2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

GP3: A 3D Geometry-Aware Policy with Multi-View Images for Robotic Manipulation

Quanhao Qian[1,2], Guoyang Zhao[1,3], Gongjie Zhang[1,2], Jiuniu Wang[1,2], Junlong Gao[1,2], Deli Zhao[1,2], Ran Xu[1,2†]

***Abstract*— Effective robotic manipulation relies on a precise understanding of 3D scene geometry, and one of the most straightforward ways to acquire such geometry is through multi-view observations. Motivated by this, we present GP3—a 3D geometry-aware robotic manipulation policy that leverages multi-view input. GP3 employs a spatial encoder to infer dense spatial features from RGB observations, which enable the estimation of depth and camera parameters, leading to a compact yet expressive 3D scene representation tailored for manipulation. This representation is fused with language instructions and translated into continuous actions via a lightweight policy head. We further introduce G-FiLM, which applies language-conditioned FiLM only to cross-view global attention. Comprehensive experiments demonstrate that GP3 consistently outperforms state-of-the-art methods on simulated benchmarks. Furthermore, GP3 transfers effectively to real-world robots in depth-challenging scenes with only minimal fine-tuning. These results highlight GP3 as a practical, sensor-agnostic solution for geometry-aware robotic manipulation.**

I. INTRODUCTION

Spatial perception is a core capability for general-purpose robotic motion, particularly in enabling robots to fully perceive and utilize 3D geometric information in the scene. A dominant line of research integrates 3D information into visuomotor policies by directly encoding point cloud data [1]–[14]. In practice, depth pipelines can still be unstable under occlusion, reflective/transparent surfaces, and IR noise. Conversely, approaches that generate implicit 3D representations from standard RGB images [15]–[21] often struggle with generalization, producing spatial representations that are not robust enough for diverse, unseen environments.

To bridge this gap, we introduce GP3, a geometry-aware policy for robotic manipulation that achieves robust multi-view spatial reasoning without requiring depth data. GP3 is built upon two key technical contributions: a robot-adapted spatial understanding module and an efficient attention mechanism for multi-view, language-guided control.

First, the core of GP3 is RoboVGGT, a spatial encoder adapted for robotic tasks. We begin with a large-scale pretrained 3D reconstruction model, VGGT [22], selected for its exceptional generalization capabilities across diverse scenes. We then finetune it on a newly curated, multi-domain robotics dataset, comprising simulated data from RLBench [23], MetaWorld [24], and RoboTwin [25], alongside real-world task data. This targeted finetuning strengthens RoboVGGT with generalizable spatial reasoning abilities grounded in multi-view geometry, enabling accurate 3D understanding across a wide spectrum of robotic scenes and tasks.

Second, as evidenced in Section IV-C, we observed that simply increasing the number of input views can paradoxically degrade performance by introducing distracting information and diluting the model's focus. To counteract this, we introduce G-FiLM (Global attention-based Feature-wise Linear Modulation). Unlike direct FiLM insertion [26], G-FiLM modulates only cross-view global attention to reduce inter-view redundancy with language guidance, improving task success.

Finally, our experimental results demonstrate that the proposed approach eliminates the need for explicit depth sensors while retaining robust 3D reconstruction capabilities for previously unseen robotic tasks. The method efficiently adapts to new environments using only minimal amounts of data, and consistently outperforms prior methods reliant on explicit 3D information [4], [7] as well as other approaches with RBG inputs [17], [27]–[30]. Under the same implementation settings, GP3 improves over the best baseline by 11.2% on MetaWorld, by 22.7% on RLBench, and by 57.5% in real-world experiments.

In summary, our contributions are threefold:

- We introduce **RoboVGGT**, a geometry-aware 3D reconstruction model fine-tuned on a curated dataset combining simulated and real-world data, achieving robust and generalizable 3D reconstruction across diverse robotic scenarios.
- We present **G-FiLM**, a global-attention-selective FiLM adaptation for multi-view geometry transformers, which injects language conditioning specifically into cross-view interaction layers to suppress redundant information and enhance task-relevant attention.
- With the above two key technical contributions, we propose **GP3**, a 3D geometry-aware policy for robotic manipulation that enables robust and efficient multi-view spatial reasoning. GP3 achieves state-of-the-art performance across multiple simulated and real-world benchmarks, setting a new standard for robust visuomotor control without relying on specific sensors.

[1]Alibaba DAMO Academy, Hangzhou, China.
[2]HuPan Lab.
[3]Tongji University, Shanghai, China.
† Corresponding author.

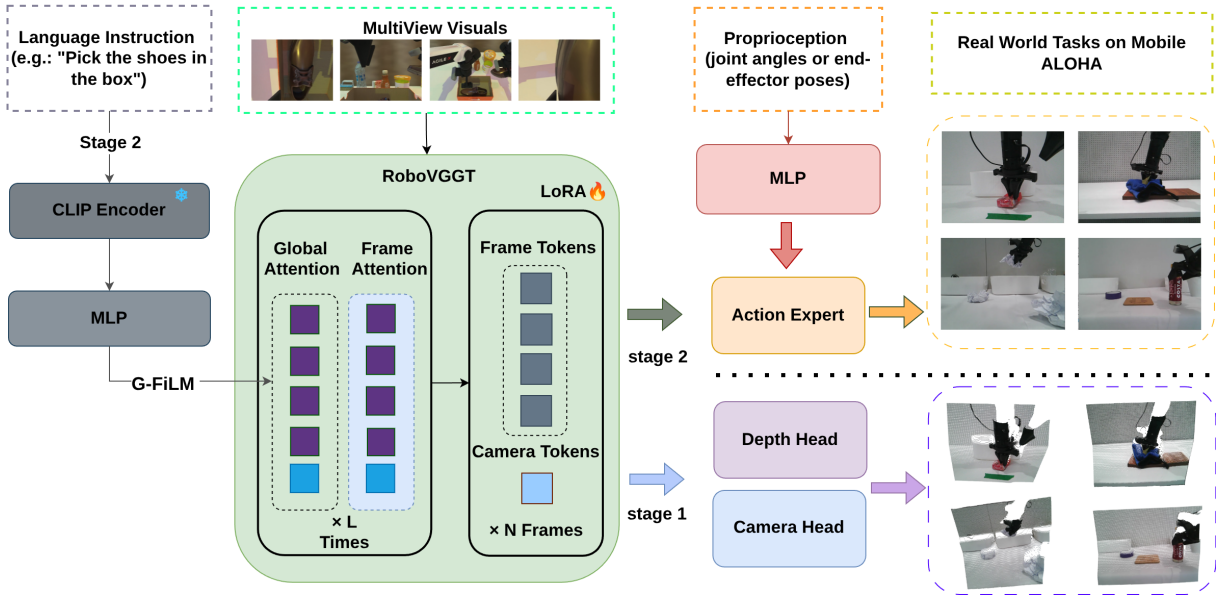


Fig. 1: Architecture overview. Our framework adopts a two-stage training pipeline. Stage 1 (*geometry model fine-tuning*): RoboVGGT is fine-tuned by equipping it with a multi-view camera parameter head and a depth estimation head. Stage 2 (*action prediction training*): We incorporate global attention-based feature-wise linear modulation (G-FiLM), enabling the encoder to focus on task-relevant regions of interest (ROIs) and suppress task-irrelevant noise through adaptive feature modulation. The extracted spatial features are then combined with proprioceptive information to predict the action.

II. RELATED WORK

A. Robotic Manipulation

Traditional robotic manipulation has largely relied on state-based reinforcement learning [31]–[33], assuming full observability of low-dimensional state representations (e.g., joint angles, end-effector poses). Although effective in simulation and controlled environments, these methods are challenging to deploy in real-world scenarios due to the difficulty of accurate state estimation and the high cost of specialized sensing hardware [34].

Recent advances [30], [35]–[38] have shifted towards learning control policies directly from visual observations, often leveraging large-scale datasets and modern architectures such as Vision Transformers (ViTs) [39]. For example, R3M [29] employs contrastive learning to obtain universal embodied representations from diverse human video data, while VC-1 [28] evaluates the effectiveness of masked auto-encoding (MAE) [40] strategies for robotic pretraining. Vision-Language-Action (VLA) models [41]–[43] further unify perception, language grounding, and action prediction within a single framework, improving task generalization but still facing challenges in the fine-grained spatial reasoning required for precision tasks such as assembly or insertion [44].

This limitation has motivated growing interest in methods that explicitly incorporate spatial and geometric features into robotic perception and control.

B. Spatial Geometry-Based Methods

To address the limitations of 2D-based methods in capturing precise spatial relationships, recent research has explored integrating 3D geometric information into robotic perception and control. Such approaches leverage depth, point clouds, or multi-view geometry to enhance spatial reasoning and improve manipulation accuracy, particularly in tasks requiring fine-grained positioning and contact reasoning. 3D-aware robotic perception research can be broadly categorized into two complementary paradigms—explicit and implicit geometry representations.

Explicit Geometry-Based Approaches. Recent advancements in robotic perception have predominantly relied on explicit spatial geometry representations, including point clouds, voxels, and Gaussian splats for 3D spatial reasoning. Representative frameworks such as 3DP [4], 3DA [3], Lift3D [7], and RVT [45], [46] demonstrate the effectiveness of point cloud-based approaches, while voxel-based methods like PerACT [12], [47], and VoxAct [11] further expand this paradigm. However, these methods require stable depth acquisition and substantial computation, and can degrade under occlusion and sensor noise in real robotic environments. To address these drawbacks, recent works, including our proposed framework, explore RGB-only alternatives that retain geometric reasoning capabilities without relying on explicit depth.

Implicit Geometry-Based Approaches. Inspired by multi-modal alignment approaches [27], [48], recent works such as [49], [50] attempt to address depth dependency through depth-augmented image features. While these models can predict depth from visual input, single-frame depth estimation remains inherently ambiguous due to missing scale information and absolute depth resolution without ad-

ditional constraints. To overcome this limitation, approaches like RoboHorizon [18] and MV-MWM [19] employ multi-view MAE to learn implicit 3D information of a robot workspace. The most relevant work to ours is SPA [17], which enhances ViT with 3D awareness through differentiable neural rendering on multi-view images. However, the volumetric features produced by SPA predominantly support coarse-grained spatial reasoning, which is insufficient for precision-critical robotic tasks. In addition, these previous methods lack generalization capability and fail to understand spatial relationships in unseen manipulation tasks or scenarios. In this paper, we propose a pixel-wise spatial reasoning framework that enables fine-grained perception crucial for complex robotic applications by transferring the pretrained geometry transformer to robotic tasks.

C. Transformer-Based 3D Vision Methods

Classical scene reconstruction methods [51]–[54] leverage multi-view geometry to recover 3D environments. A newer trend centers on transformer-based foundation models [21], [22], [55]–[59], which reconstruct dense scenes from raw RGB, even when camera parameters are unknown. DUSt3R [60] and MASt3R [56] pioneered joint geometry and camera estimation, while CUT3R [57] and Fast3R [58] improved reconstruction speed. VGGT [22] currently achieves state-of-the-art accuracy and stability in local 3D mapping.

Our model, RoboVGGT, builds upon VGGT's pretrained capacity for geometry extraction and adapts it to robotic manipulation, where performance in object-centric, small-workspace environments is critical. Since the generalization ability of transformer-based 3D models is highly sensitive to dataset diversity, we collect targeted robotic datasets and fine-tune RoboVGGT with both geometry-specific and task-specific objectives. Furthermore, we integrate language conditioning into the pipeline, enabling the extraction of high-quality, 3D-aware features specialized for robotic manipulation.

III. METHODOLOGY

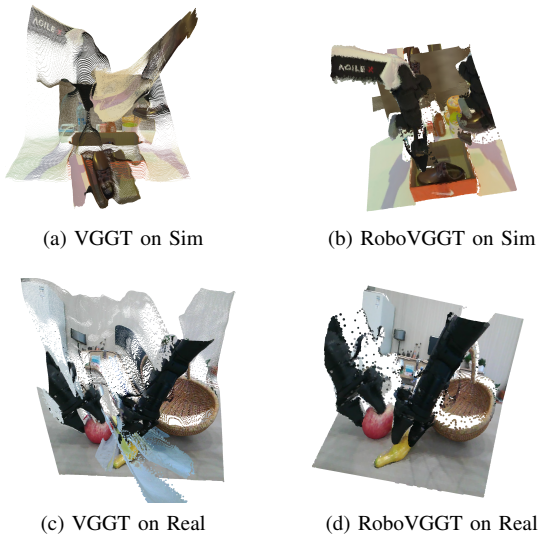
We aim to model 3D visual geometry for robotic manipulation tasks without relying on explicit dense 3D inputs. As illustrated in Figure 1, we introduce a spatial encoder that extracts rich spatial features from multiple RGB views and passes them to the action expert through a two-stage training.

Section III-A defines the problem, Section III-B describes the training of the spatial encoder, Section III-C introduces the fusion of language and spatial features by G-FiLM, and Section III-D provides a detailed explanation of the action training process.

A. Problem definition and notation

The inputs consist of a sequence of RGB images I_i captured from either onboard or third-person cameras, the robot's proprioceptive state P, and a language instruction L. The objective is to predict the corresponding sequence of robot actions A.

B. Geometry-Aware Spatial Encoder



(a) VGGT on Sim (b) RoboVGGT on Sim

(c) VGGT on Real (d) RoboVGGT on Real

Fig. 2: The reconstruction results compared the original VGGT model against our fine-tuned RoboVGGT model, along with simulations and real-world inputs. Our targeted fine-tuned spatial encoder shows clear improvement on robot manipulation scenes.

To capture 3D geometric information in multi-view inputs, we employ a geometry-aware spatial encoder, thus avoiding the dependency on point clouds as in [7]. We build the RoboVGGT based on VGGT [22] for its state-of-the-art 3D performance; unlike methods that require explicit 3D priors, VGGT inherently learns inter-frame correspondences and reconstructs scenes from multi-view RGB images.

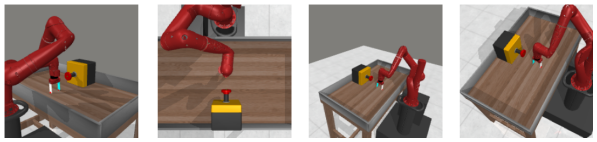
VGGT's original design for high-resolution inputs leads to slow training/inference and high memory costs. To mitigate this computational bottleneck and bridge the domain gap between its pretraining data and robotic manipulation, we fine-tune the model at a practical 224×224 resolution. Our training leverages 150K simulated frames from RL-Bench [23], MetaWorld [24] and RoboTwin [25], along with 20K real-world frames collected in robotic scenes, and incorporates explicit geometric supervision to improve adaptation.

We follow the VGGT framework to train the spatial encoder using a multi-task camera and depth loss:

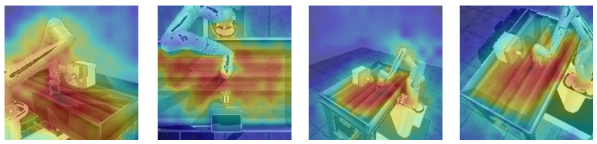
$$\mathcal{L} = \mathcal{L}_{\text{camera}} + \mathcal{L}_{\text{depth}}. \tag{1}$$

As shown in Figure 2, we input multi-view images captured by the front camera and wrist cameras to the spatial encoder. The original model lacks the ability to generalize to robotic scenes, exhibiting particularly poor reconstruction quality for the robotic arm. After our fine-tuning process, the model demonstrates a robust capability to reconstruct scenes involving robotic manipulation. Notably, the robotic task in the depicted real-world scene was unseen by the model during training, further demonstrating the generalization ability of our fine-tuned model to robotic scenes.

C. Global Attention-based Feature-wise Linear Modulation



(a) Observations



(b) G-FiLM's Attention Responses

Fig. 3: Our proposed G-FiLM effectively guides attention to the visual contents pertinent to the button-press task.

While the spatial encoder can extract implicit spatial information from multi-frame inputs, we observe that increasing the number of views does not always improve task success rates and may even degrade performance. This is primarily due to the redundant and task-irrelevant information introduced in multi-view settings.

The OpenVLA-oft [61] framework addresses this by applying FiLM [26] modules to the self-attention components of each vision transformer block.

In our work, the transformer architecture integrates both global and frame-wise attention layers, with global attention designed to capture cross-perspective associations, while frame-wise attention focuses primarily on local image feature extraction. G-FiLM differs from a direct FiLM plug-in by modulating only global attention layers for cross-view interaction, preserving local geometric encoding in frame-wise blocks. Our formulation is defined as:

$$\text{G-FiLM}(F) = \gamma \odot (M * F) + \beta \tag{2}$$

where γ and β are scaling and shifting vectors projected from language embeddings that modulate the visual features F through an affine transformation, and $\odot$ denotes element-wise multiplication. M is a mask where each element M_j corresponds to an attention layer j.

$$M_j = \begin{cases} \text{True} & \text{if } j \text{ is a global attention layer} \\ \text{False} & \text{otherwise} \end{cases} \tag{3}$$

The novelty of G-FiLM lies in this structure-aware integration for multi-view geometry transformers. The corresponding results in Table III show consistent gains over baseline methods and conventional FiLM insertion. Figure 3 displays the visualization results of task focus, obtained by normalizing global attention tokens, showing that G-FiLM helps the model focus on areas pertinent to the button-press task.

D. Action Training

For each robot, the action representation is selected to match its control interface: we adopt the end-effector pose for simulated tasks and the joint-space configuration for the real-world ALOHA system. The action loss is defined as

$$\mathcal{L}_{\text{action}} = \sum MSE(A^{gt} - A) \tag{4}$$

Specifically, we leverage LoRA [62] for efficient fine-tuning of the RoboVGGT across all training stages including action training.

IV. EXPERIMENTS

Sections IV-A and IV-B evaluate the manipulation capabilities of the proposed GP3 framework through experiments conducted in both simulated and real-world settings. Section IV-C presents an ablation study to assess the contribution of each component, and Section IV-D evaluates the accuracy of point map estimation on unseen tasks.

A. Simulation Experiment

1) Benchmarks: We evaluate GP3 on diverse tasks drawn from two widely used robotic manipulation simulation benchmarks: MetaWorld [24], running in the MuJoCo simulator, and RLBench [23], running in the CoppeliaSim simulator. In MetaWorld, which features a tabletop environment with a Sawyer robotic arm and a two-finger gripper, we select 50 tasks of varying difficulty levels. In RLBench, which uses a Franka Panda robot equipped with multi-view cameras, we evaluate on the same 6 tasks as [7].

2) Data collection: In MetaWorld, we generate 25 expert demonstrations per task using scripted policies. Each trajectory terminates immediately upon task success. Unlike [4], which records a fixed 200-step horizon, we collect images from three corner cameras and one overhead camera. For RLBench, demonstration trajectories are generated via predefined waypoints using the Open Motion Planning Library [63], resulting in 100 training demonstrations per task. We collect images from the front camera, overhead camera, and wrist camera.

3) Baselines: To assess GP3's contribution, we benchmark it against 6 representative methods, grouped in 3 categories: **(1) 2D Robotic Representation Methods**: 2D foundation model CLIP (ViT-B/16) [27] and the 2D pretrained policy backbones R3M [29] and VC-1 [28]. **(2) Explicit 3D Policies**: DP3 [4] and pretrained 3D policy Lift3D [7]. **(3) Implicit 3D Representation Methods**: SPA [17], the prior SOTA of implicit 3D pretrained representation. We adopt the three-layer MLP policy head for MetaWorld and the diffusion-policy head [30] for RLBench. For fairness, we keep the same training loss as GP3 for all models. For implicit 3D methods, we assess single-view (SPA-S and GP3-S) and multi-view configurations to investigate whether incorporating multiple viewpoints improves the quality of the learned geometric representations.

4) Training and Evaluation Details: For a fair comparison, all baselines follow a unified training and evaluation protocol, differing only in their visual modalities: 3D policy baselines additionally process 1024-point clouds generated with cropping and downsampling parameters as specified in their respective original papers [4], whereas other methods

receive 224×224 RGB images. The robot state consists of the end-effector pose and joint angles, concatenated with the visual features. For visual feature extraction, each baseline employs its original encoder, whereas our method uses RoboVGGT to obtain implicit spatial features. We optimize using Adam $(\beta_1, \beta_2) = (0.95, 0.999)$ with a cosine annealing scheduler. The initial learning rate is set to 5×10^{-4}; both schedules include a linear warm-up over the first 10% of training. Each method is trained for 100 epochs, and all training and evaluation runs are conducted on H20 GPUs.

5) Quantitative Results: As shown in Table I, across both MetaWorld and RLBench, GP3 establishes new state-of-the-art performance while relying solely on RGB inputs. It consistently surpasses prior 3D representation [17] and 3D policy methods [7], demonstrating that our implicit geometry learned from RGBs is highly effective. Notably, unlike SPA, whose performance degrades on RLBench as the number of views increases, GP3 maintains a steady improvement with additional views. Furthermore, multi-view training endows GP3 with the ability to extract 3D geometry from a single view, leading to superior results even under single-view settings.

B. Real-World Experiment

We performed a quantitative evaluation of our method in a real-world physical environment, built upon the open-source Mobile ALOHA platform [64]. The experimental setup consists of two 7-DoF manipulator arms and three calibrated Intel RealSense D435i cameras. Two cameras are rigidly mounted on the robot wrists to capture actuator-centric views, while the third is fixed in a forward-facing position to provide a global scene perspective. Although D435i provides depth streams, our policy uses RGB only because depth is less stable in close-range manipulation. This multi-view RGB configuration supplies comprehensive visual information to the policy, which is essential for tasks requiring high precision and fine hand–eye coordination. We use a diffusion policy head for ALOHA, and the training details are the same as in simulation. We compare our method with [17], [30].

1) Dataset Collection: We collected a teleoperated dataset for four common household-style tasks: Task1: Catching a desktop paper ball and throwing it away. Task2: Grasping a cloth and cleaning a table. Task3: Clearing items on a desk. Task4: Retrieving a coffee cup from the desk. For each task, we recorded 100 successful demonstration trajectories. Visual data from all three cameras was captured at 30 FPS, while the robot's proprioceptive state was recorded at the same FPS. To maintain synchronization with the visual inputs and improve efficiency during training, we downsampled the high-frequency action trajectories. Our policy model is trained to directly predict the robot's joint angles as its action output.

2) Evaluation Protocol: After training, the policy was deployed on the physical ALOHA robot for evaluation. While simulators can offer scalable and reproducible evaluation, real-world testing remains the definitive measure of performance. For each of the four tasks, we performed 20 independent trials and recorded the number of successful completions. Finally, we calculated the average success rate for each task to quantitatively assess the real-world performance and robustness of our method.

3) Quantitative Results: As shown in Table II, GP3 with multi-view input achieves state-of-the-art performance on all evaluated real-world tasks. Unlike other methods that benefit little or even perform worse in multi-perspective settings, our method shows significant improvement as the number of cameras increases. Latency is reported in Section IV-D.

C. Ablation Study

To systematically assess the contributions of each core component, the fine-tuning of the spatial encoder for robot scene reconstruction, and the language guidance mechanism within our framework, we conducted an ablation study on 15 tasks from the MetaWorld benchmark. These tasks are categorized as follows: easy tasks include button-press, drawer-open, reach, handle-pull, peg-unplug-side, lever-pull, and dial-turn; medium tasks include hammer, sweep-into, bin-picking, push-wall, and box-close; hard tasks include assembly, hand-insert, and shelf-place. Quantitative results are summarized in Table III, and the key findings are as follows:

1) Impact of Fine-tuning: The baseline model (without fine-tuning) achieves a 56.0% mean success rate with single-view inputs. After fine-tuning on robot-centric data, the single-view success rate improves to 70.9%, highlighting the importance of adaptation to robot-specific scenarios. Similarly, the 2-view and 4-view configurations show increases of 12.3% and 15.5%, respectively, after fine-tuning.

2) Multi-view Inputs: Compared to its single-view baseline (56.0%), the model's performance improves to 66.1% with two views. However, increasing to four views unexpectedly reduces performance to 61.8%. Even after fine-tuning on robot-centric datasets (w/ FT), the success rate declines from 78.4% (two views) to 77.3% (four views). This suggests that once the number of views exceeds an optimal threshold, standard geometric models without specific view-handling mechanisms may introduce redundancy and noise, which in turn leads to degraded performance.

3) Conventional FiLM-based Feature Modulation: The conventional FiLM implementation (w/ FT+F), when combined with the fine-tuned encoder, shows a positive impact, generally improving upon the w/ FT baseline: With 4 views, w/ FT+F achieves an 80.3% mean success rate, a 3.0% improvement over w/ FT (4 views, 77.3%), and 1.3% improvement over w/ FT+F (2 views, 79.0%). This indicates that FiLM, by incorporating linguistic features, helps aggregate and modulate features from different views.

4) G-FiLM Architecture: With 4 views, w/ FT+GF leads with an 83.9% mean success rate, representing a 3.6% improvement over w/ FT+F (4 views, 80.3%), and a 2.3% improvement over w/ FT+GF (2 views, 81.6%). The multi-view improvement surpasses that of using conventional FiLM,

TABLE I: Comparison of manipulation success rates on MetaWorld and RLBench benchmarks. Methods are grouped by feature and input type.

Method	Feature Type	Input Type	MetaWorld					RLBench
			Easy	Medium	Hard	Very Hard	Mean S.R.	Mean S.R.
CLIP [27]	2D Rep.	Single-view	84.7	47.6	44.6	46.4	67.9	49.3
R3M [29]	2D Rep.	Single-view	80.4	49.5	55.3	38.4	66.4	66.0
VC-1 [28]	2D Rep.	Single-view	84.5	49.1	55.3	38.4	68.6	54.7
DP3 [4]	3D Rep.	Point Cloud	79.3	34.2	44.0	29.6	60.1	57.3
Lift3D [7]	3D Rep.	Point Cloud	84.8	56.0	60.0	28.0	69.8	53.3
SPA-S [17]	3D Rep.	Single-view	86.0	53.8	48.0	46.4	70.4	60.0
SPA	3D Rep.	Multi-view	91.0	62.2	58.6	58.4	77.5	52.0
GP3-S (Ours)	3D Rep.	Single-view	88.3	57.1	75.5	46.4	75.7	70.0
GP3 (Ours)	3D Rep.	Multi-view	**95.7**	**72.4**	**82.0**	**69.2**	**86.7**	**78.7**

TABLE II: Comparison of manipulation success rates on the real-world experiments.

Method	View	Task1	Task2	Task3	Task4
Diffusion Policy-S [30]	1	8/20	5/20	3/20	1/20
Diffusion Policy	3	5/20	7/20	1/20	0/20
SPA-S [17]	1	3/20	6/20	3/20	3/20
SPA	3	7/20	7/20	5/20	4/20
GP3-S (Ours)	1	14/20	13/20	11/20	13/20
GP3 (Ours)	3	**18/20**	**19/20**	**15/20**	**17/20**

demonstrating that G-FiLM better handles noise and redundancy in multi-view inputs compared to FiLM. This validates the advantage of selectively fusing geometric and linguistic features, enabling the incorporation of semantic cues into the multi-view fusion module while preserving the geometric model's inherent capability for scene feature extraction. Such a design is particularly critical for robotic manipulation in complex environments with multi-view observations.

TABLE III: Ablation study on a 15-task MetaWorld subset. FT means using fine-tuned encoder on robot dataset, F means using FiLM, and GF means using G-FiLM.

Method	Views	Easy	Medium	Hard	Mean S.R.
Baseline	1	63.4	55.2	40.0	56.0
	2	70.2	69.6	50.7	66.1
	4	74.3	57.6	40.0	61.8
w/ FT	1	80.0	64.8	60.0	70.9
	2	83.4	74.4	73.4	78.4
	4	82.3	76.0	68.0	77.3
w/ FT+F	1	79.4	75.2	57.3	73.6
	2	85.0	73.6	72.0	79.0
	4	86.3	74.4	76.0	80.3
w/ FT+GF	1	78.8	80.8	61.3	76.0
	2	81.7	83.2	**78.6**	81.6
	4	**86.9**	**85.7**	69.3	**83.9**

D. Point Map Estimation

We further compare the accuracy of our predicted point clouds with that of the original VGGT on two unseen RoboTwin tasks, whose scenes are significantly different from those in the training dataset. We generate 20 demonstrations per task and randomly sample 10 frames for each demonstration. Predicted point clouds are aligned to the ground truth using the Umeyama [65] algorithm. Following [66], we report *Accuracy*, *Completeness*, and *Overall* (Chamfer distance) for point map estimation. To ensure the reliability of the evaluation, we filter out views with an overlap of less than 0.35 with the main view during validation. This threshold is chosen to exclude low-quality or highly dissimilar views that could negatively affect the accuracy of the comparison. As shown in Table IV, fine-tuning improves the geometry model's performance in robotic scenes.

TABLE IV: Dense MVS Estimation on unseen tasks in RoboTwin. Lower values are better (↓).

Method	Acc.↓	Comp.↓	Overall↓
Original VGGT	4.756	0.923	2.840
RoboVGGT	1.142	0.472	0.807

For practical deployment, RoboVGGT runs at 152 ms, 158 ms, and 162 ms for 1, 2, and 4 views on an RTX 4090, and the small 10 ms increase from 1 to 4 views highlights efficient shared-weight encoding and G-FiLM modulation that preserve real-time responsiveness for multi-view spatial reasoning and high-frequency robot control.

V. CONCLUSION

We present GP3, a sensor-agnostic 3D geometry-aware policy framework that learns visuomotor control from multi-view RGB by fine-tuning a large-scale 3D reconstruction model, enabling robust spatial reasoning without depth sensors or explicit 3D supervision. GP3 achieves state-of-the-art performance on MetaWorld and RLBench, transfers to real robots with minimal overhead, and suggests a lightweight, scalable path to RGB-only manipulation, while future work will target long-horizon dynamic tasks and more complex skills.

REFERENCES

[1] S. Chen, R. Garcia, I. Laptev, and C. Schmid, "Sugar: Pre-training 3d visual representations for robotics," in *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, 2024, pp. 18 049–18 060.

[2] A. Wilcox, M. Ghanem, M. Moghani, P. Barroso, B. Joffe, and A. Garg, "Adapt3r: Adaptive 3d scene representation for domain transfer in imitation learning," 2025. [Online]. Available: https://arxiv.org/abs/2503.04877

[3] T.-W. Ke, N. Gkanatsios, and K. Fragkiadaki, "3d diffuser actor: Policy diffusion with 3d scene representations," in *ICRA 2024 Workshop on 3D Visual Representations for Robot Manipulation*, 2024. [Online]. Available: https://openreview.net/forum?id=VWxztuB1rJ

[4] 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 *ICRA 2024 Workshop on 3D Visual Representations for Robot Manipulation*.

[5] R. Yang, G. Chen, C. Wen, and Y. Gao, "Fp3: A 3d foundation policy for robotic manipulation," *ArXiv*, vol. abs/2503.08950, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:276937820

[6] H. Zhu, Y. Wang, D. Huang, W. Ye, W. Ouyang, and T. He, "Point cloud matters: Rethinking the impact of different observation spaces on robot learning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 77 799–77 830, 2024.

[7] Y. Jia, J. Liu, S. Chen, C. Gu, Z. Wang, L. Luo, L. Lee, P. Wang, Z. Wang, R. Zhang *et al.*, "Lift3d foundation policy: Lifting 2d large-scale pretrained models for robust 3d robotic manipulation," *arXiv preprint arXiv:2411.18623*, 2024.

[8] H. Huang, K. Schmeckpeper, D. Wang, O. Biza, Y. Qian, H. Liu, M. Jia, R. Platt, and R. Walters, "Imagination policy: Using generative point cloud models for learning manipulation policies," *arXiv preprint arXiv:2406.11740*, 2024.

[9] S. Haldar and L. Pinto, "Point policy: Unifying observations and actions with key points for robot manipulation," *arXiv preprint arXiv:2502.20391*, 2025.

[10] B. Eisner, H. Zhang, and D. Held, "Flowbot3d: Learning 3d articulation flow to manipulate articulated objects," *CoRR*, vol. abs/2205.04382, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2205.04382

[11] I.-C. A. Liu, S. He, D. Seita, and G. Sukhatme, "Voxact-b: Voxel-based acting and stabilizing policy for bimanual manipulation," 2024. [Online]. Available: https://arxiv.org/abs/2407.04152

[12] M. Grotz, M. Shridhar, Y.-W. Chao, T. Asfour, and D. Fox, "Peract2: Benchmarking and learning for robotic bimanual manipulation tasks," in *CoRL 2024 Workshop on Whole-body Control and Bimanual Manipulation: Applications in Humanoids and Beyond*, 2024.

[13] X. Miao, H. Duan, Q. Qian, J. Wang, Y. Long, L. Shao, D. Zhao, R. Xu, and G. Zhang, "Towards scalable spatial intelligence via 2D-to-3D data lifting," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[14] Y. Ravan, Z. Yang, T. Chen, T. Lozano-Pérez, and L. P. Kaelbling, "Combining planning and diffusion for mobility with unknown dynamics," 2024. [Online]. Available: https://arxiv.org/abs/2410.06911

[15] T. Gervet, Z. Xian, N. Gkanatsios, and K. Fragkiadaki, "Act3d: 3d feature field transformers for multi-task robotic manipulation," in *7th Annual Conference on Robot Learning*, 2023. [Online]. Available: https://openreview.net/forum?id=-HFJuX1uqs

[16] Z. Xian and N. Gkanatsios, "Chaineddiffuser: Unifying trajectory diffusion and keypose prediction for robotic manipulation," in *Conference on Robot Learning/Proceedings of Machine Learning Research*. Proceedings of Machine Learning Research, 2023.

[17] H. Zhu, H. Yang, Y. Wang, J. Yang, L. Wang, and T. He, "SPA: 3d spatial-awareness enables effective embodied representation," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=6TLdqAZgzn

[18] Z. Chen, J. Huo, Y. Chen, and Y. Gao, "Robohorizon: An llm-assisted multi-view world model for long-horizon robotic manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2501.06605

[19] Y. Seo, J. Kim, S. James, K. Lee, J. Shin, and P. Abbeel, "Multi-view masked world models for visual robotic manipulation," in *International Conference on Machine Learning*. PMLR, 2023, pp. 30 613–30 632.

[20] S. Qian, K. Mo, V. Blukis, D. F. Fouhey, D. Fox, and A. Goyal, "3d-mvp: 3d multiview pretraining for robotic manipulation," *arXiv preprint arXiv:2406.18158*, 2024.

[21] G. Zhang, J. Lin, S. Wu, Z. Luo, Y. Xue, S. Lu, Z. Wang *et al.*, "Online map vectorization for autonomous driving: A rasterization perspective," *Advances in Neural Information Processing Systems*, vol. 36, pp. 31 865–31 877, 2023.

[22] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "Vggt: Visual geometry grounded transformer," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5294–5306.

[23] 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.

[24] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine, "Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning," in *Conference on robot learning*. PMLR, 2020, pp. 1094–1100.

[25] Y. Mu, T. Chen, Z. Chen, S. Peng, Z. Lan, Z. Gao, Z. Liang, Q. Yu, Y. Zou, M. Xu *et al.*, "Robotwin: Dual-arm robot benchmark with generative digital twins," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 27 649–27 660.

[26] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

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

[28] A. Majumdar, K. Yadav, S. Arnaud, J. Ma, C. Chen, S. Silwal, A. Jain, V.-P. Berges, T. Wu, J. Vakil *et al.*, "Where are we in the search for an artificial visual cortex for embodied intelligence?" *Advances in Neural Information Processing Systems*, vol. 36, pp. 655–677, 2023.

[29] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3m: A universal visual representation for robot manipulation," in *6th Annual Conference on Robot Learning*, 2022. [Online]. Available: https://openreview.net/forum?id=tGbpgz6yOrI

[30] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Robotics: Science and Systems*, 2023. [Online]. Available: https://doi.org/10.15607/RSS.2023.XIX.026

[31] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," 2015.

[32] Y. Li, "Deep reinforcement learning: An overview," *arXiv preprint arXiv:1701.07274*, 2017.

[33] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[34] S. Gu, E. Holly, T. Lillicrap, and S. Levine, "Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 3389–3396.

[35] 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*. PMLR, 2022, pp. 991–1002.

[36] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[37] J. Shang, K. Schmeckpeper, B. B. May, M. V. Minniti, T. Kelestemur, D. Watkins, and L. Herlant, "Theia: Distilling diverse vision foundation models for robot learning," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=ylZHvlwUcI

[38] V. Mazzia, S. Angarano, F. Salvetti, F. Angelini, and M. Chiaberge, "Action transformer: A self-attention model for short-time pose-based human action recognition," *Pattern Recognition*, vol. 124, p. 108487, 2022.

[39] M. Shridhar, L. Manuelli, and D. Fox, "Cliport: What and where pathways for robotic manipulation," in *Conference on robot learning*. PMLR, 2022, pp. 894–906.

[40] K. He, X. Chen, S. Xie, Y. Li, P. Doll'ar, and R. B. Girshick, "Masked autoencoders are scalable vision learners," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15 979–15 988, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:243985980

[41] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. Sanketi, Q. Vuong *et al.*, "Openvla: An

open-source vision-language-action model," in *8th Annual Conference on Robot Learning*.

[42] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning*. PMLR, 2023, pp. 2165–2183.

[43] J. Liu, M. Liu, Z. Wang, P. An, X. Li, K. Zhou, S. Yang, R. Zhang, Y. Guo, and S. Zhang, "Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation," *Advances in Neural Information Processing Systems*, vol. 37, pp. 40 085–40 110, 2024.

[44] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3929–3945, 2023.

[45] A. Goyal, J. Xu, Y. Guo, V. Blukis, Y.-W. Chao, and D. Fox, "Rvt: Robotic view transformer for 3d object manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 694–710.

[46] A. Goyal, V. Blukis, J. Xu, Y. Guo, Y.-W. Chao, and D. Fox, "Rvt-2: Learning precise manipulation from few demonstrations," *arXiv preprint arXiv:2406.08545*, 2024.

[47] M. Shridhar, L. Manuelli, and D. Fox, "Perceiver-actor: A multi-task transformer for robotic manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 785–799.

[48] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 11 975–11 986.

[49] X. Pang, W. Xia, Z. Wang, B. Zhao, D. Hu, D. Wang, and X. Li, "Depth helps: Improving pre-trained rgb-based policy with depth information injection," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 7251–7256.

[50] J. Li, W. Wang, Y. Peng, C. Shen, Y. Zhu, and Z. Xu, "Visual robotic manipulation with depth-aware pretraining," in *2024 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2024, pp. 843–850.

[51] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd, "3d reconstruction of complex structures with bundle adjustment: an incremental approach," in *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, 2006, pp. 3055–3061.

[52] D. Bradley, T. Boubekeur, and W. Heidrich, "Accurate multi-view reconstruction using robust binocular stereo and surface meshing," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[53] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[54] L. Pan, D. Baráth, M. Pollefeys, and J. L. Schönberger, "Global structure-from-motion revisited," in *European Conference on Computer Vision*. Springer, 2024, pp. 58–77.

[55] H. Wang and L. Agapito, "3d reconstruction with spatial memory," in *International Conference on 3D Vision 2025*, 2025.

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

[57] Q. Wang, Y. Zhang, A. Holynski, A. A. Efros, and A. Kanazawa, "Continuous 3d perception model with persistent state," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 10 510–10 522.

[58] J. Yang, A. Sax, K. J. Liang, M. Henaff, H. Tang, A. Cao, J. Chai, F. Meier, and M. Feiszli, "Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 21 924–21 935.

[59] B. Smart, C. Zheng, I. Laina, and V. A. Prisacariu, "Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs," *ArXiv*, vol. abs/2408.13912, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:271957263

[60] 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*, 2024, pp. 20 697–20 709.

[61] M. J. Kim, C. Finn, and P. Liang, "Fine-tuning vision-language-action models: Optimizing speed and success," *arXiv preprint arXiv:2502.19645*, 2025.

[62] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[63] I. A. Sucan, M. Moll, and L. E. Kavraki, "The open motion planning library," *IEEE Robotics Automation Magazine*, vol. 19, no. 4, pp. 72–82, 2012.

[64] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation using low-cost whole-body teleoperation," in *8th Annual Conference on Robot Learning*, 2024.

[65] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 4, pp. 376–380, 1991.

[66] J. Wang, C. Rupprecht, and D. Novotny, "Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 9739–9749.