When Would Vision-Proprioception Policy Fail in Robotic Manipulation?

Anonymous Author(s)
Affiliation
Address
email

Abstract

Proprioceptive information is critical for precise servo control by providing real-time robotic states. Its collaboration with vision is highly expected to enhance performances of the manipulation policy in complex tasks. However, recent studies have reported *confused observations* that vision-proprioception policies frequently suffer from poor generalization. In this work, we attempt to answer the question: **When would vision-proprioception policy fail?** To this end, we conducted temporally controlled experiments and found that during task sub-phases that robot's motion transitions, which require target localization, the vision modality of the vision-proprioception policy fails to take effect. Further analysis reveals that the policy naturally gravitates toward concise proprioceptive signals that offer faster loss reduction when training, thereby dominating the optimization and suppressing the learning of the visual modality during motion-transition phases. To alleviate this, we propose the Gradient Adjustment with Phase-guidance (GAP) algorithm that adaptively modulates the optimization of proprioception, enabling dynamic collaboration within vision-proprioception policy. Specifically, we leverage proprioception to capture robotic states and estimate the probability of each timestep in the trajectory belonging to motion-transition phases. During policy learning, we apply fine-grained adjustment that reduces the magnitude of proprioception's gradient based on estimated probabilities, leading to improved generalization of vision-proprioception policies. The comprehensive experiments demonstrate GAP is applicable in both simulated and real-world environments, across one-arm and dual-arm setups, and compatible with both conventional and Vision-Language-Action models. We believe this work can offer valuable insights into the development of vision-proprioception policies for robotic manipulation.

1 Introduction

Proprioceptive information has long been recognized as a cornerstone of low-level robotic control, enabling smooth motor behavior through immediate access to the robot's internal state. This capability is especially critical in tasks requiring high accuracy and fast correction, such as posture control Allum et al. (1998); Henze et al. (2014) and locomotion Bjelonic et al. (2016); Lee et al. (2020); Yang et al. (2023). In recent years, there has been growing interest in introducing proprioception to learning-based manipulation Levine et al. (2016); Cong et al. (2022); Jiang et al. (2025). Despite the expectations that its inclusion will empower manipulation policies to maintain precision and robustness across various scenarios, existing works have reported *confused observations*: HPT Wang et al. (2024) demonstrated clear improvements under the joint utilization of vision and proprioception, while Octo Octo Model Team et al. (2024) observed policies trained with additional propioception seemed generally worse than vision-only policies. This discrepancy exposes a critical obstacle to understanding: **when vision-proprioception policy would fail in robotic manipulation?**

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

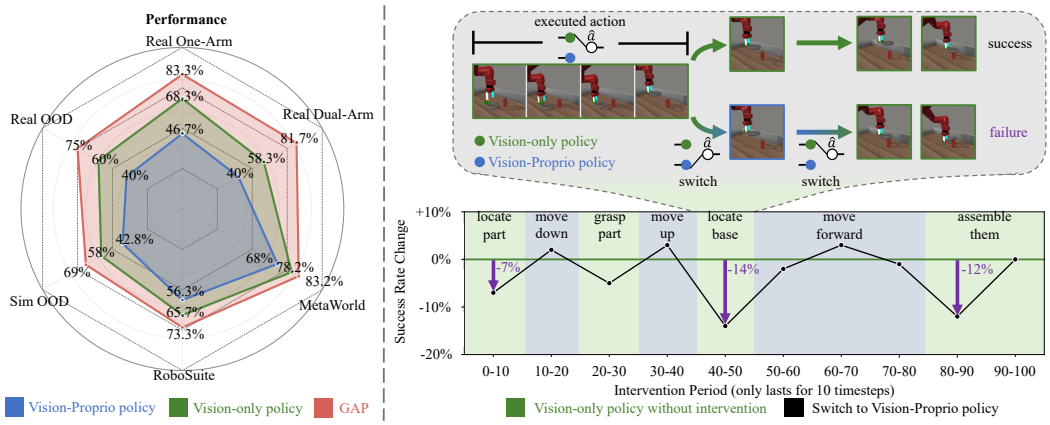

Figure 1: When would vision-proprioception policy fail? (left) Vision-Proprioception policies perform 15.8% worse than Vision-only policies. (right) We explore this through intervening the task execution of vision-only policy during different periods, by switching to vision-proprioception policy. Such intervention has minimal impact during motion-consistent phases like "move forward". However, during motion-transition phases like "locate base" and "assemble them", switching leads to noticeable degradation, indicating the vision modality fails to take effect during these phases.

Extensive prior studies have revealed that the importance of visual and proprioceptive information could change over time within manipulation Sarlegna and Sainburg (2009); Feng et al. (2024); He et al. (2025), which referred to as *Modality Temporality*. For example, during motion-consistent phases where the robot performs ongoing movements, the policy can benefit more from proprioceptive signals. In contrast, during the transition intervals where the robot's motion shifts, it is required to rely more on visual cues for accurate target localization. To verify whether the vision-proprioception policy exhibits such collaboration, we conduct an intervention experiment in the controlled simulation environment. Concretely, we execute the "assembly" task using the vision-only policy, but for a specific 10-timestep period, we replace executed actions with those predicted by the vision-proprioception policy under the same observations. As shown in Figure 1 (right), the intervention brings minimal impact during motion-consistent phases like "move forward", during motion-transition phases like "locate base" and "assemble them", the switching leads to noticeable degradation. It suggests that the vision modality of the vision-proprioception policy fails to take effect during motion-transition phases.

We further investigate the underlying cause from an optimization perspective. During motion-transition phases, visual cues tend to be subtle and may only differ at the pixel level Tsagkas et al. (2025). As a result, the vision-proprioception policy naturally gravitates toward the more concise proprioceptive signals to minimize the training loss, thereby dominating the optimization Huang et al. (2022); Fan et al. (2023). This dominance suppresses the learning of the vision modality and ultimately leads to under-utilized visual information during motion-transition phases.

To alleviate this, we propose the Gradient Adjustment with Phase-guidance (GAP) algorithm that adaptively modulates the optimization of proprioception, enabling dynamic collaboration between vision and proprioception. Specifically, we first define the motion of the robot using the concise proprioception signals and segment the trajectory into motion-consistent phases. Motion of the robot transits within the intervals between these phases, we thus employ an temporal network like LSTM to model transition processes and estimate the probability that each timestep belongs to motion-transition phases. During policy learning, we guide the vision-proprioception policy to focus on essential visual cues of motion-transition phases, by applying fine-grained gradient adjustment that reduces the magnitude of proprioception's gradient based on estimated probabilities.

Our GAP algorithm facilitates the vision-proprioception policy to effectively utilize proprioception without suppressing the learning of visual modality. GAP is compatible with both conventional and Vision-Language-Action models, and its versatility and effectiveness have been validated by extensive experiments in both simulated and real-world environments. The evaluations cover a wide range of manipulation tasks and includes one-arm and dual-arm robotic setups. We believe this work can offer valuable insights into the development of vision-proprioception policies for robotic manipulation.

2 Related Work

Vision-Proprioception Policy in Manipulation. Vision has been the most commonly used modality in robotic manipulation policies Zitkovich et al. (2023); Kim et al. (2024); Zeng et al. (2024). While it provides sufficient information to complete many manipulation tasks, visual data often includes a large amount of noise, such as irrelevant background distractions Tsagkas et al. (2025). Therefore, more concise proprioceptive information has been introduced by many works to assist robotic manipulation policy, with the expectation that it can provide complementary and physically grounded information for precise and robust task execution Cong et al. (2022); Mandlekar et al. (2022); Chi et al. (2023); Fu et al. (2024); Wang et al. (2024); Liu et al. (2024). However, existing studies have reported confused observations: some works demonstrate clear improvements when integrating proprioceptive information with vision Cong et al. (2022); Wang et al. (2024), others observe limited gains or even detrimental effects Mandlekar et al. (2022); Octo Model Team et al. (2024). Fu et al. (2024) attributes this to overfitting while Octo Model Team et al. (2024) suggests it arises from causal confusion between the proprioceptive information and the target actions. In this study, we further explore when vision-proprioception policy would fail and introduce a modality-temporality perspective to offer valuable insights into the development of vision-proprioception policies for robotic manipulation.

Modality Temporality. In manipulation tasks, each modality's contribution to decision-making can vary significantly over time. For example, in "pick-place" task, policy must first rely on vision to locate the target object. When moving toward the object, proprioception becomes more critical for executing consistent and precise actions. It is proven by strong correlations between variations in modality data and task stages Lee et al. (2019); He et al. (2025); Jiang et al. (2025). Feng et al. (2024) summarizes such property of manipulation tasks as modality temporality. Given this nature of robotic manipulation tasks, recent works have proposed approaches based on dynamic fusion Li et al. (2023); Feng et al. (2024); He et al. (2025) and modality selection Jiang et al. (2025) to improve the performance of multimodal manipulation policies. In this study, we introduce the modality-temporality perspective to understand the roles of vision and propriocetion and propose the gradient adjustment algorithm to enhance dynamic collaboration within the vision-proprioception policy.

3 When Would Vision-Proprioception Policy Fail?

In this section, we first formalize the problem and further analyze when vision-proprioception policy would fail from an optimization perspective. The vision-proprioception policy is learned under the Behavior Cloning (BC) paradigm, which can be formulated as the Markov Decision Process (MDP) framework Torabi et al. (2018). Formally, the policy π takes the environment observation $o_t \in O$ as input at each timestep t. In this work, o_t includes RGB-sensor readings v_t, and for vision-proprioception policy π_{v+s}, it includes robot proprioceptive information s_t additionally. This proprioceptive information consists of the 6D pose of robot's gripper $(p_t^x, p_t^y, p_t^z, \theta_t^x, \theta_t^y, \theta_t^z) \in \mathbb{R}^6$ in Cartesian space and orientation, and a continuous value $g_t \in [0, 1]$ representing the degree of gripper opening, with $g_t = 1$ denoting fully open and $g_t = 0$ denoting fully closed.

The policy π maps the observation history to a sequence of actions: $\hat{a}_{t+L} = \pi(o_{t-H:t})$, where L and H indicate the length of predicted action sequence and observation history respectively. For simplicity, we set omit them in the following discussion. The training objective can be formulated as:

$$\pi^* = \text{argmin}_\pi \mathbb{E}_{(o_t, a_t) \sim \tau_e} [\mathcal{L}_{BC}(\pi(o_t)), a_t], \tag{1}$$

where τ_e is expert demonstration dataset and a_t is action labels. In vanilla BC, $\mathcal{L}$ typically represents the Mean Squared Error (MSE) loss for continuous action spaces, or Cross-Entropy (CE) loss for discrete action spaces. We focus solely on the vanilla MSE loss here.

In this work, we adopt standard joint-learning architecture to design the vision-proprioception policy, which extracts features from both vision and proprioception modalities using two separate chunks ϕ_v, ϕ_s. These features from two modalities are then concatenated and fed into the policy head ψ. Although some recent works have tried exploring alternative modality fusion approaches Wang et al. (2024); Feng et al. (2024), concatenation remains the most widely used approach Levine et al. (2016); Cong et al. (2022); Mandlekar et al. (2022). To support our analyze under this fusion approach, we split the first layer of MLP-based policy head ψ into ψ_s, ψ_v and rewrite the action prediction as:

$$\hat{a} = (\psi_s(f_s) + \psi_v(f_v)) \cdot W_{share} + b, \tag{2}$$

where f_s, f_v is the feature extracted by $\phi_s(o), \phi_v(o)$ respectively. Under Gradient Descent (GD)-based policy learning, the optimization of the vision chunk's parameters ω_v is influenced by:

$$\frac{\partial \mathcal{L}_{BC}}{\partial \omega_v} = \frac{\partial ||\hat{a} - a||_2^2}{\partial \hat{a}} \cdot \frac{\partial (\psi_s(f_s) + \psi_v(f_v)) \cdot W_{share} + b)}{\partial f_v} \cdot \frac{\partial f_v}{\partial \omega_v}. \tag{3}$$

Within the execution trajectory of the task, changes in visual cues are usually subtle compared to proprioceptive signals. For example, when the gripper is closing, visual cues differ only at pixel-level Tsagkas et al. (2025), while concise and low-dimension proprioceptive signals directly represent this process via changes in opening degree g. As a result, the vision-proprioception policy naturally gravitates toward proprioceptive signals to minimize the training loss. It leads to optimization dominated by proprioception and suppresses the learning of ω_v due to vision modality's low contribution to action prediction Huang et al. (2022); Fan et al. (2023).

As shown in Figure 1 (right), such overreliance to proprioception brings negligible impact during motion-consistent phases, since the execution of ongoing movements benefits significantly from proprioceptive signals. However, the initial positions of the target objects vary during testing and the proprioceptive signal does not contain object-related information. During motion-transition phases, the policy is required to accurately locate the target objects. The suppressed learning of vision modality thus regretfully impairs generalization of the vision-proprioception policy.

4 Method

To alleviate the suppression of the learning of vision modality during motion-transition phases, we propose the Gradient Adjustment with Phase-guidance (GAP) algorithm. As shown in Figure 2, we initially define the representation of robot's motion and identify motion-consistent phases. Motion-transition phase indicators are then predicted to estimate the probability that each timestep belongs to motion-transition phases. Based on these indicators, we apply fine-grained gradient adjustment during policy learning, facilitating dynamic collaboration within the vision-proprioception policy.

4.1 Motion Representation of Robot

Proprioceptive signals of the trajectory $[s_1, s_2, ..., s_N]$ directly provide the state of the gripper's position p, orientation θ, and opening degree g. The variations in them effectively capture the motion of the robot arm over time. We first define the representation of motion for further motion-transition phase estimation. Specifically, the motion between timestep i and timestep j is defined as: $m_{i:j} = \{p_{i:j}, \theta_{i:j}, g_{i:j}\}$, where $p_{i:j} = p_j - p_i$ denotes the change in the gripper's 3D position, $\theta_{i:j} = \theta_j - \theta_i$ denotes the change in orientation, and $g_{i:j} = g_j - g_i$ denotes the change in gripper opening. Together, these three dimensions provide a complete representation of the robot's motion.

4.2 Motion-Transition Phase Estimation

The represented motion captures the movement of robot arm, allowing expert demonstrations to be segmented into sequences of continuous states that correspond to semantically similar motions. To leverage this property for identifying motion-consistent phases, we employ the simple yet effective Change Point Detection (CPD) algorithm Liu et al. (2013); Aminikhanghahi and Cook (2017). The overall motion of a trajectory phase $\tau_{t_1:t_2}$ can be characterized by $m_{t_1:t_2}$. Based on whether the directions of these changes are consistent, we define the following distance between phase motion $m_{t_1:t_2}$ and adjacent motion $m_{i:i+1}$:

$$d(m_{t_1:t_2}, m_{i:i+1}) = -\cos(p_{t_1:t_2}, p_{i:i+1}) - \alpha\cos(\theta_{t_1:t_2}, \theta_{i:i+1}) - \beta(\text{sgn}(g_{t_1:t_2}) == \text{sgn}(g_{i:i+1})), \tag{4}$$

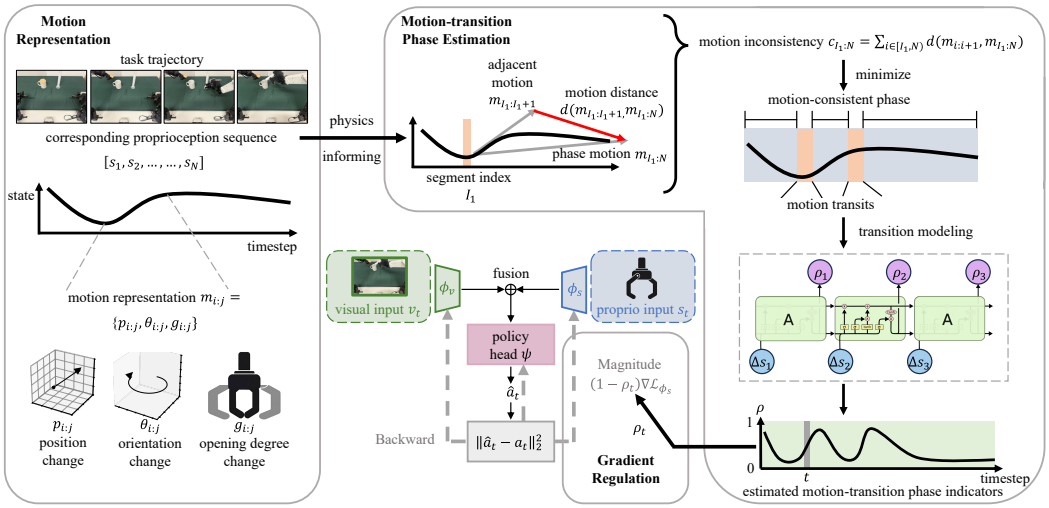

Figure 2: The pipeline of our Gradient Adjustment with Phase-guidance (GAP) algorithm. We define the motion representation and identify the motion-consistent phases by minimizing the total cost between phase motion and each adjacent motion. Motion-transition phase indicators are then estimated to reduce the magnitude of proprioception's backward gradient. GAP facilitates vision-proprioception policies to effectively utilize proprioception without suppressing vision modality.

where $\mathrm{sgn}(\cdot)$ denotes the sign function, and α, β are weighting factors for the orientation and opening degree component respectively. The statistic that measure the motion inconsistency of phase $\tau_{t_1:t_2}$ is defined as $c_{t_1:t_2} = \sum_{i=t_1}^{t_2-1} d(m_{t_1:t_2}, m_{i:i+1})$. The Change Point Detection algorithm leverages dynamic programming to identify a set of indices I that minimize the total cost $\sum_I c_{\tau_I}$, segmenting the trajectory into motion-consistent phases.

Motion of the robot transits within the intervals between these phases, requiring the policy to locate target object. Vision is therefore expected to play a more significant role. To model motion transitions, we further utilize the temporal differences of proprioceptive information $\Delta s_i = s_{i+1} - s_i$ and leverage their sequential context with an temporal network such as LSTM. It predicts motion-transition phase indicators ρ_i to estimate the probability that timestep i belongs to motion-transition phases. The predicted indicators ρ is under the supervision of indices set I. Additionally, for timesteps within a range near the transition, we reduce the penalty applied to them in order to better capture the inherently continuous and gradual transition process.

4.3 Gradient adjustment for Modality Collaboration

The vision-proprioception policy extracts features from both vision and proprioception modalities using two separate chunks ϕ_v, ϕ_s, which consist of an encoder and a temporal transformer, these features are then fused and fed into policy head to predict the action. However, since visual cues during motion-transition phases may be subtle, the policy tends to rely heavily on features of proprioception. As a result, the gradient optimization for corresponding samples becomes dominated by proprioceptive inputs, which in turn constrains the learning of the vision modality chunk ϕ_v.

To mitigate this, we employ gradient adjustment to control the optimization of proprioceptive chunk ϕ_s during motion-transition phases, thereby guiding the vision-proprioception policy to focus more on visual cues and preventing the degradation of its generalization. Concretely, in the j-th epoch of Gradient Descent (GD)-based optimization, the parameters of the proprioceptive feature chunk ω_s^j are updated according to the following formula:

$$\omega_s^{j+1} = \omega_s^j - \lambda \cdot (1-\rho) \cdot \eta \nabla \omega_s^j \mathcal{L}_{BC}(\omega_s^j), \tag{5}$$

where η is the learning rate, λ is a hyper-parameter that controls the degree of adjustment. For each timestep, we modulate the magnitude of the proprioception backward gradient based on its indicator ρ of belonging to motion-transition phases. The higher value of ρ leads to greater degree of modulation.

By applying gradient adjustment with phase-guidance as illustrated in Algorithm 1, the vision-proprioception policy is enabled to effectively leverage proprioceptive information without compromising its generalization ability.

Algorithm 1 Vision-Proprioception Policy Learning with Gradient Adjustment

Notations: Expert demonstrations o_e, proprioceptive signals s_e, epoch number T, vision-proprioception policy π_{v+s}, proprioception chunk parameters ω_s, vision chunk parameters ω_v.

Motion-Transition Phase Estimation
Identify motion-consistent phases by Change Point Detection $I \leftarrow \text{CPD}(s_e)$;
Predict motion-transition phase indicators $\rho \leftarrow \text{LSTM}(\Delta s_e)$;

Gradient Adjustment during Policy Learning
for $j = 0, 1, \cdots, T-1$ **do**
  Sample a fresh mini-batch B_j from expert demonstrations o_e;
  Feed-forward the batched data B_j to π_{v+s};
  Calculate average indicator ρ_j of B_j;
  Update proprioception chunk ω_s^{j+1} using Equation 5;
  Update vision chunk ω_v^{j+1}.
end for

5 Experiments

In this section, we conduct validate the versatility and effectiveness of our proposed Gradient Adjustment with Phase-guidance (GAP) algorithm through a series of question-driven experiments. The evaluations comprehensively cover a wide range of manipulation tasks, including simple pick-and-place tasks, rotation-sensitive tasks, as well as long-horizon and contact-rich tasks.

5.1 Experimental Setup

We select two simulated environments as our benchmarks: MetaWorld Yu et al. (2020) and RoboSuite Zhu et al. (2020). Tasks in MetaWorld are relatively simple, featuring a 4-dimensional action space that includes the gripper's position and its opening degree, while tasks in RoboSuite involve complex scenarios, longer task sequence horizons and richer physical interactions, with the action space further including the orientation of the gripper. For real-world experiments shown in Figure 3, we use a 6-DoF xArm 6 robotic arm equipped with a Robotiq gripper for all one-arm tasks. Moreover, we utilize the open-source Cobot Magic platform to support tasks that require dual-arm collaboration. In all tasks, the initial position of target object varies randomly in each validation, while the initial position of gripper remains fixed.

5.2 Can vision-proprioception policies outperform after GAP?

Vision-Proprioception policies perform generally worse than vision-only policies. Can they outperform vision-only policies after our GAP algorithm is applied? To answer this, we conducted comparative analyses between our algorithm and the following baselines:

- MS-Bot Feng et al. (2024): this method uses state tokens with stage information to guide the dynamic collaboration of modalities within multi-modality policy.
- Auxiliary Loss (Aux): following HumanPlus Fu et al. (2024), we use visual feature to predict the next frames as an auxiliary loss, which tries to enhance the vision modality.
- Mask: to prevents the overfitting to specific modality, RDT-1B Liu et al. (2024) randomly and independently masks each uni-modal input with a certain probability during encoding. We adapt the algorithm by masking only proprioception modality instead.

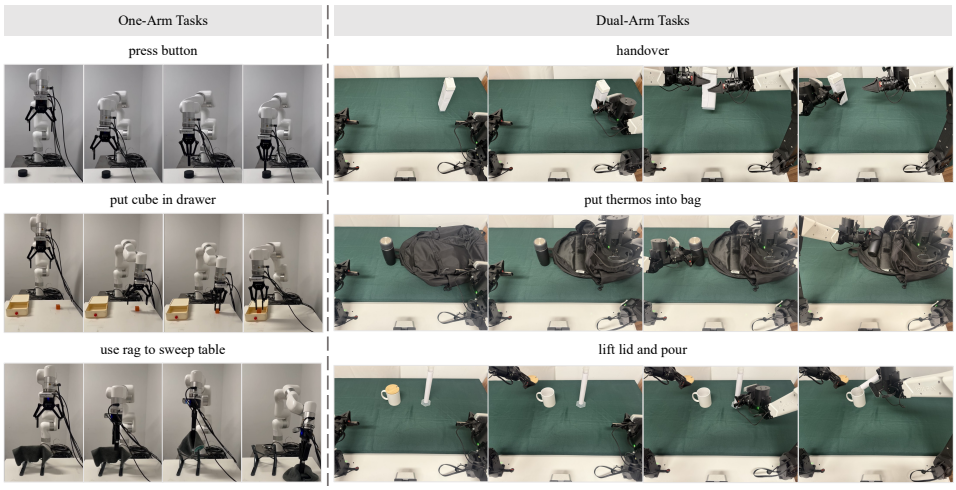

Figure 3: Visualization of real-world tasks. Our experiments cover a wide range of manipulation tasks, including both One-Arm and Dual-Arm Setups.

Results in Table 1 demonstrate that vision-proprioception policies with our GAP applied outperform vision-only policies and other methods. Although MS-Bot achieves overall improvements over the vision-only policy by incorporating stage information, it focuses on the semantic stage instead of motion-transition phases. As a result, its benefits are marginal in tasks like "push-wall" and "lift lid and pour", where motion frequently transits. This highlights the necessity of fine-grained gradient adjustment during motion-transition phases. Auxiliary loss forces the vision-proprioception policy to concentrate on visual input during the whole task, which falls short in tasks requiring proprioception to enhance the precision and robustness of manipulation, such as "threading". Meanwhile, masking the proprioceptive input with a fixed probability overlooks the modality temporality of manipulation tasks, resulting in minimal improvement. By adaptively applying fine-grained gradient adjustment during motion-transition phases, GAP enables the vision-proprioception policy to effectively leverage these two modalities and outperform both the vision-only policy and other methods.

Table 1: Comparisons with other methods in both simulated and real-world environments. The vision-proprioception policies after our gradient adjustment significantly outperform other methods.

Suite	Meta-World					RoboSuite		
Task / Method	pick-place	assembly	disassemble	push-wall	bin-picking	hammer	stack	threading
Vision-only	92%	82%	85%	64%	63%	86%	67%	44%
Concatenation	79%	76%	80%	56%	49%	79%	56%	34%
MS-Bot Feng et al. (2024)	90%	93%	88%	67%	70%	88%	70%	51%
Aux Fu et al. (2024)	89%	93%	78%	51%	55%	72%	55%	47%
Mask Liu et al. (2024)	86%	90%	84%	82%	61%	79%	62%	48%
GAP (Ours)	94%	96%	91%	73%	70%	91%	77%	52%

Setup	Real One-Arm			Real Dual-Arm		
Task / Method	press button	put cube in drawer	use rag to sweep table	handover	put thermos into bag	lift lid and pour
Vision-only	18/20	14/20	9/20	15/20	11/20	9/20
Concatenation	12/20	11/20	5/20	12/20	7/20	5/20
MS-Bot Feng et al. (2024)	20/20	16/20	11/20	16/20	13/20	10/20
Aux Fu et al. (2024)	19/20	16/20	11/20	15/20	13/20	8/20
Mask Liu et al. (2024)	18/20	14/20	7/20	15/20	9/20	7/20
GAP (Ours)	20/20	17/20	13/20	18/20	16/20	15/20

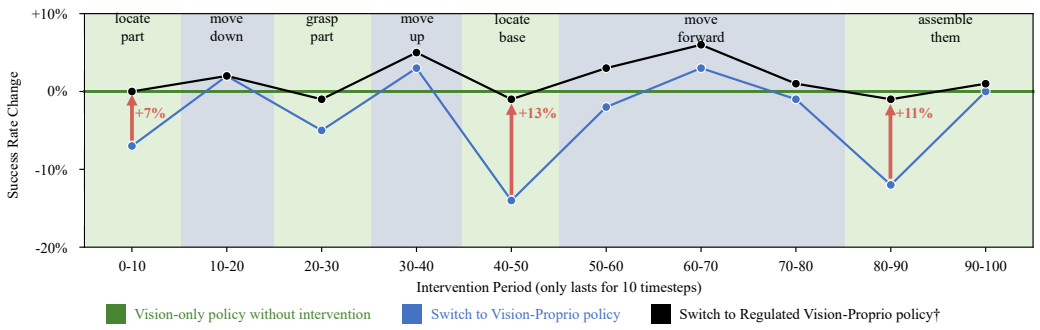

Figure 4: The Intervention experiment we conduct with the regulated vision-proprioception policy. The sight changes in success rate indicate that GAP does enhance the utilization of vision modality.

5.3 Does GAP enhance the utilization of vision modality?

Although vision-proprioception policies outperform vision-only policies after apply GAP, it remains unclear whether GAP truly enhances the utilization of the vision modality within vision-proprioception policies. To answer this, we first conducted intervention experiment under the same settings as described in Section 1. As shown in the Figure 4, the degrees of suppression of vision modality during motion-transition phases are significantly reduced after applying GAP, indicating GAP does enhance the utilization of vision modality. We further evaluated the generalization of the vision-proprioception policies in out-of-distribution (OOD) scenarios. In each scenario, the initial distribution of object positions differs from that in the training dataset of expert demonstrations. The vision-only policies are less affected by such changes due to well-utilized vision modality as demonstrated in Tabel 2. Vision-proprioception policies exhibit poor generalization with suppressed vision. Meanwhile, Our algorithm alleviates this by regulating the optimization of the proprioceptive, preventing the suppression. The maintained superior performance over vision-only policy also indicates the effectiveness of introducing proprioception modality for precise and robust manipulation.

Table 2: Experiments under out-of-distribution settings. For each task, our proposed GAP algorithm enhances the generalization of the vision-proprioception policies.

Setup	Meta-World		RoboSuite		Real One-Arm	Real Dual-Arm
Method \ Task	assembly	bin-picking	stack	threading	put cube in drawer	handover
Vision-only	78%	59%	63%	32%	12/20	12/20
Concatenation	62%	32%	49%	28%	7/20	9/20
Ours	88%	67%	72%	49%	15/20	15/20

5.4 Is GAP compatible with Vision-Language-Action models?

Above experiments have demonstrated that our algorithm facilitates dynamic collaboration within conventional vision-proprioception models. We further investigate is GAP compatible with Vision-Language-Action (VLA) models. Specifically, we compare fine-tuned Octo model Octo Model Team et al. (2024) using only visual information (Octo-V) versus using both vision and proprioception (Octo-VP), and tries to apply our gradient adjustment algorithm during fine-tuning. As reported in the original paper, policies trained with additional propioception seemed generally worse than vision-only policies. We observe the same trend across various tasks in Table 3. However, after applying our gradient adjustment algorithm, Octo-VP† achieves an average improvement of 17% and exhibits stronger generalization ability than Octo-V. These results suggest that our algorithm effectively enhances dynamic collaboration between vision and proprioception within VLA models.

Table 3: Performances of fine-tuned Octo. † indicates GAP is applied.

Suite	Meta-World		RoboSuite	
Model \ Task	disassemble	push-wall	hammer	threading
Octo-V	95%	77%	92%	69%
Octo-VP	82%	65%	88%	57%
Octo-VP†	100%	85%	97%	78%

5.5 Can our algorithm be applied to various modality fusion approaches?

The preliminary results in Section 3 reveal that the vision-proprioception policy using straightforward concatenation tends to perform worse than the vision-only policy. We further explore a broader set of fusion approaches and validate the versatility of our algorithm. Specifically, we apply GAP to three typical and widely used fusion approaches: concatenation, summation and FiLM Perez et al. (2018).

As reported in Table 4, vision-only policies outperforms all three fusion approaches in tasks such as "pick-place" and "hammer", indicating that vision modality suffice for certain tasks. However, they fail drastically in task "threading" due to demands for precise manipulation and exhibits suboptimal performance in task "push-wall", which involves visual occlusions at the target location, highlighting the necessity of the inclusion of proprioceptive information for precise and robust manipulation.

Table 4: Performance of typical fusion approaches combined with GAP. † indicates GAP is applied.

Suite	Meta-World					RoboSuite		
Method \ Task	pick-place	assembly	disassemble	push-wall	bin-picking	hammer	stack	threading
Vision-only	92%	82%	85%	64%	63%	86%	67%	44%
Concatenation	79%	76%	80%	56%	49%	79%	56%	34%
Summation	78%	95%	80%	54%	61%	75%	49%	30%
FiLM	75%	91%	47%	67%	59%	76%	53%	41%
Concatenation†	94%	96%	91%	73%	70%	91%	77%	52%
Summation†	92%	97%	93%	66%	70%	88%	82%	48%
FiLM†	90%	94%	85%	74%	68%	95%	72%	46%

Concatenation preserves raw features from both modalities, but the high-dimensional redundancy hinders the policy to dynamically utilize each modality. As a result, it underperforms in tasks like "push-wall", where effective coordination is required. Simple summation may obscure critical details, whose limitation is evident in precise manipulation tasks such as "threading" and "push-wall". Meanwhile, FiLM applies affine transformations to conditionally adjust features, making it more suitable for tasks requiring modality collaboration. For instance, it achieves a notably higher score in "push-wall" task. However, its performance tends to degrade in simpler tasks where such complex conditioning may be unnecessary. Conversely, GAP successfully unlocked the full potential of the vision-proprioception policy, outperforming vision-only policies in all three fusion approaches.

6 Conclusion and Limitation

In this study, we illustrate that vision-proprioception policy would fail during motion-transition phases due to its suppressed vision modality. To alleviate this, we propose the Gradient Adjustment with Phase-guidance (GAP) algorithm, enabling dynamic collaboration between vision and proprioception within vision-proprioception policy. We believe this work can offer valuable insights into the development of vision-proprioception policies for robotic manipulation.

Limitations. All vision-proprioception policies are trained on single embodiment in this work. As existing large-scale datasets often contain diverse embodiments, exploring the role of proprioception in cross-embodiment datasets would be promising for future research.

References

J. Allum, B. Bloem, M. Carpenter, M. Hulliger, and M. Hadders-Algra. Proprioceptive control of posture: a review of new concepts. *Gait & posture*, 8(3):214–242, 1998.

B. Henze, C. Ott, and M. A. Roa. Posture and balance control for humanoid robots in multi-contact scenarios based on model predictive control. In *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, pages 3253–3258, 2014.

M. Bjelonic, N. Kottege, and P. Beckerle. Proprioceptive control of an over-actuated hexapod robot in unstructured terrain. In *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, pages 2042–2049, 2016.

J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter. Learning quadrupedal locomotion over challenging terrain. *Science robotics*, 5(47):eabc5986, 2020.

Y. Yang, J. Norby, J. K. Yim, and A. M. Johnson. Proprioception and tail control enable extreme terrain traversal by quadruped robots. In *Proceedings of International Conference on Intelligent Robots and Systems (IROS)*, pages 735–742, 2023.

S. Levine, C. Finn, T. Darrell, and P. Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.

L. Cong, H. Liang, P. Ruppel, Y. Shi, M. Görner, N. Hendrich, and J. Zhang. Reinforcement learning with vision-proprioception model for robot planar pushing. *Frontiers in Neurorobotics*, 16:829437, 2022.

G. Jiang, Y. Sun, T. Huang, H. Li, Y. Liang, and H. Xu. Robots pre-train robots: Manipulation-centric robotic representation from large-scale robot datasets. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2025.

L. Wang, X. Chen, J. Zhao, and K. He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. *Advances in Neural Information Processing Systems (NeurIPS)*, 37: 124420–124450, 2024.

Octo Model Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, C. Xu, J. Luo, T. Kreiman, Y. Tan, L. Y. Chen, P. Sanketi, Q. Vuong, T. Xiao, D. Sadigh, C. Finn, and S. Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, 2024.

F. R. Sarlegna and R. L. Sainburg. The roles of vision and proprioception in the planning of reaching movements. *Progress in Motor Control: A Multidisciplinary Perspective*, pages 317–335, 2009.

R. Feng, D. Hu, W. Ma, and X. Li. Play to the score: Stage-guided dynamic multi-sensory fusion for robotic manipulation. In *Proceedings of Conference on Robot Learning (CoRL)*, 2024.

Z. He, H. Fang, J. Chen, H.-S. Fang, and C. Lu. Foar: Force-aware reactive policy for contact-rich robotic manipulation. *IEEE Robotics and Automation Letters*, 2025.

N. Tsagkas, A. Sochopoulos, D. Danier, S. Vijayakumar, C. X. Lu, and O. Mac Aodha. When pre-trained visual representations fall short: Limitations in visuo-motor robot learning. *arXiv preprint arXiv:2502.03270*, 2025.

Y. Huang, J. Lin, C. Zhou, H. Yang, and L. Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *Proceedings of International Conference on Machine Learning (ICML)*, pages 9226–9259. PMLR, 2022.

Y. Fan, W. Xu, H. Wang, J. Wang, and S. Guo. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20029–20038, 2023.

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 *Proceedings of Conference on Robot Learning (CoRL)*, pages 2165–2183. PMLR, 2023.

M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, et al. Openvla: An open-source vision-language-action model. In *Proceedings of Conference on Robot Learning (CoRL)*, 2024.

J. Zeng, Q. Bu, B. Wang, W. Xia, L. Chen, H. Dong, H. Song, D. Wang, D. Hu, P. Luo, et al. Learning manipulation by predicting interaction. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *Proceedings of Conference on Robot Learning (CoRL)*, pages 1678–1690, 2022.

C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn. Humanplus: Humanoid shadowing and imitation from humans. In *Proceedings of Conference on Robot Learning (CoRL)*, 2024.

S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024.

M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *Proceedings of International Conference on Robotics and Automation (ICRA)*, pages 8943–8950, 2019.

J. Jiang, K. Ota, D. K. Jha, and A. Kanezaki. Modality selection and skill segmentation via cross-modality attention. *https://arxiv.org/abs/2504.14573*, 2025.

H. Li, Y. Zhang, J. Zhu, S. Wang, M. A. Lee, H. Xu, E. Adelson, L. Fei-Fei, R. Gao, and J. Wu. See, hear, and feel: Smart sensory fusion for robotic manipulation. In *Proceedings of Conference on Robot Learning (CoRL)*, pages 1368–1378, 2023.

F. Torabi, G. Warnell, and P. Stone. Behavioral cloning from observation. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4950–4957, 2018.

S. Liu, M. Yamada, N. Collier, and M. Sugiyama. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43:72–83, 2013. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2013.01.012. URL `https://www.sciencedirect.com/science/article/pii/S0893608013000270`.

S. Aminikhanghahi and D. J. Cook. A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51(2):339–367, 2017.

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 *Proceedings of Conference on Robot Learning (CoRL)*, pages 1094–1100, 2020.

Y. Zhu, J. Wong, A. Mandlekar, R. Martín-Martín, A. Joshi, S. Nasiriany, Y. Zhu, and K. Lin. robosuite: A modular simulation framework and benchmark for robot learning. In *arXiv preprint arXiv:2009.12293*, 2020.

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*, volume 32, 2018.

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: In this work, we answer when would vision-proprioception policy fail and introduce a modality-temporality perspective, as thoroughly presented in Abstract and Section 1.

   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 discuss the limitations of our work in Section 6.

   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: [NA]

Justification: There is no theoretical results in our work.

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 details of our proposed algorithm are described in Section 2

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: The code of the paper will be released upon acceptance.

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: Training and test details are specified in Section 5.1 and Supplementary Materials.

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: : In the Supplementary Material, we report the average accuracy and standard deviation using three different random seeds. Due to computational resource limitations, we calculate these only in main experiments of simulation environments.

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: In the Supplementary Materials, we provide detailed information on the type of GPUs, the number of uesd GPUs, and their memory capacities.

   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: We have reviewed the NeurIPS Code of Ethics, and our work fully complies with it.

   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 adjustments 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: [NA]

    Justification: This paper focuses on answering when vision-proprioception policy would fail in robotic manipulation, therefore it has no societal impact.

    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: All experiments in our work are conducted on publicly available datasets, and our algorithm does not involve any associated 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: All experiments in our work are conducted on publicly available datasets, and we have cited the original papers in Section 5.1.

    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: Our work 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: Our work does not involve crowdsourcing or 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: Our work does not involve crowdsourcing or 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 our work 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.