View-Imagination: Enhancing Visuomotor Control with Adaptive View Synthesis

Dohyeok Lee[1], Munkyung Kim[1], Jung Min Lee[1], Seungyub Han[1], Jungwoo Lee[1*]

Abstract

In robotic manipulation tasks, visuomotor control suffers from limited spatial understanding problems with limited camera installation and visual imperfections, such as occlusion. In this paper, we propose view-imagination, a novel framework with incorporating viewpoint policy. We train a dynamic scene NeRF and a learnable viewpoint policy, enabling the robot to generate a maximum value viewpoint to improve affordance. In experiments, we demonstrate that view-imagination outperforms across various training configurations.

1. Introduction

Manipulation with visual observation, especially image-based deep reinforcement learning (RL) and imitation learning have yielded significant advances in robot learning [4, 7, 13]. A robot learning agent should understand spatial space from image observation to solve manipulation tasks. Previous works [9, 12, 27] using single-view image observation for RL agents suffer from the lack of enough spatial information for the environment due to the absence of diverse viewpoint image observations. To solve the lack of spatial information, multi-view RL methods for robot learning [5, 15, 22, 25] have been proposed to provide an agent with the input observation from various viewpoints.

However, existing studies have attempted to use observations from a fixed set of multiple viewpoints or 3D-aware latent vectors without fully understanding the differences between viewpoints. Given the benefits of 3D awareness, we hypothesize that the most beneficial viewpoint is scene-dependent and can resolve visual ambiguities such as occlusions.

In this paper, we propose *view-imagination*, a framework that leverages valuable novel viewpoint, considering the unique information of each viewpoint. Our method generates synthetic adaptive viewpoints based on current scene state using learned viewpoint policy, and exploits them when solving manipulation tasks.

*Corresponding author, [1]Seoul National University

Key Insight: Traditional multi-view approaches use fixed cameras that cannot adapt to changing occlusions or task phases. For example, during door opening, the optimal viewpoint shifts as the robot moves and the door handle becomes visible from different angles. Our adaptive approach dynamically selects the most informative viewpoint for each scene state, leading to more robust visuomotor control.

To summarize, the main contributions are as follows. **(1)** While previous works use fixed multi-view setups, we propose the first approach to dynamically select viewpoints based on current task state and predicted value, enabling better spatial understanding. **(2)** To generate a more gainful viewpoint, we propose *value learning* to train a viewpoint policy to select a viewpoint with maximum value using the critic model.

Note that while our current implementation uses NeRF trained on multi-view data, this represents a proof-of-concept that can be replaced with few-shot or zero-shot novel view synthesis methods [8, 21], eliminating the need for multi-view data collection entirely.

2. View-Imagination

In this section, we propose *view-imagination* for robotic manipulation tasks. To clarify our novelty which focuses on viewpoint, we evaluate the performance of baseline on `robosuite Lift` [30] task for the viewpoint of front, side, and bird in Figure 1, which demonstrates performance difference across different viewpoint settings. In the following sections, we utilize and denote DreamerV1 [9] as a baseline algorithm for convenience. The key insight is that different viewpoints provide different amounts of task-relevant information. For instance, when grasping a door handle, a side view may be more informative than a front view depending on the robot's current position. We formalize this by defining the value of a viewpoint as the expected return when using that viewpoint's observation.

To leverage difference between viewpoints, view-imagination operates in two phases: (1) **NeRF Training**: We first collect multi-view data and train a dynamic scene NeRF f_θ that can synthesize novel views from arbitrary camera poses. While our current proof-of-concept requires multi-view training data, our framework is readily exten-

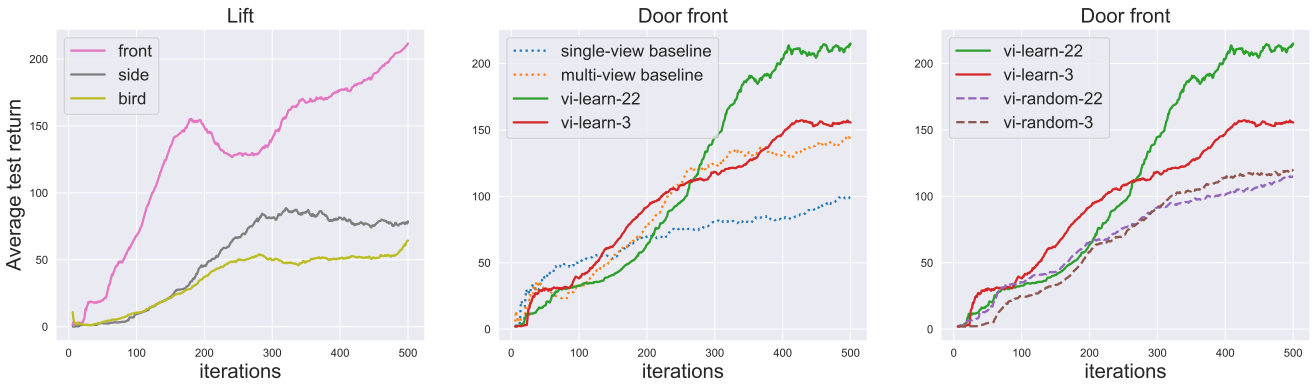

Figure 1. All plot has iteration as x-axis and average test return as y-axis. **Left**: Performance of baseline with front view, side view, and bird view observation for Lift task. **Middle**: Performance of single-view baseline (front), multi-view baseline (front,side), vi-learn-22, and vi-learn-3 (front,adaptive) **Right**: Performance of vi-learn-22, vi-learn-3, vi-random-22, and vi-random-3.

sible to zero-shot novel view synthesis models that eliminate this requirement entirely, making deployment significantly more practical. (2) **Viewpoint Policy Learning**: We then train a viewpoint policy π_ω that selects most beneficial viewpoints based on current scene state. For clarity, *view* represents an observation obtained from a certain viewpoint, and *viewpoint* means a certain camera pose. Detailed explanations are given in Appendix D.

Algorithm 1 View Imagination

Initialize viewpoint policy ω, RSSM θ, policy ϕ, critic ψ, replay buffer $\mathcal{D}$
for $l = 1, \ldots, M$ **do**
  for $c = 1, \ldots, C$ **do**
    $\mathcal{B} = \{(o_t, \tilde{o}_t(v), a_t, r_t, d_t)\}_{t=k}^{t=k+L} \sim D$
    `// dynamics learning`
    Update θ by minimizing $\mathcal{L}^{\text{wm}}(\theta)$ using $\mathcal{B}$
    `// behavior learning`
    Dreaming $\tau = \{(\hat{s}_i, \hat{a}_i, \hat{r}_i)\}_{i=1}^{H}$
    Update ϕ, ψ by $\mathcal{L}^{\text{actor}}(\phi), \mathcal{L}^{\text{critic}}(\psi)$ using τ
  `// value learning`
  for $i = 1, \ldots, N$ **do**
    Compute target y^v using $V_\psi(s_t^{v^i})$
    Update ω by minimizing $\mathcal{L}^{\text{vi}}(\omega)$
  `// env interaction`
  for $t = 1, \ldots, T$ **do**
    Synthesize $v \sim \pi_\omega(\cdot|s_t), \tilde{o}_t(v) = \text{NeRF}(o_t, v)$
    Sample action $a_t \sim \pi_\phi(\cdot|o_t, \tilde{o}_t(v))$
    $D \leftarrow D \cup \{o_t, \tilde{o}_t(v), a_t, r_t, d_t\}$

3. Experimental Results

We first train a dynamic scene NeRF model using data collected from random policy, as shown in Figure 2 in Appendix. After training the NeRF model, the viewpoint policy is optimized through value learning in conjunction with a baseline algorithm as described in Algorithm 1. The test return curve in Figure 1 compares the baseline with view-imagination combined with value learning across 22 viewpoints, referred to as *vi-learn-22*. Note that our algorithm receives only a single-view image from the environment and generates an additional adaptive view, while the baseline receives multi-view observations from the environment.

Our results show that vi-learn-22 achieves superior performance, outperforming both single-view baseline (2.2× improvement) and fixed multi-view approaches (1.4x improvement). Importantly, value learning is crucial for this performance gain: while random viewpoint selection (*vi-random-22*) outperforms single-view baselines, it falls short of fixed multi-view methods. This demonstrates that adaptive viewpoint selection requires value-driven decisions rather than random sampling to surpass carefully positioned fixed cameras.

In the context of affordance, we visualize the saliency map of both the baseline and view-imagination with value learning using 3 viewpoints (*vi-learn-3*) from the front view in Figure 2 to investigate how view-imagination aids in identifying informative objects. Qualitatively, the saliency map of vi-learn-3 tends to focus on both the manipulator and the door, whereas the baseline only concentrates on the manipulator. This suggests that adding an additional, even synthesized, adaptive view enhances the agent's affordance.

Ablation Study: Figure 1 presents comprehensive ablation results. Comparing vi-learn-22 and vi-random-22 demonstrates the importance of value-based selection over random sampling. The comparison between vi-learn-22 and vi-learn-3 shows that having more viewpoint candidates allows better selection, even though some may be suboptimal.

To sum up, view-imagination demonstrates significant performance improvements over the baseline, even with the various training configurations. The performance gain can be explained by the enhanced affordance of the view-imagination agent, as shown in the saliency map.

Acknowledgements This work is in part supported by the National Research Foundation of Korea (NRF, RS-2024-00451435(20%), RS-2024-00413957(20%)), Institute of Information & communications Technology Planning & Evaluation (IITP, RS-2021-II212068(10%), RS-2025-02305453(15%), RS-2025-02273157(15%), 2021-0-00180(10%), Artificial Intelligence Graduate School Program (Seoul National University)(10%)) grant funded by the Ministry of Science and ICT (MSIT), Institute of New Media and Communications(INMAC), the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2025, and Samsung Electronics Co., Ltd(IO210202-08370-01).

References

[1] Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. Uncertainty-based offline reinforcement learning with diversified q-ensemble, 2021. 6

[2] Michel Breyer, Lionel Ott, Roland Siegwart, and Jen Jen Chung. Closed-loop next-best-view planning for target-driven grasping. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1411–1416. IEEE, 2022. 5

[3] Lawrence Yunliang Chen, Chenfeng Xu, Karthik Dharmarajan, Zubair Irshad, Richard Cheng, Kurt Keutzer, Masayoshi Tomizuka, Quan Vuong, and Ken Goldberg. Rovi-aug: Robot and viewpoint augmentation for cross-embodiment robot learning. *arXiv preprint arXiv:2409.03403*, 2024. 5

[4] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023. 1

[5] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. *arXiv preprint arXiv:1910.11215*, 2019. 1

[6] Danny Driess, Ingmar Schubert, Pete Florence, Yunzhu Li, and Marc Toussaint. Reinforcement learning with neural radiance fields. *Advances in Neural Information Processing Systems*, 35:16931–16945, 2022. 5

[7] Zipeng Fu, Tony Z. Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. In *Conference on Robot Learning (CoRL)*, 2024. 1

[8] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T. Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models, 2024. 1, 6

[9] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019. 1, 5

[10] Danijar Hafner, Timothy P Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *International Conference on Learning Representations*, 2020. 5

[11] Wenbo Hu, Yuling Wang, Lin Ma, Bangbang Yang, Lin Gao, Xiao Liu, and Yuewen Ma. Tri-miprf: Tri-mip representation for efficient anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19774–19783, 2023. 6

[12] Wenlong Huang, Igor Mordatch, Pieter Abbeel, and Deepak Pathak. Generalization in dexterous manipulation via geometry-aware multi-task learning. *arXiv preprint arXiv:2111.03062*, 2021. 1

[13] Dmitry Kalashnikov, Jacob Varley, Yevgen Chebotar, Benjamin Swanson, Rico Jonschkowski, Chelsea Finn, Sergey Levine, and Karol Hausman. Mt-opt: Continuous multi-task robotic reinforcement learning at scale. *arXiv preprint arXiv:2104.08212*, 2021. 1

[14] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 6

[15] Akira Kinose, Masashi Okada, Ryo Okumura, and Tadahiro Taniguchi. Multi-view dreaming: Multi-view world model with contrastive learning. *Advanced Robotics*, 37(19):1212–1220, 2023. 1, 5

[16] Dohyeok Lee, Seungyub Han, Taehyun Cho, and Jungwoo Lee. Spqr: Controlling q-ensemble independence with spiked random model for reinforcement learning, 2024. 6

[17] Yunzhu Li, Shuang Li, Vincent Sitzmann, Pulkit Agrawal, and Antonio Torralba. 3d neural scene representations for visuomotor control. In *Conference on Robot Learning*, pages 112–123. PMLR, 2022. 5

[18] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 5

[19] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 6

[20] Oleh Rybkin, Chuning Zhu, Anusha Nagabandi, Kostas Daniilidis, Igor Mordatch, and Sergey Levine. Model-based reinforcement learning via latent-space collocation. In *International Conference on Machine Learning*, pages 9190–9201. PMLR, 2021. 5

[21] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9420–9429, 2024. 1, 5, 6

[22] Younggyo Seo, Junsu Kim, Stephen James, Kimin Lee, Jinwoo Shin, and Pieter Abbeel. Multi-view masked world models for visual robotic manipulation. In *International Conference on Machine Learning*, pages 30613–30632. PMLR, 2023. 1, 5

[23] Tim Seyde, Wilko Schwarting, Sertac Karaman, and Daniela Rus. Learning to plan optimistically: Uncertainty-guided

deep exploration via latent model ensembles. *arXiv preprint arXiv:2010.14641*, 2020. 5

[24] Dongseok Shim, Seungjae Lee, and H Jin Kim. Snerl: Semantic-aware neural radiance fields for reinforcement learning. In *International Conference on Machine Learning*, pages 31489–31503. PMLR, 2023. 5

[25] Mel Vecerik, Jean-Baptiste Regli, Oleg Sushkov, David Barker, Rugile Pevceviciute, Thomas Rothörl, Raia Hadsell, Lourdes Agapito, and Jonathan Scholz. S3k: Self-supervised semantic keypoints for robotic manipulation via multi-view consistency. In *Proceedings of the 2020 Conference on Robot Learning*, pages 449–460. PMLR, 2021. 1

[26] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *Conference on robot learning*, pages 2226–2240. PMLR, 2023. 5

[27] Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Mastering visual continuous control: Improved data-augmented reinforcement learning. *arXiv preprint arXiv:2107.09645*, 2021. 1

[28] Xuechao Zhang, Dong Wang, Sun Han, Weichuang Li, Bin Zhao, Zhigang Wang, Xiaoming Duan, Chongrong Fang, Xuelong Li, and Jianping He. Affordance-driven next-best-view planning for robotic grasping. In *7th Annual Conference on Robot Learning*, 2023. 5

[29] Allan Zhou, Moo Jin Kim, Lirui Wang, Pete Florence, and Chelsea Finn. Nerf in the palm of your hand: Corrective augmentation for robotics via novel-view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17907–17917, 2023. 5

[30] Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Soroush Nasiriany, and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293*, 2020. 1

Appendix

A. Related Work

Our work is related to three areas: image-based RL for manipulation, Neural Radiance Fields (NeRF) for dynamic scenes, and novel view synthesis for robot learning. Image-based RL for manipulation uses image observations as input and understands the environment from given images or learned latent representations. DreamerV1 [9] proposed a method that learns an actor-critic model by dreaming in the learned latent space for single-view image observation. However, this approach lacks consideration of multiple image observations. MV-MWM [22] and Multi-View Dreaming [15] enhanced DreamerV2 [10] by using multi-view representation learning. Multi-view approaches use fixed camera configurations that cannot adapt to dynamic scene changes or task-specific requirements. In contrast, our approach dynamically selects viewpoints based on current state, providing better coverage of task-relevant regions.

NeRF [18] represents 3D scenes using volumetric rendering. Since NeRF is only suitable for static scenes, prior works have attempted to adopt NeRF for dynamic scenes. NeRF-dy [17] first proposed an encoder-decoder framework that handles dynamics modeling. NeRF-RL [6] and SNeRL [24] learn policies using the representation from a encoder. However, the representation used is only 3D-aware, and does not consider the most beneficial viewpoints.

SPARTN [29] and Rovi-Aug [3] introduced an algorithm that generates images from new viewpoints using NeRF and ZeroNVS [21]. Both algorithms are data augmentation techniques for visual control, whereas our algorithm exploits the viewpoint itself. Next-best-view algorithms [2, 28] are designed to alleviate occlusions by anticipating the best viewpoint. However, these works have primarily focused on relocating the egocentric camera viewpoints, rather than the remote camera system or synthesizing additional view.

B. Background

Neural Radiance Fields for Dynamic Scenes. To represent a 3D scene for novel view synthesis, NeRF shows significant improvement compared to prior works. NeRF represents a 3D scene by modeling volumetric fields using neural network. Specifically, given 3D world coordinate $\mathbf{x}$ and unit direction vector $\mathbf{d}$, f_θ estimates RGB color $\mathbf{c}$ and volume density σ: $f_\theta(\mathbf{x}, \mathbf{d}) = (\mathbf{c}, \sigma)$. From volumetric rendering, pixel rendering is given as $C(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(r(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d})$ where $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ is camera ray with camera origin $\mathbf{o}$, and $T(t) = \exp(-\int_{t_n}^{t} \sigma(s)ds)$. Since original NeRF architecture can represent only static scene, Driess et al. [6], Li et al. [17], Shim et al. [24] used the encoder-decoder architecture to represent dynamic scene for NeRF. Formally, encoder Ω embeds observation $o^{1:V}$ and corresponding camera matrix $K^{1:V}$ by $z = \Omega(o^{1:V}, K^{1:V})$. For the decoder, we implement a latent-conditioned NeRF model for the volumetric rendering of dynamic scene: $f_\theta(z, \mathbf{x}, \mathbf{d}) = (\mathbf{c}, \sigma)$.

Dreamer. Dreamer is an effective world model method for robot learning as shown in [20, 22, 23, 26]. Hafner et al. [9] introduces DreamerV1, with the concept of dreaming which trains agents using a learnable latent dynamics model without interaction with the environment. DreamerV1 consists of a recurrent state space model (RSSM) to model latent space dynamics, the encoder-decoder network for image observation, and the actor-critic model which is trained by dreaming with a given latent dynamics model.

C. Algorithmic Details

For a set of observations $\mathcal{O} = \{o_0, o_1, ...\}, o_i \in \mathbb{R}^{3\times H\times W}$, the most beneficial observation is defined by the one with the largest expected return from selecting that observation. We propose the *value learning* algorithm to train *viewpoint policy* for generating *adaptive viewpoint*. Value learning enforces viewpoint policy to generate a viewpoint that has maximum value computed by the critic model. Formally, we defined viewpoint policy as $v \sim \pi_\omega(v|s_t)$ where $v \in$ SE(3). Given a set of viewpoints $\mathcal{V}$ and critic model $V_\psi(s_t)$ for state s_t, loss function is formulated by cross-entropy of viewpoint policy π_ω relative to a target vector y^v which represents a maximum value viewpoint v:

$$\mathcal{L}^{vi}(\omega) = -\sum_{v_i\in\mathcal{V}} y^{v_i} \log(\pi_\omega(v_i|s_t))$$
$$v^* = \arg\max_{v_i\in\mathcal{V}} V_\psi(s_t(o_t, \tilde{o}_t(v_i)))$$

D. Implementation Details

We implement the baseline with a multi-view encoder to cover both environmental observation and synthesized observation from our model. To distinguish whether each observation is given from the environment or synthesized by the NeRF model, we denoted environmental observation as o_t and synthesized observation from viewpoint v as $\tilde{o}_t(v)$. Following the notation of baseline, we denote the parameters of RSSM, reward predictor, and encoder-decoder model as θ, actor model as ϕ, and critic model as ψ.

In our experiments, using DreamerV1 with a multi-view encoder as the baseline, we implemented view-imagination with varying numbers of viewpoints, without value learning methods, and different environmental observations. We also visualized saliency maps for qualitative affordance evaluation. The algorithms were evaluated in the `robosuite Door` environment, with operational space as the action

Figure 2. **Left:** Two sequences of view with four different viewpoints of the ground-truth observations **(top)** and the imagined results **(bottom)**, respectively. **Right:** Two saliency maps over policy, the baseline and our *view-imagination*. We note that *view-imagination* focuses better on two important objects, door and robot arm.

space. The `Door` environment, a complex long-horizon robotic manipulation task, requires the agent to grasp and manipulate a handle of a randomly located door to open it. To successfully complete this task, the agent must infer affordances from visual inputs, even when the door handle is partially occluded from certain viewpoints. We selected the `Door` environment to demonstrate the effectiveness of the view-imagination framework in scenarios with imperfect visual observations, such as occlusion.

E. Limitation and Future Work

While view-imagination demonstrates significant performance improvements, our current implementation has two key limitations: multi-view data requirement and computational overhead. The NeRF training phase requires collecting multi-view observations, which can be impractical for real-world deployment. Additionally, NeRF rendering during inference introduces latency that may limit real-time applications.

Future work will address these limitations along two directions: (1) **Reducing data requirements** by integrating few-shot or zero-shot novel view synthesis methods such as CAT3D [8] and ZeroNVS [21], eliminating the need for multi-view training data entirely. (2) **Improving computational efficiency** by adopting faster rendering algorithms including Instant NGP [19], Tri-MipRF [11], and 3D Gaussian Splatting [14]. These advances will enable practical deployment of view-imagination in real-world robotic systems where camera setup flexibility and real-time performance are crucial.

Additionally, following benefits of randomness in RL [1, 16], viewpoint policy trained viewpoint based on not only how valuable but also how independent view are. Also, our framework can be applied to various robot learning algorithms beyond DreamerV1, including imitation learning methods and Vision-Language-Action models. For algorithms that do not inherently use value functions (e.g., behavior cloning), alternative metrics such as action prediction confidence or attention maps can guide viewpoint selection.