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

Ziye Wang[1,2†*] **Li Kang**[3†] **Yiran Qin**[1,4†] **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`.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

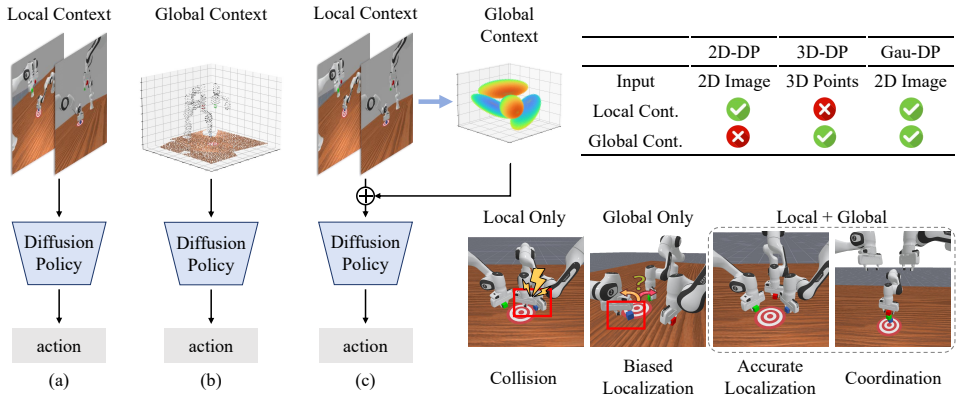

Figure 1: Both local and global context are essential in multi-agent collaboration. Comparison of multi-agent decision-making using different types of contextual information. (a) Using only local context leads to miscoordination such as collisions due to lack of global awareness. (b) Using only global 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.

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 [48] 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.

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.

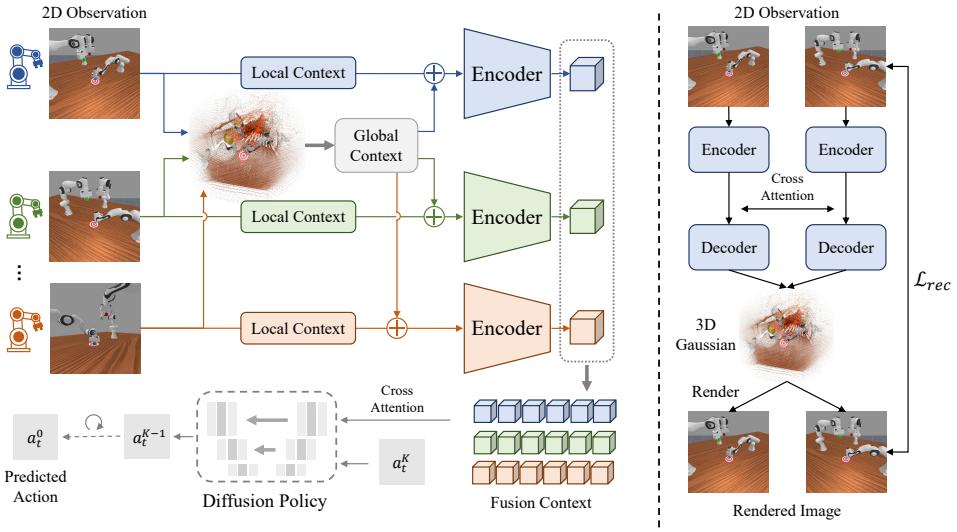

Figure 2: **(a)** Overview of the proposed ***GauDP*** framework for multi-agent imitation learning. Each agent extracts a local context from its 2D observation. A shared 3D Gaussian field is constructed from all views to form the global context, which is fused with the local context and passed through 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.

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 [49], 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 [50] 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 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.

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**

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 [48] and 2D Dense Policy [51]; for 3D input modalities, we include 3D Diffusion Policy [11] and 3D Dense Policy [51] 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.

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

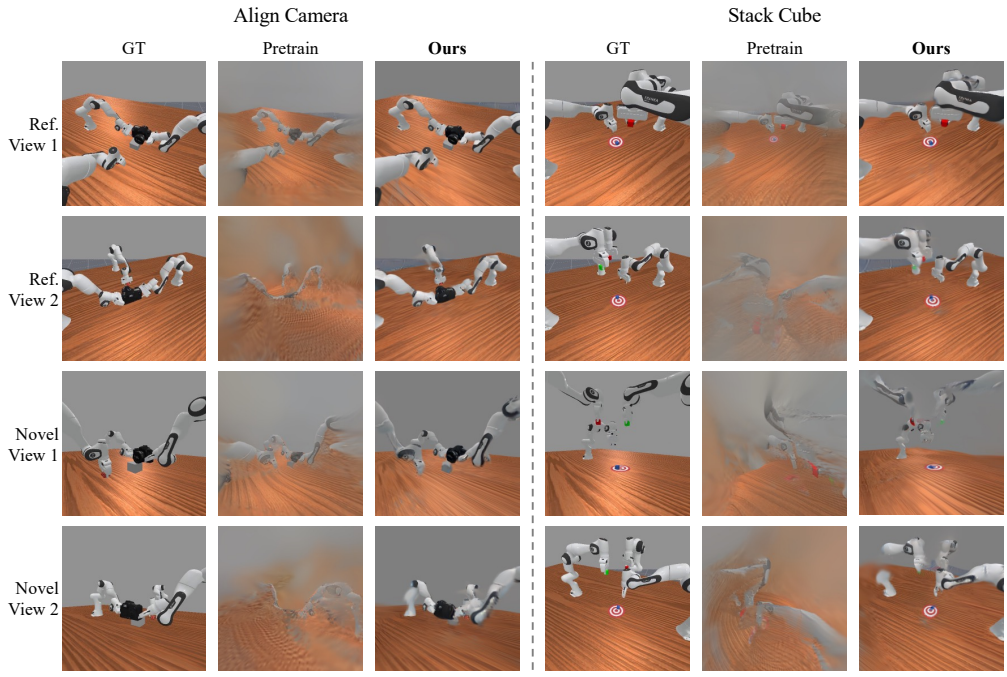

Figure 3: Visualization of Reconstruction Results. Our method achieves significantly improved reconstruction quality.

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 [51]	28%	18%	0%	0%	0%	7%	8.83%
DP [48]	9%	12%	**6%**	3%	0%	0%	5.00%
2D Dense Policy [51]	3%	2%	0%	0%	0%	**9%**	2.33%
GauDP	**72%**	**15%**	2%	**26%**	0%	3%	**19.67%**

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

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

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**

Policy[51], which struggle across the board (mostly below 10%), *GauDP* achieves a remarkable 72% 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 [51]. 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 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

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%**

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] 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.

[49] 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.

[50] 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.

[51] 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.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We give the limitation of our work in the supplementary material.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The paper provide the full set of assumptions and a complete (and correct) proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper fully disclose all the information needed to reproduce the main experimental results of the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: For some reason, we haven't provided the open access to code now. But we will release our source code once the paper is accepted.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper specify all the training and test details necessary to understand the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provide sufficient information on the computer resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discuss both potential positive societal impacts in supplementary materials.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We explicitly mentioned and properly respected the license and terms of assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.