GAP: Geometric Anchor Pre-training for Data-Efficient Visuomotor Learning of Manipulation Tasks

Davide Buoso, Andrea Protopapa, Stefano Di Carlo, Francesca Pistilli, Giuseppe Averta
Politecnico di Torino

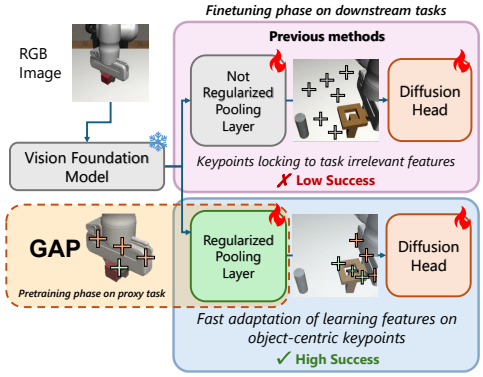
***Abstract*— Learning visuomotor policies from scarce expert demonstrations remains a core challenge in robotic manipulation. A primary hurdle lies in distilling high-dimensional representations into control-relevant geometry without overfitting. While using frozen pretrained Vision Foundation Models (VFMs) improves data efficiency, it also shifts most task adaptation onto a small spatial pooling module, which can latch onto task-irrelevant shortcuts and lose geometric grounding when finetuned with few data samples. We propose Geometric Anchor Pre-training (GAP), a simple, action-free warm-up stage that regularizes the spatial adapter *before* downstream imitation learning. GAP pre-trains the pooling layer on a lightweight simulated proxy task where object masks are available at no cost, encouraging the adapter to produce keypoints that cover its spatial extent (instead of collapsing). This yields stable *geometric anchors* that provide a reliable coordinate interface for few-shot policy learning, while keeping the VFM frozen. We evaluate GAP on RoboMimic and ManiSkill under severe data scarcity (15–50 demonstrations) and domain shift. A simple adapter regularized with GAP consistently outperforms stronger attention-based poolers and end-to-end fine-tuning, achieving 62% success on RoboMimic `Can` (+16% over baselines), 63% on `Tool Hang`, and 61% on ManiSkill `StackCube` (+11% over full fine-tuning). The proxy stage is lightweight (about 40 minutes on a single consumer GPU) and fully decoupled from downstream tasks, making it practical to reuse across environments.**

I. Introduction

Vision-based Imitation Learning (IL) for robotic manipulation is a highly studied setting where Diffusion Policies [1] emerged as powerful architectures. However, learning from scarce expert demonstrations remains a core challenge in robotic manipulation. To improve data efficiency, a common recipe is freezing a pretrained Vision Foundation Model (VFM) and training a lightweight spatial bottleneck. Yet, in the low-data regime, this spatial bottleneck is one of the main point of failure. The pooling module can lock onto visual shortcuts instead of learning stable, object-centric geometry– a failure mode we call *bottleneck collapse*. To address this, we propose **Geometric Anchor Pre-training (GAP)**, an action-free warm-up stage that regularizes the spatial adapter *before* downstream policy learning (Figure 1). GAP pretrains the bottleneck on a cheap simulated proxy task using free object masks, encouraging keypoints to lie on the object and cover its spatial extent. This produces stable *geometric anchors* that serve as a reliable coordinate interface for few-shot downstream policy learning (15–50 demonstrations). In summary, our main contributions are:

- **GAP**: A lightweight, mask-supervised proxy pretraining strategy that mitigates bottleneck collapse.

Fig. 1: **Geometric Anchor Pretraining.** GAP is a pretraining strategy applied to the spatial pooling layer on a cheap proxy task, which consistently outperforms other pooling techniques when data is scarce.



- We empirically prove that GAP learns domain-agnostic geometric anchors that improve robustness across tasks and backbones.
- We prove that GAP learns also view-agnostic priors by testing its cross-domain and simulator ability.

II. Methodology

Geometric Anchor Pretraining (GAP) addresses spatial overfitting in data-scarce regimes ($N \leq 50$) by explicitly supervising a coordinate-based adapter via a masked proxy task.

A. Spatial Adapter & Proxy Task

To extract robust features, a frozen pretrained backbone f_ϕ is followed by a lightweight adapter f_A, which applies a Spatial Softmax to produce K 2D keypoints $P_t = \{p_{k,t}\}_{k=1}^{K}$. Without explicit supervision, these keypoints latch onto spurious visual cues. GAP pretrains f_A on a cheap simulated proxy task (e.g., `LiftCube` from [2]), leveraging ground-truth object masks $\mathcal{M}_t$ without expert action labels. We supervise f_A using a multi-objective spatial loss to enforce object-centric, distributed, and non-redundant keypoints:

Centroid Alignment ($\mathcal{L}_{center}$): Pulls the predicted keypoint centroid $\bar{p}_t$ to the ground-truth mask centroid c_t to ground keypoints on the target object

$$\mathcal{L}_{center} = \|\bar{p}_t - c_t\|_2^2$$

Geometric Spread ($\mathcal{L}_{spread}$): Matches the spatial variance of keypoints σ_p to the normalized object scale σ_{target} to prevent collapse:

$$\mathcal{L}_{spread} = \|\sigma_p - \sigma_{target}\|_2^2$$

TABLE I: **Multi-Task Evaluation Results.** For all tasks, we pre-train on *LiftCube* from Robomimic. Results on the ManiSkill simulator environment are shaded in gray to denote the domain shift. GAP achieves state-of-the-art average performance. For GAP the best performing VFM is VC1 with ViT-B while for AFA is VC1 for *Can* and R3M for the other tasks.

	Robomimic: Can					Robomimic: Square				
Method	**15**	**20**	**30**	**50**	**Avg**	**15**	**20**	**30**	**50**	**Avg**
E-E (Full FT)	0.55 (0.04)	0.76 (0.02)	0.88 (0.03)	0.95 (0.01)	0.785	0.15 (0.03)	0.19 (0.02)	0.29 (0.03)	0.38 (0.02)	0.253
R3M + SS	0.50 (0.02)	0.75 (0.05)	0.78 (0.04)	0.86 (0.02)	0.723	0.12 (0.02)	0.13 (0.04)	0.17 (0.04)	0.22 (0.02)	0.158
DINOv2 + SS	0.51 (0.08)	0.68 (0.03)	0.73 (0.03)	0.86 (0.00)	0.695	0.10 (0.00)	0.22 (0.02)	0.23 (0.04)	0.26 (0.03)	0.203
VC-1 + SS	0.49 (0.06)	0.64 (0.10)	0.82 (0.05)	0.89 (0.02)	0.710	0.07 (0.06)	0.24 (0.06)	0.23 (0.03)	0.32 (0.05)	0.215
AFA	0.46 (0.01)	0.74 (0.02)	0.78 (0.02)	0.93 (0.05)	0.728	0.15 (0.02)	0.19 (0.03)	0.32 (0.04)	0.43 (0.04)	0.273
GAP (Ours)	**0.62** (0.06)	**0.80** (0.02)	**0.94** (0.04)	**0.96** (0.02)	**0.830**	**0.20** (0.03)	**0.33** (0.03)	**0.37** (0.01)	**0.53** (0.03)	**0.358**
	Robomimic: Tool Hang					**ManiSkill: StackCube**				
Method	**15**	**20**	**30**	**50**	**Avg**	**15**	**20**	**30**	**50**	**Avg**
E-E (Full FT)	0.06 (0.05)	0.2 (0.1)	0.13 (0.06)	0.33 (0.06)	0.18	**0.22** (0.10)	0.23 (0.05)	0.50 (0.08)	0.66 (0.08)	0.403
R3M + SS	0.16 (0.06)	0.17 (0.06)	0.27 (0.06)	0.50 (0.10)	0.275	0.06 (0.01)	0.09 (0.02)	0.15 (0.05)	0.38 (0.09)	0.171
DINOv2 + SS	0.23 (0.06)	0.13 (0.06)	0.20 (0.17)	0.23 (0.23)	0.198	0.07 (0.01)	0.10 (0.03)	0.15 (0.00)	0.60 (0.05)	0.230
VC-1 + SS	0.20 (0.05)	0.17 (0.11)	0.20 (0.10)	0.43 (0.05)	0.250	0.04 (0.01)	0.08 (0.05)	0.28 (0.02)	0.63 (0.03)	0.258
AFA	0.20 (0.10)	0.23 (0.10)	0.30 (0.06)	0.45 (0.10)	0.295	0.09 (0.02)	0.14 (0.03)	0.25 (0.02)	0.44 (0.08)	0.230
GAP (Ours)	**0.27** (0.06)	**0.33** (0.06)	**0.37** (0.06)	**0.63** (0.05)	**0.400**	0.20 (0.03)	**0.24** (0.04)	**0.61** (0.10)	**0.80** (0.02)	**0.463**

Keypoint Diversity ($\mathcal{L}_{div}$): Penalizes redundancy by enforcing a minimum separation margin δ_{min} between any pair of keypoints:

$$\mathcal{L}_{div} = \frac{1}{K}\sum_{k=1}^{K}\left[\max\left(0, \delta_{min} - \min_{j\neq k}\|p_{k,t} - p_{j,t}\|_2\right)\right]^2$$

The final objective is $\mathcal{L}_{GAP} = \lambda_c\mathcal{L}_{center} + \lambda_s\mathcal{L}_{spread} + \lambda_d\mathcal{L}_{div}$, creating a "push-pull" dynamic. We partition the K keypoints into M disjoint subsets corresponding to semantic entities. The spatial losses are applied independently to each subset using its mask, allowing keypoints to deploy as independent semantic trackers that rapidly re-anchor to novel geometries even during partial occlusions or active reorientation.

Following GAP phase, the regularized adapter and frozen encoder are transferred to downstream. We fine-tune only f_A alongside the diffusion head via the standard action-prediction objective $\mathcal{L}_{diff}$.

III. Experiments

We evaluate GAP on four tasks across RoboMimic [3] and ManiSkill3 [2] (`Can`, `Square`, `ToolHang`, `StackCube`) under severe data scarcity (15–50 demonstrations). We compare against end-to-end full fine-tuning (E-E), standard Spatial Softmax (SS)-based pooling [4], and Attentive Feature Aggregation (AFA) [5] using R3M [6], DINOv2 [7], and VC-1 [8] backbones. All GAP models are pretrained on the same mask-supervised `LiftCube` proxy task (≈40 min on one GPU). We mainly measure the success rate of the final policies using different spatial adapters and pretraining strategies. Comprehensive metrics for all tasks are provided in Table I. In our evaluations, GAP consistently demonstrates strong performance across various tasks, outperforming both AFA and E-E baselines on standard environments like `Can` and `Square`. Most notably, in the highly complex `Tool Hang` task (50 demos), GAP achieves a success rate of 0.63, whereas E-E largely fails to learn the task efficiently.

Finally, to assess domain robustness, we evaluated the models on `StackCube` from another simulator [2]. Under severe data scarcity (15 demos), AFA achieves a success rate of only 0.09, whereas GAP more than doubles this performance to reach 0.20. This proves that GAP learns simulator-agnostic general priors rather than relying on environmental artifacts, achieving superior performance across all demonstration counts.

Ablations To thoroughly evaluate the contributing factors to our method's success, we performed extensive ablation studies focusing on backbone compatibility, loss formulation, pretraining strategies, and sim-to-real transferability of the priors. Overall, the results demonstrate that GAP serves as a highly effective and universal regularizer that consistently improves downstream performance across different vision backbones, avoiding the degradation seen with standard semantic pooling. Furthermore, we prove that all the loss terms are necessary to prevent representation collapse. We also found that GAP's purely geometric pretraining objective successfully adapts to downstream tasks, heavily outperforming standard action-supervised pretraining approaches. Finally, deploying the GAP-pretrained adapter on real-world robot videos reveals strong sim-to-real potential, evidenced by robust, object-centric keypoint initialization in zero-shot settings.

Fig. 2: **VC-1 backbone with GAP in the wild.** GAP provides robust zero-shot keypoint initialization on real-world video [9].

IV. Conclusion

This paper identifies *bottleneck collapse* as a key failure mode of frozen-VFM pipelines, particularly when confronted with the viewpoint shifts inherent to task transfer. To prevent the spatial pooling layer from overfitting to static visual shortcuts, we introduce **Geometric Anchor Pre-training (GAP)**. By applying a geometric objective ($\mathcal{L}_{GAP}$) on a cheap proxy task, GAP produces viewpoint-invariant geometric anchors that consistently outperform baselines across tasks and visual domains. While GAP is not yet validated on highly deformable objects, its explicit spatial structure opens direct pathways for active perception. A key future extension is utilizing real-time collapses in GAP's geometric spread as an intrinsic uncertainty metric to trigger purposeful sensing actions—such as sweeping a wrist camera to disambiguate occluded geometry.

REFERENCES

[1] 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*, vol. 44, no. 10-11, pp. 1684–1704, 2025.

[2] S. Tao, F. Xiang, A. Shukla, Y. Qin, X. Hinrichsen, X. Yuan, C. Bao, X. Lin, Y. Liu, T.-k. Chan, *et al.*, "Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai," *arXiv preprint arXiv:2410.00425*, 2024.

[3] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, F.-F. Li, S. Savarese, Y. Zhu, and R. Martín-Martín, "What matters in learning from offline human demonstrations for robot manipulation," in *Conference on Robot Learning*. PMLR, 2021, pp. 1678–1690.

[4] C. Finn, X. Y. Tan, Y. Duan, T. Darrell, S. Levine, and P. Abbeel, "Deep spatial autoencoders for visuomotor learning," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 512–519.

[5] N. Tsagkas, A. Sochopoulos, D. Danier, S. Vijayakumar, A. Kouris, O. Mac Aodha, and C. X. Lu, "Attentive feature aggregation or: How policies learn to stop worrying about robustness and attend to task-relevant visual cues," *arXiv preprint arXiv:2511.10762*, 2025.

[6] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3M: A universal visual representation for robot manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 892–909.

[7] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, *et al.*, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2024.

[8] A. Majumdar, K. Yadav, S. Arnaud, J. Ma, V. Chen, S. Silwal, A. Jain, V.-P. Berges, T. Wu, J. Vakil, *et al.*, "Where are we in the search for an effective robot motor control foundation model?" in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[9] Y. Fang, Y. Yang, X. Zhu, K. Zheng, G. Bertasius, D. Szafir, and M. Ding, "Rebot: Scaling robot learning with real-to-sim-to-real robotic video synthesis," *arXiv preprint arXiv:2503.14526*, 2025.