping or placement, thus reducing individual agent performance.

To address this dilemma, effectively integrating both global and local observations is crucial. However, naive fusion of these signals typically lacks 3D structural constraints, making it difficult for the model to reason about spatial relationships and agent-specific contexts. This motivates the need for a unified representation that can simultaneously encode global consistency and local precision.

To this end, we propose ***GauDP***, a unified image-Gaussian representation for multi-agent embodied collaboration (Fig. 1c). Our framework first reconstructs a 3D Gaussian [10] field from the agents' local-view RGB images captured from arbitrary viewpoints. This allows the system to build a globally consistent yet spatially detailed scene representation. Each agent then dynamically queries the shared Gaussian representation to extract task-relevant features for decision-making, enabling coordination while preserving fine-grained control. Importantly, our design naturally scales to more agents without requiring architectural changes, thanks to the flexibility of the Gaussian representation.

We evaluate *GauDP* on the RoboFactory [1] benchmark across diverse multi-arm collaboration tasks. Experiments show that our method significantly outperforms image-based imitation learning methods and achieves performance comparable to point-cloud-based methods such as 3D Diffusion Policy [11], despite using only RGB input. Further ablation studies demonstrate *GauDP*'s robustness and scalability as the number of agents increases. Visualization results confirm its ability to integrate multi-agent observations into a high-quality 3D global representation that improves decision accuracy.

Our contributions are summarized as follows: (1) We introduce *GauDP*, a unified framework that integrates local and global observations via 3D Gaussian fields for multi-agent embodied collaboration. (2) We design a dynamic representation selection mechanism that enables each agent to reason over shared 3D context while maintaining individual precision. (3) We demonstrate the effectiveness and scalability of *GauDP* on the RoboFactory benchmark, achieving strong performance with only RGB input.

2 Related Work

2.1 3D Reconstruction from Multi-View Images

The advent of Neural Radiance Fields (NeRF) [12] and 3D Gaussian Splatting (3DGS) [13] has significantly advanced 3D scene reconstruction by representing entire environments as a unified set

of spatially distributed primitives. These methods are capable of not only accurately reconstructing photorealistic visual appearances but also capturing the underlying 3D geometric structure of a scene from multi-view images. However, achieving high-fidelity reconstructions typically requires densely sampled input views and lengthy optimization time for each scene.

To address this, a growing body of work has focused on adapting NeRF [14, 15, 16, 17, 18] and 3DGS [10, 19, 20, 21, 22, 23] to operate under sparse-view conditions. These approaches typically introduce additional priors, such as semantic information or geometric constraints, to regularize the inherently ill-posed problem of reconstruction from limited viewpoints. Besides, they still often rely on accurate camera poses and sufficient overlap among the views, which are difficult to obtain in real-world robotic manipulation scenarios.

Beyond reconstructing high-fidelity 3D scenes, traditional Structure-from-Motion (SfM) pipelines [24] estimate both 3D structure and camera poses based on sparse feature correspondences. While SfM remains effective in many cases, its performance significantly degrades under wide baselines or extremely sparse views, where reliable feature matching becomes challenging. Recently, learning-based approaches have emerged that directly infer dense 3D geometry from a small number of images [25, 26, 27, 28]. These methods mark a shift toward end-to-end systems that implicitly learn geometric relationships, enabling 3D structure estimation even from as few as two input views.

2.2 Robot Manipulation

Behavioral Cloning (BC)[29, 30, 31, 32, 33] trains policies using pre-recorded human demonstrations to directly imitate expert behaviors, whereas Offline Reinforcement Learning (ORL)[34, 35, 36] refines action selection through reward maximization over large-scale fixed datasets. While BC directly mimics demonstrated behavior, ORL enables further policy improvement by optimizing over offline rewards. Generative approaches have expanded the landscape of policy learning: Action Chunking with Transformers (ACT) combines Transformer architectures with conditional variational autoencoders to capture temporal dependencies in sequential decision-making [37, 38, 39]. More recently, diffusion-based frameworks have shown strong potential in robotic imitation learning due to their high-fidelity trajectory generation. Notable examples include Diffusion Policy [40] and its 3D extension [11], which leverages point cloud inputs to enhance spatial reasoning. The stochastic generative nature of these models makes them especially effective in capturing multimodal action distributions. Demonstration acquisition primarily relies on human-operated robotic systems across diverse tasks [41, 32, 42, 30], while simulator-based trajectory synthesis has emerged as a scalable alternative [43, 44, 45, 46, 47]. Simulated environments allow for controlled task variation and embodiment flexibility. However, existing systems largely focus on single-agent scenarios. Effective data-driven policy learning for multi-agent robotic manipulation—particularly in settings involving coordination among multi-agent—remain significantly underexplored.

2.3 Embodied Multi-Agent Cooperation

Many real-world embodied tasks inherently demand cooperation among multiple agents to achieve complex objectives. A variety of studies [48, 49, 50, 1, 51, 52, 53, 54, 55, 56, 57, 58] have investigated this problem across diverse application domains. Existing research primarily focuses on aspects such as multi-agent task allocation [59, 60, 61] and collaborative decision-making [62, 63, 64]. Recent progress has been driven by the integration of large language models (LLMs), which enable high-level reasoning, communication, and coordination among agents [65, 66, 67, 68, 69, 70, 71]. In parallel, several works employ video generation models [72, 73, 74, 75, 76, 77] to construct multi-agent world models [78], demonstrating promising results in specific scenarios. Nevertheless, these approaches often lack robust visual grounding, which limits their capability to reason about spatial constraints and perceptual affordances. Although recent attempts [79, 80, 81] introduce vision-language models (VLMs) to provide a more grounded perception of the environment, comprehensive research on heterogeneous multi-agent cooperation—particularly in contexts requiring fine-grained visual reasoning and embodied perception—remains limited. In this work, we explicitly combine agent heterogeneity with visual reasoning to enable complex, perception-driven collaboration among embodied agents.

3 Method

3.1 Preliminary

3D Gaussian Splatting (3DGS). 3D Gaussian Splatting [13] represents a 3D scene as a collection of spatially distributed anisotropic Gaussian primitives. Each Gaussian is parameterized by a 3D mean position $\boldsymbol{\mu} \in \mathbb{R}^3$, a scaling vector $\boldsymbol{s} \in \mathbb{R}^3$, a rotation represented by a unit quaternion $\boldsymbol{r} \in \mathbb{R}^4$, an opacity value $\alpha \in \mathbb{R}$, and color information $\boldsymbol{c} \in \mathbb{R}^3$. To model view-dependent effects, the RGB color can be modulated by additional spherical harmonics coefficients $\boldsymbol{h} \in \mathbb{R}^k$. Therefore, a group of Gaussians can be represented as $\mathcal{G} = \{\cup(\boldsymbol{\mu}_i, \boldsymbol{s}_i, \boldsymbol{r}_i, \alpha_i, \boldsymbol{c}_i, \boldsymbol{h}_i)\}$.

Rendering in 3DGS is performed through a differentiable rasterization process that computes each Gaussian's contribution to the image plane. First, all Gaussians are projected into the camera coordinate system. The screen space is then divided into tiles, and Gaussians falling outside the view frustum are efficiently culled to reduce computation. Finally, for each pixel, visible Gaussians are sorted in view-space depth order, and their contributions are composited via alpha blending.

This rendering pipeline enables accurate and efficient supervision of the underlying 3D structure. During optimization, the parameters of the Gaussians are updated to minimize the discrepancy between the rendered images and the ground-truth observations across multiple camera views. Importantly, **when the optimized Gaussian representation yields rendered images that are consistent with the ground-truth across views, it can be regarded as a faithful reconstruction of the true scene geometry**. Formally, let $\mathcal{G}_A$ and $\mathcal{G}_B$ denote two distinct 3D Gaussian configurations, and let $\mathcal{R}(\mathcal{G}, v)$ denote the rendered image of $\mathcal{G}$ under viewpoint v. If the renderings of $\mathcal{G}_A$ and $\mathcal{G}_B$ are identical across all training views $\mathcal{V} = v_1, v_2, \ldots, v_N$:

$$\forall v \in \mathcal{V}, \quad \mathcal{R}(\mathcal{G}_A, v) = \mathcal{R}(\mathcal{G}_B, v),$$

then $\mathcal{G}_A$ and $\mathcal{G}_B$ must be geometrically equivalent, i.e., $\mathcal{G}_A \equiv \mathcal{G}_B$. Conversely, if they are not geometrically equivalent, there must exist at least one view $v \in \mathcal{V}$ such that their renderings differ:

$$\mathcal{G}_A \not\equiv \mathcal{G}_B \quad \Rightarrow \quad \exists v \in \mathcal{V}, \quad \mathcal{R}(\mathcal{G}_A, v) \neq \mathcal{R}(\mathcal{G}_B, v).$$

This implies that multi-view consistency effectively constrains the optimization to a unique, geometrically faithful solution in a self-supervised manner.

Diffusion Policy (DP). Diffusion Policy [82] formulates action generation as a conditional denoising diffusion process. Given a sequence of past observations $\mathcal{O} = \{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_N\}$, the goal is to generate a future action sequence $\boldsymbol{a} = \{a_1, a_2, \ldots, a_L\}$.

The target action sequence $\boldsymbol{a}$ is gradually perturbed by Gaussian noise through a forward diffusion process:

$$q(\boldsymbol{a}^k \mid \boldsymbol{a}^{k-1}) = \mathcal{N}\Big(\sqrt{1-\beta_k}\,\boldsymbol{a}^{k-1},\, \beta_k \mathbf{I}\Big), \quad k = 1, \ldots, K,$$

where β_k is the noise variance schedule. The reverse process learns to iteratively denoise a random sample $\boldsymbol{a}^K \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ back to a clean action sequence, conditioned on observations $\mathcal{O}$:

$$p_\Phi(\boldsymbol{a}^{k-1} \mid \boldsymbol{a}^k, \mathcal{O}) = \mathcal{N}\big(\boldsymbol{\mu}_\Phi(\boldsymbol{a}^k, \mathcal{O}, k),\, \Sigma_\Phi(\boldsymbol{a}^k, \mathcal{O}, k)\big),$$

where Φ denotes the parameters of the conditional denoising network. Through iterative denoising, the learned policy π_Φ generates action trajectories by:

$$\pi_\Phi(\boldsymbol{a} \mid \mathcal{O}) = \mathbb{E}_{\boldsymbol{a}^K \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\prod_{k=1}^{K} p_\Phi(\boldsymbol{a}^{K-k} \mid \boldsymbol{a}^{K-k+1}, \mathcal{O}) \right].$$

3.2 Problem Formulation

We consider the problem of predicting future action sequences for multi-arm embodied agents based on multi-view visual observations. Let $\mathcal{O} = \{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_N\}$ denote a set of synchronized observations captured from N views. Each image $\mathcal{I}_i$ captures the scene from a unique perspective, offering complementary geometric information. The prediction target is a sequence of future actions $\boldsymbol{a} = \{a_1, a_2, \ldots, a_L\}$, where $a_t \in \mathbb{R}^d$ represents the control signal at timestep t.

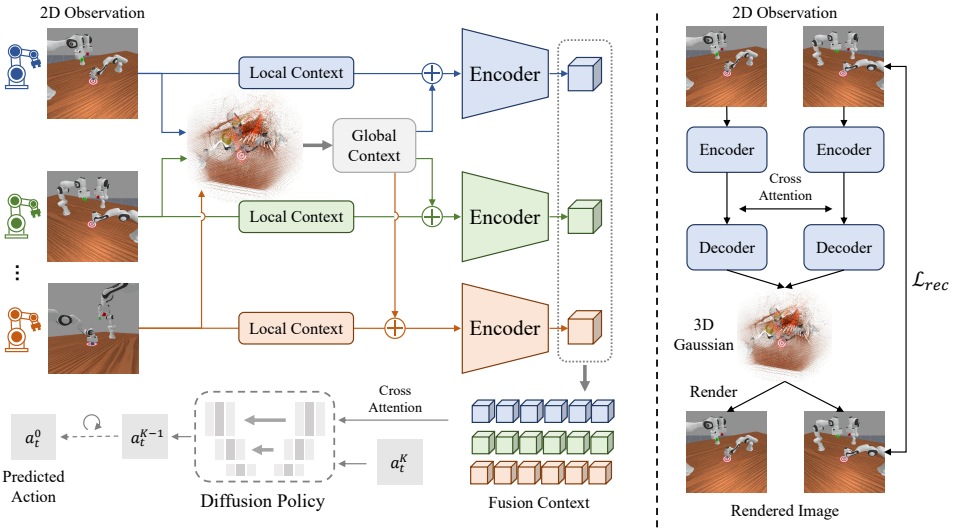

Figure 2: **(a)** Overview of the proposed ***GauDP*** framework for multi-agent imitation learning. Each agent extracts a local context from its 2D observation. A shared 3D Gaussian field is constructed from all views to form the global context, which is fused with the local context and passed through an encoder. The resulting per-agent features are processed by a diffusion policy via cross-attention to predict actions. **(b)** Pipeline for constructing the global Gaussian field. Multi-view images are encoded and aggregated via cross-attention, followed by a reconstruction loss $\mathcal{L}_{\text{rec}}$ between rendered and input views to ensure consistency.

Rather than directly predicting $\boldsymbol{a}$ from 2D image features, we first reconstruct a compact and differentiable 3D Gaussian representation $\mathcal{G}$ as a global context from $\mathcal{O}$. We define a conditional policy π_Φ parameterized by Φ that generates the action sequence $\boldsymbol{a}$ conditioned on both the raw visual inputs $\mathcal{O}$ and the reconstructed 3D representation $\mathcal{G}$:

$$\pi_\Phi(\boldsymbol{a} \mid \mathcal{O}) := \pi_\Phi(\boldsymbol{a} \mid \mathcal{O}, \mathcal{G}).$$

Where, $\mathcal{G} = \mathcal{F}(\mathcal{O})$ and $\mathcal{F}(\cdot)$ is a mapping from multi-view observations to a set of Gaussians. To model the complex conditional distribution over action sequences, we adopt the Diffusion Policy framework. Let $\boldsymbol{a}$ denote the target action sequence, and let $\boldsymbol{a}^K \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ be a sample from an isotropic Gaussian prior. The generative process denoises $\boldsymbol{a}^K$ through a learned reverse process conditioned on $(\mathcal{O}, \mathcal{G})$:

$$\pi_\Phi(\boldsymbol{a}_t \mid \mathcal{O}, \mathcal{G}) = \mathbb{E}_{\boldsymbol{a}_t^K \sim \mathcal{N}(0,\mathbf{I})} \left[\prod_{i=1}^{K} p_\Phi(\boldsymbol{a}_t^{K-i} \mid \boldsymbol{a}_t^{K-i+1}, \mathcal{O}, \mathcal{G}) \right],$$

where p_Φ denotes the learned denoising transition at each timestep. This framework allows the policy to generate realistic and context-aware action trajectories without relying on strong priors over the action space.

3.3 Overview

Mainstream approaches typically train policies to predict future actions directly from either image observations or global point clouds. However, these methods face limitations in multi-agent collaborative tasks: they either rely solely on local observations for each agent or require access to a centralized global point cloud, which restricts both accurate localization and effective coordination among agents. To address these challenges, we propose a novel framework solely based on image observations that enables collaborative multi-agent manipulation through effective global context integration. Specifically, our method fuses multi-view observations to reconstruct a unified global context and selectively distributes task-relevant features back to individual agents as needed. This design not only enhances inter-agent coordination but also improves precise perception and localization. An overview of the proposed framework is illustrated in Fig. 2(a).

3.4 Global Context Reconstruction

In this work, we define a global context as a unified, view-independent representation built within a common 3D coordinate space. This representation should not only preserve color information from raw multi-view image observations, but also restore the underlying 3D structure of the scene reconstructed from these views. To achieve this, we design a framework that reconstructs 3D scenes in a self-supervised manner using only the multi-view 2D observations typically employed for training diffusion policies. Our reconstruction framework is built upon 3D Gaussian Splatting (3DGS). However, conventional 3DGS methods suffer from two major limitations: First, they require densely sampled views with accurate camera poses. Second, they demand scene-specific optimization that can take several minutes per scene. These constraints render them impractical for embodied scenarios, where rapid adaptation and generalization are essential.

To overcome these challenges, we adopt Noposplat [83], a feed-forward network capable of directly reconstructing 3D Gaussian representations from sparse and unposed views. We further fine-tune the pretrained Noposplat model using multi-view observations collected from our robotic manipulation scenarios, which are the same data used to train our downstream diffusion policy.

As illustrated in Fig.2(b), each RGB image is independently encoded by a shared-weight ViT [84] encoder across all views. The resulting per-view features are then passed through a cross-view ViT decoder, which fuses information across different perspectives using cross-attention layers in each transformer block. Finally, a Gaussian parameter prediction head estimates a set of 3D Gaussians for each pixel based on the fused features. This process can be expressed as:

$$\mathcal{G}_i = \mathcal{F}(\mathbf{x}_i), \quad \forall i \in \mathcal{I},$$

where $\mathbf{x}_i$ denotes the fused feature at pixel i, $\mathcal{F}(\cdot)$ is the mapping network, and $\mathcal{G}_i \in \mathbb{R}^{C_{\mathcal{G}} \times H \times W}$ represents the estimated parameters of the corresponding 3D Gaussian.

To further improve the fidelity of the reconstructed 3D structure, we introduce an additional depth supervision during fine-tuning. Specifically, in the rendering process, each estimated 3D Gaussian is projected onto the camera coordinate system. Instead of computing the RGB contribution of each Gaussian to the image pixels, we compute the contribution of its projected depth to the corresponding pixel. This depth rendering process yields a synthetic depth map $\hat{D}$, which can then be supervised using available ground-truth depth D via a reconstruction loss $\mathcal{L}_{\text{depth}}$. This depth-based supervision provides stronger geometric guidance and encourages the model to recover 3D Gaussians that are more consistent with the actual scene geometry. The overall reconstruction loss is defined as:

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{rgb}} + \alpha \cdot \mathcal{L}_{\text{depth}},$$

where α is a balancing weight that controls the influence of the depth supervision.

It is worth noting that depth maps and camera poses are used only during the fine-tuning stage of Noposplat. During policy training and inference, our framework solely relies on multi-view RGB observations to infer the 3D Gaussians, making it lightweight and pose-free at deployment time.

3.5 Global Context Allocation and Pixel-level Synergy

The reconstructed global context encodes rich multi-view and multi-agent information, capturing both the semantic and geometric structure of the scene. However, directly feeding the entire global context to each agent is suboptimal, as it introduces irrelevant information and may interfere with the agent's ability to focus on task-relevant cues from its own perspective. Moreover, effective synergy between global and local context remains underexplored. Existing approaches typically aggregate global and local information only at a coarse level, which fails to capture fine-grained spatial alignment and task-specific dependencies. This coarse fusion strategy may lead to diluted feature representations and impaired action reasoning, especially in densely interactive multi-agent scenarios.

To address these challenges, we introduce a selective global context dispatch mechanism along with a pixel-aligned fusion strategy for fine-grained integration of global and local information. Recall that in Section 3.4, we reconstruct 3D Gaussians by aggregating multi-view image tokens via cross-attention. Each predicted Gaussian encodes both visual appearance and geometry within a unified global coordinate system, while remaining naturally aligned with the input image pixels from which it was derived. Leveraging this alignment, we selectively dispatch the predicted 3D Gaussians back

Table 1: Quantitative Comparison of 3D Gaussian Reconstruction. Improved visual quality reflects higher accuracy in 3D reconstruction.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Pretrain	17.918	0.580	0.492
Ours	**23.424**	**0.779**	**0.148**

to the corresponding agent's observation frame based on their image of origin. Instead of distributing the entire global context indiscriminately, each agent receives only the subset of Gaussians associated with its own view. These Gaussians have already integrated information from other views during reconstruction, thus providing a distilled and relevant global summary for that agent.

For synergistic fusion, we transform the selected Gaussians back into a 2D grid that matches the spatial dimensions of the original image. These global context features are then concatenated with the agent's local image features and passed through a lightweight convolutional fusion module, which learns to combine the complementary strengths of local perception and global understanding.

This design ensures that each agent benefits from a targeted and contextually relevant global representation, while preserving spatial consistency and enabling pixel-level synergy between local and global cues, both of which are critical for precise and coordinated action planning in multi-agent manipulation tasks.

4 Experiment

4.1 Experiment Setup

Dataset. Imitation learning in multi-agent collaborative manipulation presents significant challenges due to the high levels of complexity, coordination, synchronization, and symmetry awareness required, making it difficult to collect high-quality data in real-world scenarios. To address this issue, we leverage the RoboFactory benchmark [1], an automated data collection framework specifically designed for embodied multi-agent systems. We select 6 tasks from RoboFactory involving collaborative manipulation using two to four robotic arms. These tasks are designed to cover a range of coordination complexities and physical interaction patterns. Please refer to the Appendix for detailed descriptions of each task.

Baseline. Existing diffusion-policy-based approaches predominantly focus on either 2D or 3D modalities. To ensure a comprehensive and fair comparison, we evaluate our method against several representative baselines in both domains. For 2D vision-based observations, we adopt Diffusion Policy [82] and 2D Dense Policy [85]; for 3D input modalities, we include 3D Diffusion Policy [11] and 3D Dense Policy [85] as baselines. For fair comparison, we maintain a consistent visual backbone across all methods: ResNet-18 is used for 2D visual inputs following the original Diffusion Policy, and a lightweight MLP is used for 3D data as in 3D Diffusion Policy.

Experiment setting. We use success rate as the primary evaluation metric to assess the effectiveness of each policy. Evaluation is performed every 100 training epochs over 100 episodes per policy. All experiments are implemented using the PyTorch framework and conducted on a single NVIDIA A800 GPU. Policies are trained for 100 epochs using a batch size of 32. We adopt the Adam optimizer with an initial learning rate of 10^{-4}, combined with a warm-up phase followed by cosine decay scheduling. To ensure a fair comparison, all baseline and ablation models are trained using the same set of hyperparameters and optimization settings.

For fair comparison, our proposed *GauDP* and all baseline methods are trained under identical hyperparameters and optimization settings following standard Diffusion Policy benchmarks. Specifically, we use an action prediction horizon of 8, 3 observation steps, and 6 action execution steps. Both *GauDP* and DP adopt DDPM with 100 denoising steps, while DP3 employs DDIM with the same number of steps.

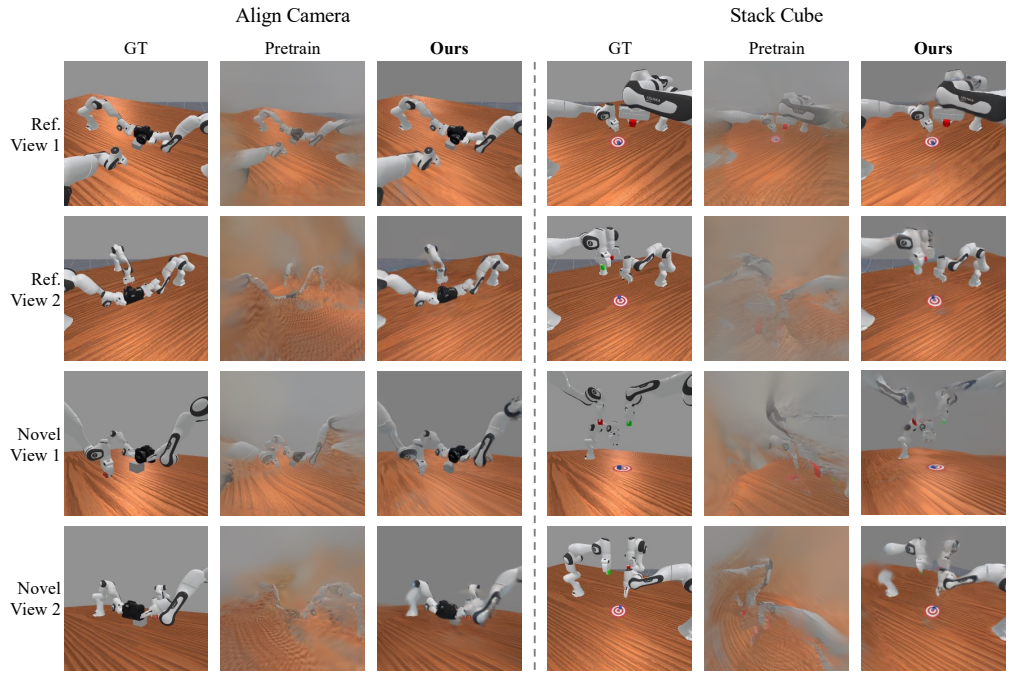

Figure 3: Visualization of Reconstruction Results. Our method achieves significantly improved reconstruction quality.

4.2 Experiment Results

Reconstruction Results. As discussed in Section 3.1, higher-quality rendered images indicate more accurate reconstruction of the underlying 3D scene, including both geometry and color information. To evaluate reconstruction performance, we conduct experiments where the full scene is reconstructed using observations from only two reference viewpoints. Quantitative results are summarized in Table 1, and qualitative comparisons of the rendered Gaussians from both reference and novel views are shown in Figure 3.

Our finetuned model significantly outperforms the pretrained baseline, producing reconstructions that are not only sharper and more detailed, but also more faithful to the original scene geometry and appearance. As shown in Figure 3, our method yields consistent and high-fidelity renderings across both reference and novel views, with clearly defined object boundaries. In contrast, the pretrained model often produces blurry and distorted results, with noticeable discrepancies between the reconstructions and the original images. Besides, in the first two columns, the reconstructed positions of the robot arms differ considerably from those in the ground truth. Our method consistently maintains high structural fidelity and visual consistency, demonstrating its effectiveness in capturing the accurate 3D structure of the scene.

Point Cloud-based Diffusion Policy. The DP3 model leverages 3D point cloud observations to inform multi-agent manipulation. As shown in the table 2, this policy achieves moderate performance across two-arm tasks, such as 30% on Lift Barrier and 21% on Place Food, indicating that point cloud inputs capture sufficient geometric structure for spatially grounded actions. However, its effectiveness drops sharply in tasks with more agents and higher coordination requirements—such as only 1% on Stack Cubes (2 arms). These results suggest that point cloud-based methods lack the fine-grained local control afforded by geometric inputs like point clouds, which limits their effectiveness in precision-critical tasks such as stacking cubes, especially in multi-arm settings.y in tasks with more agents and higher coordination requirements.

Image-based Diffusion Policy. As shown in the table 2, our method GauDP-prefuse significantly outperforms prior 2D diffusion policies across multiple tasks. Compared to DP[40] and 2D Dense Policy[85], which struggle across the board (mostly below 10%), *GauDP* achieves a remarkable 72%

Table 2: Success rate comparison across multi-agent manipulation tasks with different 2–4 arms setting. Underlines denote the best baseline of point cloud-based diffusion policy(DP); **bold** highlights the best of image-based DP. ***GauDP*** 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	
DP3(XYZ) [11]	30%	21%	1%	3%	0%	9%	10.67%
DP3(XYZ+RGB) [11]	31%	25%	1%	18%	0%	11%	14.33%
3D Dense Policy [85]	28%	18%	0%	0%	0%	7%	8.83%
DP [82]	9%	12%	**6%**	3%	0%	0%	5.00%
2D Dense Policy [85]	3%	2%	0%	0%	0%	**9%**	2.33%
GauDP	**72%**	**15%**	2%	**26%**	0%	3%	**19.67%**

Table 3: Training and inference efficiency on the *Lift Barrier* task (Training on A100 GPU; inference on NVIDIA RTX 5090 GPU).

Method	Training Time (GPU h)	Inference Speed (FPS)
DP	4.8	1.49
DP3	2.5	1.57
GauDP	6.5	1.28

success rate in the Lift Barrier task, indicating its superior ability to model geometry-aware visual representations. Additionally, *GauDP* shows strong performance in tasks requiring semantic alignment, such as Align Camera (26%), where other methods fail to generalize. These results demonstrate that integrating geometric priors into image-based pipelines enables better spatial understanding and more effective multi-agent coordination, especially in tasks with complex embodiment and scene variability.

Training and Inference Efficiency. We evaluate the training and inference efficiency of *GauDP* compared with diffusion-based baselines. As shown in Table 3, GauDP requires slightly longer training time due to its geometry-aware modules, but maintains comparable inference speed while offering superior performance.

Real-World Experiments. To further verify the practicality, we conduct real-robot experiments on three representative multi-agent collaboration tasks: *Card Box Stacking*, *Card Box Handover*, and *Grab Roller*. As shown in Table 4, *GauDP* consistently outperforms DP across all tasks, particularly in stacking and handover scenarios that require precise spatial coordination.

Discussion. As shown in Table 2, our method ***GauDP*** achieves the highest average success rate of 19.67% across diverse multi-agent manipulation tasks, significantly outperforming all baselines, including those relying on 3D point cloud inputs such as DP3 [11] and 3D Dense Policy [85]. Notably, while our method operates solely on 2D RGB inputs, it surpasses several 3D-based counterparts in tasks that require global coordination (e.g., Align Camera: 26% vs. 18%) and even matches them in fine-grained manipulation tasks (e.g., Stack Cube: 2% vs. 1%). These results highlight that our approach's strength lies not in the modality itself, but in the design of visual representations that fuse both local visual details and global spatial context. This balance enables agents to perceive geometric structure from images and reason over scene-level relationships, allowing for generalizable cooperation without relying on explicit 3D geometry.

4.3 Ablation Study

Ablation on the Coordinate System of Gaussians. We replace the original camera-coordinate parameterization of Gaussians with a unified world-coordinate system aligned to the first observation frame. Results show that using local coordinates performs better, as it preserves agent-centric spatial

Table 4: Real-robot performance comparison across three multi-agent collaboration tasks. Each score in the table is reported as m/n, where m denotes the number of successful executions and n represents the total number of rollouts performed for that task.

Method	Card Box Stacking			Card Box Handover		Grab Roller	
	Place Succ.	Stack Succ.	Succ.	Place Succ.	Handover Succ.	Succ.	Succ.
DP	19/30	11/19	11/30	22/30	14/22	14/30	22/30
GauDP	**23/30**	**17/23**	**17/30**	**24/30**	**19/24**	**19/30**	**27/30**

Table 5: Ablation study on key components of ***GauDP*** across multi-agent manipulation tasks. **Bold** highlights the best among image-based configurations. Our full model achieves the highest average success rate, demonstrating the effectiveness of combining geometric and visual cues.

Method	2 Arms			3 Arms		4 Arms	Avg
	Lift Barrier	Place Food	Stack Cube	Align Camera	Stack Cube	Take Photo	
w/ unify coor.	30%	1%	**8%**	26%	0%	0%	10.83%
w/o prefuse	2%	4%	0%	1%	0%	0%	1.17%
w/o Image	32%	7%	0%	**28%**	0%	0%	11.17%
w/o Gaussian	9%	12%	6%	3%	0%	0%	5.00%
Ours	**72%**	**15%**	2%	26%	0%	**3%**	**19.67%**

relationships and avoids alignment errors across diverse viewpoints. **Ablation on Fusion Strategies for Local and Global Context.** We replace the default fine-grained, pixel-level fusion with a coarse feature-level concatenation of independently encoded modalities. Performance declines with this coarse fusion, likely due to the loss of spatial alignment and fine-grained cross-modal reasoning. **Ablation on the Role of Image and Gaussian.** We remove either the image input or the Gaussian representation during policy training and inference. Using both inputs achieves the best results, as images provide appearance cues while Gaussians supply global geometric context.

5 Conclusion

In this paper, we present *GauDP* a novel framework for multi-agent collaboration through Gaussian-image synergy in diffusion policies. Specifically, *GauDP* a globally consistent 3D Gaussian representation from the local RGB observations of each agent and reallocates the Gaussian information back to individual agents. This process significantly enhances each agent's perception of the global task information, thereby boosting the success rate of complex collaborative tasks. We evaluate the performance of *GauDP* using the RoboFactory benchmark, which features a diverse set of multi-arm manipulation tasks. As the number of agents increases, *GauDP* not only outperforms existing image-based methods but also matches the effectiveness of point-cloud-driven methods. Future directions will focus on: (1) designing Gaussian representations that are more suitable as inputs for the Vision-Language-Action model to enhance its capabilities in multi-agent collaboration; and (2) leveraging Gaussians to improve the representation of dynamic scenes, enabling them to play a role in world models designed for multi-agent environments.

References

[1] Y. Qin, L. Kang, X. Song, Z. Yin, X. Liu, X. Liu, R. Zhang, and L. Bai, "Robofactory: Exploring embodied agent collaboration with compositional constraints," *arXiv preprint arXiv:2503.16408*, 2025.

[2] B. Huang, Y. Chen, T. Wang, Y. Qin, Y. Yang, N. Atanasov, and X. Wang, "Dynamic handover: Throw and catch with bimanual hands," *arXiv preprint arXiv:2309.05655*, 2023.

[3] Z. Mandi, S. Jain, and S. Song, "Roco: Dialectic multi-robot collaboration with large language models," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 286–299, IEEE, 2024.

[4] M. Chang, G. Chhablani, A. Clegg, M. D. Cote, R. Desai, M. Hlavac, V. Karashchuk, J. Krantz, R. Mottaghi, P. Parashar, *et al.*, "Partnr: A benchmark for planning and reasoning in embodied multi-agent tasks," *arXiv preprint arXiv:2411.00081*, 2024.

[5] D. Jiang, H. Wang, and Y. Lu, "Mastering the complex assembly task with a dual-arm robot: A novel reinforcement learning method," *IEEE Robotics & Automation Magazine*, vol. 30, no. 2, pp. 57–66, 2023.

[6] J. W. Kim, T. Z. Zhao, S. Schmidgall, A. Deguet, M. Kobilarov, C. Finn, and A. Krieger, "Surgical robot transformer (srt): Imitation learning for surgical tasks," *arXiv preprint arXiv:2407.12998*, 2024.

[7] T. Zhang, D. Li, Y. Li, Z. Zeng, L. Zhao, L. Sun, Y. Chen, X. Wei, Y. Zhan, L. Li, *et al.*, "Empowering embodied manipulation: A bimanual-mobile robot manipulation dataset for household tasks," *arXiv preprint arXiv:2405.18860*, 2024.

[8] J.-J. Jiang, X.-M. Wu, Y.-X. He, L.-A. Zeng, Y.-L. Wei, D. Zhang, and W.-S. Zheng, "Rethinking bimanual robotic manipulation: Learning with decoupled interaction framework," *arXiv preprint arXiv:2503.09186*, 2025.

[9] Q. Lv, H. Li, X. Deng, R. Shao, Y. Li, J. Hao, L. Gao, M. Y. Wang, and L. Nie, "Spatial-temporal graph diffusion policy with kinematic modeling for bimanual robotic manipulation," *arXiv preprint arXiv:2503.10743*, 2025.

[10] H. Xiong, S. Muttukuru, R. Upadhyay, P. Chari, and A. Kadambi, "Sparsegs: Real-time 360deg sparse view synthesis using gaussian splatting," *arXiv preprint arXiv:2312.00206*, 2023.

[11] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," in *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[12] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[13] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering.," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[14] A. Jain, M. Tancik, and P. Abbeel, "Putting nerf on a diet: Semantically consistent few-shot view synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5885–5894, 2021.

[15] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, "pixelnerf: Neural radiance fields from one or few images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4578–4587, 2021.

[16] W. Yang, G. Chen, C. Chen, Z. Chen, and K.-Y. K. Wong, "S^3-nerf: Neural reflectance field from shading and shadow under a single viewpoint," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1568–1582, 2022.

[17] B. G. Gerats, J. M. Wolterink, and I. A. Broeders, "Nerf-or: neural radiance fields for operating room scene reconstruction from sparse-view rgb-d videos," *International journal of computer assisted radiology and surgery*, pp. 1–10, 2024.

[18] H. Xu, A. Chen, Y. Chen, C. Sakaridis, Y. Zhang, M. Pollefeys, A. Geiger, and F. Yu, "Murf: multi-baseline radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20041–20050, 2024.

[19] Q. Zhao and S. Tulsiani, "Sparse-view pose estimation and reconstruction via analysis by generative synthesis," *Advances in Neural Information Processing Systems*, vol. 37, pp. 111899–111922, 2024.

[20] D. Charatan, S. L. Li, A. Tagliasacchi, and V. Sitzmann, "pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19457–19467, 2024.

[21] Y. Chen, H. Xu, C. Zheng, B. Zhuang, M. Pollefeys, A. Geiger, T.-J. Cham, and J. Cai, "Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images," in *European Conference on Computer Vision*, pp. 370–386, Springer, 2024.

[22] Y. Wan, M. Shao, Y. Cheng, and W. Zuo, "S2gaussian: Sparse-view super-resolution 3d gaussian splatting," *arXiv preprint arXiv:2503.04314*, 2025.

[23] H. Huang, Y. Wu, C. Deng, G. Gao, M. Gu, and Y.-S. Liu, "Fatesgs: Fast and accurate sparse-view surface reconstruction using gaussian splatting with depth-feature consistency," *arXiv preprint arXiv:2501.04628*, 2025.

[24] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

[25] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "Dust3r: Geometric 3d vision made easy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024.

[26] J. Edstedt, Q. Sun, G. Bökman, M. Wadenbäck, and M. Felsberg, "Roma: Robust dense feature matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19790–19800, 2024.

[27] V. Leroy, Y. Cabon, and J. Revaud, "Grounding image matching in 3d with mast3r," in *European Conference on Computer Vision*, pp. 71–91, Springer, 2024.

[28] Y. Cabon, L. Stoffl, L. Antsfeld, G. Csurka, B. Chidlovskii, J. Revaud, and V. Leroy, "Must3r: Multi-view network for stereo 3d reconstruction," *arXiv preprint arXiv:2503.01661*, 2025.

[29] M. Dalal, A. Mandlekar, C. R. Garrett, A. Handa, R. Salakhutdinov, and D. Fox, "Imitating task and motion planning with visuomotor transformers," in *Conference on Robot Learning*, 2023.

[30] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "Bc-z: Zero-shot task generalization with robotic imitation learning," in *Conference on Robot Learning*, pp. 991–1002, PMLR, 2022.

[31] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan, "Vima: General robot manipulation with multimodal prompts," in *Fortieth International Conference on Machine Learning*, 2023.

[32] A. Mandlekar, D. Xu, R. Martín-Martín, S. Savarese, and L. Fei-Fei, "Learning to generalize across long-horizon tasks from human demonstrations," *arXiv preprint arXiv:2003.06085*, 2020.

[33] T. Ma, J. Zhou, Z. Wang, R. Qiu, and J. Liang, "Contrastive imitation learning for language-guided multi-task robotic manipulation," *arXiv preprint arXiv:2406.09738*, 2024.

[34] D. Kalashnikov, J. Varley, Y. Chebotar, B. Swanson, R. Jonschkowski, C. Finn, S. Levine, and K. Hausman, "Mt-opt: Continuous multi-task robotic reinforcement learning at scale," *arXiv preprint arXiv:2104.08212*, 2021.

[35] A. Kumar, A. Singh, F. Ebert, M. Nakamoto, Y. Yang, C. Finn, and S. Levine, "Pre-training for robots: Offline rl enables learning new tasks from a handful of trials," *arXiv preprint arXiv:2210.05178*, 2022.

[36] Y. Chebotar, Q. Vuong, K. Hausman, F. Xia, Y. Lu, A. Irpan, A. Kumar, T. Yu, A. Herzog, K. Pertsch, *et al.*, "Q-transformer: Scalable offline reinforcement learning via autoregressive q-functions," in *Conference on Robot Learning*, pp. 3909–3928, PMLR, 2023.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[38] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.

[39] T. Buamanee, M. Kobayashi, Y. Uranishi, and H. Takemura, "Bi-act: Bilateral control-based imitation learning via action chunking with transformer," in *2024 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*, pp. 410–415, IEEE, 2024.

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

[41] F. Ebert, Y. Yang, K. Schmeckpeper, B. Bucher, G. Georgakis, K. Daniilidis, C. Finn, and S. Levine, "Bridge data: Boosting generalization of robotic skills with cross-domain datasets," *arXiv preprint arXiv:2109.13396*, 2021.

[42] A. Mandlekar, Y. Zhu, A. Garg, J. Booher, M. Spero, A. Tung, J. Gao, J. Emmons, A. Gupta, E. Orbay, *et al.*, "Roboturk: A crowdsourcing platform for robotic skill learning through imitation," in *Conference on Robot Learning*, pp. 879–893, PMLR, 2018.

[43] Y. Mu, T. Chen, S. Peng, Z. Chen, Z. Gao, Y. Zou, L. Lin, Z. Xie, and P. Luo, "Robotwin: Dual-arm robot benchmark with generative digital twins (early version)," *arXiv preprint arXiv:2409.02920*, 2024.

[44] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "Rlbench: The robot learning benchmark & learning environment," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3019–3026, 2020.

[45] S. Tao, F. Xiang, A. Shukla, Y. Qin, X. Hinrichsen, X. Yuan, C. Bao, X. Lin, Y. Liu, T. kai Chan, Y. Gao, X. Li, T. Mu, N. Xiao, A. Gurha, Z. Huang, R. Calandra, R. Chen, S. Luo, and H. Su, "Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai," *arXiv preprint arXiv:2410.00425*, 2024.

[46] S. Nambiar, M. Jonsson, and M. Tarkian, "Automation in unstructured production environments using isaac sim: A flexible framework for dynamic robot adaptability," *Procedia CIRP*, vol. 130, pp. 837–846, 2024.

[47] S. Nasiriany, A. Maddukuri, L. Zhang, A. Parikh, A. Lo, A. Joshi, A. Mandlekar, and Y. Zhu, "Robocasa: Large-scale simulation of everyday tasks for generalist robots," in *Robotics: Science and Systems*, 2024.

[48] X. An, C. Wu, Y. Lin, M. Lin, T. Yoshinaga, and Y. Ji, "Multi-robot systems and cooperative object transport: Communications, platforms, and challenges," *IEEE Open Journal of the Computer Society*, vol. 4, pp. 23–36, 2023.

[49] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang, "Large language model based multi-agents: A survey of progress and challenges," *arXiv preprint arXiv:2402.01680*, 2024.

[50] R. Li, Z. Hu, W. Qu, J. Zhang, Z. Yin, S. Zhang, X. Huang, H. Wang, T. Wang, J. Pang, *et al.*, "Labutopia: High-fidelity simulation and hierarchical benchmark for scientific embodied agents," *arXiv preprint arXiv:2505.22634*, 2025.

[51] Y. Qin, E. Zhou, Q. Liu, Z. Yin, L. Sheng, R. Zhang, Y. Qiao, and J. Shao, "Mp5: A multi-modal open-ended embodied system in minecraft via active perception," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16307–16316, 2024.

[52] H. Zhang, C. Zhu, X. Wang, Z. Zhou, S. Hu, and L. Y. Zhang, "Badrobot: Jailbreaking llm-based embodied ai in the physical world," *arXiv preprint arXiv:2407.20242*, 2024.

[53] E. Zhou, Y. Qin, Z. Yin, Y. Huang, R. Zhang, L. Sheng, Y. Qiao, and J. Shao, "Minedreamer: Learning to follow instructions via chain-of-imagination for simulated-world control," *arXiv preprint arXiv:2403.12037*, 2024.

[54] Y. Qin, A. Sun, Y. Hong, B. Wang, and R. Zhang, "Navigatediff: Visual predictors are zero-shot navigation assistants," *arXiv preprint arXiv:2502.13894*, 2025.

[55] W. Zheng, W. Chen, Y. Huang, B. Zhang, Y. Duan, and J. Lu, "Occworld: Learning a 3d occupancy world model for autonomous driving," in *European Conference on Computer Vision*, pp. 55–72, Springer, 2025.

[56] H. Zhou, H. Geng, X. Xue, Z. Yin, and L. Bai, "Reso: A reward-driven self-organizing llm-based multi-agent system for reasoning tasks," *arXiv preprint arXiv:2503.02390*, 2025.

[57] E. Zhou, Q. Su, C. Chi, Z. Zhang, Z. Wang, T. Huang, L. Sheng, and H. Wang, "Code-as-monitor: Constraint-aware visual programming for reactive and proactive robotic failure detection," *arXiv preprint arXiv:2412.04455*, 2024.

[58] J. Ma, Y. Qin, Y. Li, X. Liao, Y. Guo, and R. Zhang, "Cdp: Towards robust autoregressive visuomotor policy learning via causal diffusion," *arXiv preprint arXiv:2506.14769*, 2025.

[59] J. Liu, C. Xu, P. Hang, J. Sun, M. Ding, W. Zhan, and M. Tomizuka, "Language-driven policy distillation for cooperative driving in multi-agent reinforcement learning," *IEEE Robotics and Automation Letters*, 2025.

[60] K. Obata, T. Aoki, T. Horii, T. Taniguchi, and T. Nagai, "Lip-llm: Integrating linear programming and dependency graph with large language models for multi-robot task planning," *IEEE Robotics and Automation Letters*, 2024.

[61] Y. Wang, R. Xiao, J. Y. L. Kasahara, R. Yajima, K. Nagatani, A. Yamashita, and H. Asama, "Dart-llm: Dependency-aware multi-robot task decomposition and execution using large language models," *arXiv preprint arXiv:2411.09022*, 2024.

[62] W. Wang, I. Obi, and B.-C. Min, "Multi-agent llm actor-critic framework for social robot navigation," *arXiv preprint arXiv:2503.09758*, 2025.

[63] H. Zhang, W. Du, J. Shan, Q. Zhou, Y. Du, J. B. Tenenbaum, T. Shu, and C. Gan, "Building cooperative embodied agents modularly with large language models," *arXiv preprint arXiv:2307.02485*, 2023.

[64] L. Kang, X. Song, H. Zhou, Y. Qin, J. Yang, X. Liu, P. Torr, L. Bai, and Z. Yin, "Viki-r: Coordinating embodied multi-agent cooperation via reinforcement learning," *arXiv preprint arXiv:2506.09049*, 2025.

[65] X. Bo, Z. Zhang, Q. Dai, X. Feng, L. Wang, R. Li, X. Chen, and J.-R. Wen, "Reflective multi-agent collaboration based on large language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 138595–138631, 2024.

[66] X. Guo, K. Huang, J. Liu, W. Fan, N. Vélez, Q. Wu, H. Wang, T. L. Griffiths, and M. Wang, "Embodied llm agents learn to cooperate in organized teams," *arXiv preprint arXiv:2403.12482*, 2024.

[67] S. S. Kannan, V. L. Venkatesh, and B.-C. Min, "Smart-llm: Smart multi-agent robot task planning using large language models," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 12140–12147, IEEE, 2024.

[68] S. Nasiriany, A. Maddukuri, L. Zhang, A. Parikh, A. Lo, A. Joshi, A. Mandlekar, and Y. Zhu, "Robocasa: Large-scale simulation of everyday tasks for generalist robots," *arXiv preprint arXiv:2406.02523*, 2024.

[69] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language models can teach themselves to use tools," *Advances in Neural Information Processing Systems*, vol. 36, pp. 68539–68551, 2023.

[70] Z. Zhao, W. S. Lee, and D. Hsu, "Large language models as commonsense knowledge for large-scale task planning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 31967–31987, 2023.

[71] X. Zhou, H. Zhu, L. Mathur, R. Zhang, H. Yu, Z. Qi, L.-P. Morency, Y. Bisk, D. Fried, G. Neubig, *et al.*, "Sotopia: Interactive evaluation for social intelligence in language agents," *arXiv preprint arXiv:2310.11667*, 2023.

[72] A. Bar, G. Zhou, D. Tran, T. Darrell, and Y. LeCun, "Navigation world models," *arXiv preprint arXiv:2412.03572*, 2024.

[73] Y. Qin, Z. Shi, J. Yu, X. Wang, E. Zhou, L. Li, Z. Yin, X. Liu, L. Sheng, J. Shao, *et al.*, "Worldsimbench: Towards video generation models as world simulators," *arXiv preprint arXiv:2410.18072*, 2024.

[74] J. Yu, J. Bai, Y. Qin, Q. Liu, X. Wang, P. Wan, D. Zhang, and X. Liu, "Context as memory: Scene-consistent interactive long video generation with memory retrieval," *arXiv preprint arXiv:2506.03141*, 2025.

[75] J. Yu, Y. Qin, H. Che, Q. Liu, X. Wang, P. Wan, D. Zhang, K. Gai, H. Chen, and X. Liu, "A survey of interactive generative video," *arXiv preprint arXiv:2504.21853*, 2025.

[76] J. Yu, Y. Qin, H. Che, Q. Liu, X. Wang, P. Wan, D. Zhang, and X. Liu, "Position: Interactive generative video as next-generation game engine," *arXiv preprint arXiv:2503.17359*, 2025.

[77] J. Yu, Y. Qin, X. Wang, P. Wan, D. Zhang, and X. Liu, "Gamefactory: Creating new games with generative interactive videos," *arXiv preprint arXiv:2501.08325*, 2025.

[78] H. Zhang, Z. Wang, Q. Lyu, Z. Zhang, S. Chen, T. Shu, B. Dariush, K. Lee, Y. Du, and C. Gan, "Combo: compositional world models for embodied multi-agent cooperation," *arXiv preprint arXiv:2404.10775*, 2024.

[79] Y. Wang, Q. Liu, Z. Jiang, T. Wang, J. Jiao, H. Chu, B. Gao, and H. Chen, "Rad: Retrieval-augmented decision-making of meta-actions with vision-language models in autonomous driving," *arXiv preprint arXiv:2503.13861*, 2025.

[80] S. Zheng, J. Liu, Y. Feng, and Z. Lu, "Steve-eye: Equipping llm-based embodied agents with visual perception in open worlds," *arXiv preprint arXiv:2310.13255*, 2023.

[81] E. Zhou, J. An, C. Chi, Y. Han, S. Rong, C. Zhang, P. Wang, Z. Wang, T. Huang, L. Sheng, *et al.*, "Roborefer: Towards spatial referring with reasoning in vision-language models for robotics," *arXiv preprint arXiv:2506.04308*, 2025.

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

[83] B. Ye, S. Liu, H. Xu, X. Li, M. Pollefeys, M.-H. Yang, and S. Peng, "No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images," *arXiv preprint arXiv:2410.24207*, 2024.

[84] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[85] Y. Su, X. Zhan, H. Fang, H. Xue, H.-S. Fang, Y.-L. Li, C. Lu, and L. Yang, "Dense policy: Bidirectional autoregressive learning of actions," *arXiv preprint arXiv:2503.13217*, 2025.

A Task Description

We use 6 task in RoboFactory benchmark dataset. Table 6 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.

Table 6: 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.

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 7, 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 8 and 9. 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.

Table 7: 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%
Global GauDP	**72%**	**15%**	**2%**	**26%**	0%	**3%**	**19.67%**

Table 8: Parameter comparison across different policy variants.

Setting	DP	**GauDP-policy**	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

Table 9: 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%	**6%**	3%	0%	0%	5.00%
LargeDP	60%	12%	4%	**29%**	0%	0%	17.50%
GauDP	**72%**	**15%**	2%	26%	0	**3%**	**19.67%**

Table 10: Robustness evaluation under random lighting and distractor objects.

Model	DP	DP3	**GauDP**
Grab Roller	20%	46%	**50%**

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.

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 items placed near the target objects), following the RoboTwin 2.0 benchmark. As shown in Table 10, 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.