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

Zero-Shot Exocentric Viewpoint-Robust Imitation Learning (VIL): Bridging Handheld Gripper and Exocentric Views

Boyan Li[1], Peilin Meng[1], Chang Liu[1], Yulin Chen[1], Qi Zhou[1] and Youyi Bi[1,†]

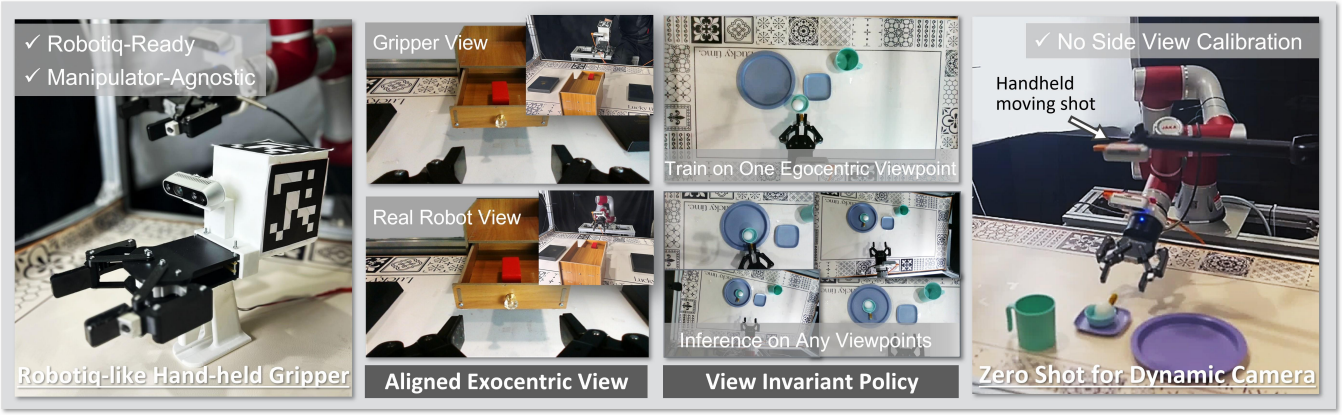

Fig. 1: Zero-shot exocentric viewpoint-robust imitation learning (VIL) integrates handheld gripper with exocentric observation and viewpoint-robust policy. Left: prototype of our Robotiq-like handheld gripper. Mid: VIL aligns egocentric views while adopting algorithm to generalize across exocentric viewpoints. Right: VIL delivers stable performance under dynamic views.

***Abstract*— Recent advances in robot learning have motivated integrated pipelines that combine hardware for data collection with imitation learning algorithms. Existing data collection methods like leader–follower, VR/AR, and exoskeletons rely on costly hardware and exhibit limited scalability, while imitation learning algorithms built on them remain highly sensitive to viewpoint shifts, further constraining generalizability. Handheld grippers provide a low-cost, robot-agnostic alternative, but prior systems bypass exocentric view alignment by relying solely on wrist-mounted cameras, resulting in narrowed observation and reduced policy robustness. We propose VIL, a framework pairing customized handheld gripper with zero-shot, exocentric viewpoint-robust imitation learning algorithm, bridging the handheld gripper with exocentric views. Our approach employs adapters for appearance alignment and a hybrid encoder design to extract view-consistent representations for an ACT-style policy, enabling robust execution across diverse perspectives. We further optimize the data collection pipeline and validate the system in both simulation and real-world tasks. Experiments show that VIL achieves stable performance under viewpoint shifts, challenging low-horizon scenarios, and dynamic perspectives, outperforming SOTA methods and demonstrating a scalable pipeline for manipulator-independent, viewpoint-robust policy learning. The project repository containing code and hardware is available at https://github.com/liboyan233/VIL.git.**

I. Introduction

The rapid progress of robot learning has given rise to diverse pipelines that combine advances in hardware design for data collection with imitation learning algorithms for policy generation [1–3]. Compared with pipelines based on teleoperation (e.g. leader–follower arms [1,4]), VR systems [5,6], or exoskeletons [7], demonstrating through handheld grippers [2,8,9] offers distinct advantages: they are low-cost, scalable without the need for a physical robot arm, and naturally produce manipulator-agnostic policies [2,10]. Despite these benefits, handheld grippers face a critical challenge, namely, limited compatibility with third-person (exocentric) views due to two primary factors. First, human operators or the robot arm itself often occlude or misalign exocentric observations, creating visual gap between training and inference that risks unstable or unreliable policy execution. Moreover, in-the-wild handheld setups—where robots are absent during data collection—cannot ensure strict calibration or consistently aligned exocentric viewpoints across training and inference, whereas most imitation learning algorithms still rely on the assumption of fixed and perfectly aligned perspectives [1,11,12]. Even minor shifts in exocentric viewpoints can cause substantial performance degradation, greatly undermining the reliability and convenience of redeployment [13,14]. Currently, handheld grippers with imitation learning pipelines typically rely solely on the first-person (egocentric) views [2]. This restricted view narrows observational coverage, limiting access to global context and depth cues, and reduces compatibility with current datasets widely adopted with third-person cameras, ultimately hindering scalability.

In this work, we propose VIL, a novel zero-shot exocentric viewpoint-robust imitation learning framework built on a custom-designed handheld gripper. Our pipeline bridges the gap between handheld egocentric pipelines and exocentric

† Correponding author
[1]Shanghai Jiao Tong University, Shanghai, China. Emails: {liboyan, helloworld.m, fzcowen1, chenyuli, zhouqi1998, youyi.bi}@sjtu.edu.cn.

The authors would like to acknowledge the financial support from the National Natural Science Foundation of China (52575300) and the National Key R&D Program of China (2022YFB4702400).

observations across viewpoints. Specifically, we adopt a novel vision-encoder design that exploits the strong viewpoint generalization capability of Swin Transformer (Swin-T) [15], integrated with ACT policy generation block. This allows the model to learn view-consistent representations, supporting more reliable and transferable policy learning across diverse perspectives. To further mitigate domain gaps between training and inference views, we incorporate state-of-the-art foundation models (e.g., SAM2) together with generalizable inpainting algorithm to remove manipulator- and human-related visual artifacts, ensuring consistent observations. We further design a Robotiq-2f-85-like handheld gripper inspired by UMI [2] while introducing several modifications to enable exocentric view integration and fit Robotiq's structure. The resulting design adopts widely available components and supports hybrid trajectory tracking, enabling robust data collection with additional exocentric views. Benefiting from the handheld gripper and proposed algorithm, our pipeline enables policies learned from handheld gripper demonstrations to be deployed—without fine-tuning—on any manipulator equipped with Robotiq-2f-85 gripper and under seen or unseen exocentric viewpoint, providing a scalable and generalizable solution for real-world imitation learning pipeline. Our main contributions are as follows:

- **Bridging the handheld–exocentric view gap**: Design a customized Robotiq-like handheld gripper and enable exocentric view observations for the handheld gripper pipeline via inpainting-based visual alignment.
- **Zero-shot viewpoint-robust**: Our visuomotor imitation algorithm attains robust generalization across unseen exocentric viewpoints, decouples the learned policy from specific views.
- **Robust visuomotor performance**: Experiments demonstrate accurate, functional, and stable action generation under novel or dynamic exocentric view with exocentric-egocentric complementary observations.

II. Related Work

A. Data Collection for Robotic Manipulation

A variety of data collection strategies have been explored to support robot learning pipelines. To resolve delicate manipulation tasks, leader-follower systems are adopted [1,4,16], as they provide intuitive control by directly mirroring human motions onto a robot arm. However, these systems inherently rely on physical robot arms during data collection, often requiring dual-arm setups or specially designed scaled-down leaders, which results in high costs and makes the data manipulator-specific. To reduce such dependencies, researchers have explored VR/AR systems [6,17–19] and exoskeleton-based interfaces [20,21], enabling demonstration collection without a robot in the loop. While these approaches allow more flexible data acquisition, they often suffer from either high latency with sim-to-real discrepancies or reliance on specialized hardware that limits scalability.

In contrast, handheld grippers [2,22–25] have emerged as a lightweight alternative, offering low cost, ease of use, and high scalability. They enable manipulator-agnostic demonstrations by decoupling data from robot-specific kinematics and minimized visual domain gap between training and inference with only wrist-mounted camera equipped. Nevertheless, existing handheld solutions still face a critical limitation: their egocentric perspective provides only partial observational coverage and is inherently mismatched with exocentric views that supply global context. In this work, we introduce both hardware and algorithmic modifications to bridge this gap, enabling our designed handheld gripper to seamlessly operate under exocentric viewpoint shifts, thus removing the need for viewpoint alignment and facilitating convenient collection-to-deployment transfer.

B. Imitation Learning under Viewpoint Variability

Building on these collected demonstrations, a broad range of imitation learning (IL) algorithms have been proposed. A representative baseline is Action Chunking with Transformers (ACT) [1], which leverages an encoder–decoder transformer to predict action chunks, thereby reducing compounding errors and producing smoother trajectories from noisy demonstrations. **Diffusion Policy (DP)** [11] reformulates policy generation as a conditional denoising process, enabling multimodal action prediction and more dexterous control. Despite their advances, these methods remain highly sensitive to viewpoint changes [1,11,26]: policies trained from fixed camera perspectives often degrade sharply under even slight deviations, leading to instability and outright task failures. This sensitivity poses a major obstacle for real-world deployment, as it requires duplicating training viewpoints during inference.

Several approaches attempt to alleviate this limitation. For instance, UMI [2] relies only on fisheye wrist-mounted cameras to sidestep viewpoint mismatch, but this design is incompatible with standard datasets' format with exocentric observations and risks losing essential global context. Other works pair robotic arms with active vision or "head motion prediction" to simulate dynamic exocentric perspectives [27,28]. However, such perspectives remain predefined and follow similar motion patterns, merely shifting the coupling between policy and view from static to dynamic, rather than removing it. Further researches attempt to utilize pointcloud as coordinate-aligned context [20,29,30], but requires strict calibration, making it challenging for in-the-wild settings where robots are missing.

On another front, data augmentation and domain randomization [14,31–34] have been used to create viewpoint diversity during training, yet these approaches involve tedious preprocessing and yield only limited generalization. In contrast, our proposed **viewpoint-robust Imitation Learning (VIL)** framework eliminates the need for augmentation or viewpoint calibration. By combining handheld gripper demonstrations with novel vision encoders and inpainting-based alignment module, VIL achieves robust, zero-shot generalization under exocentric viewpoint shifts. One may collect data at any exocentric viewpoint and perform inference

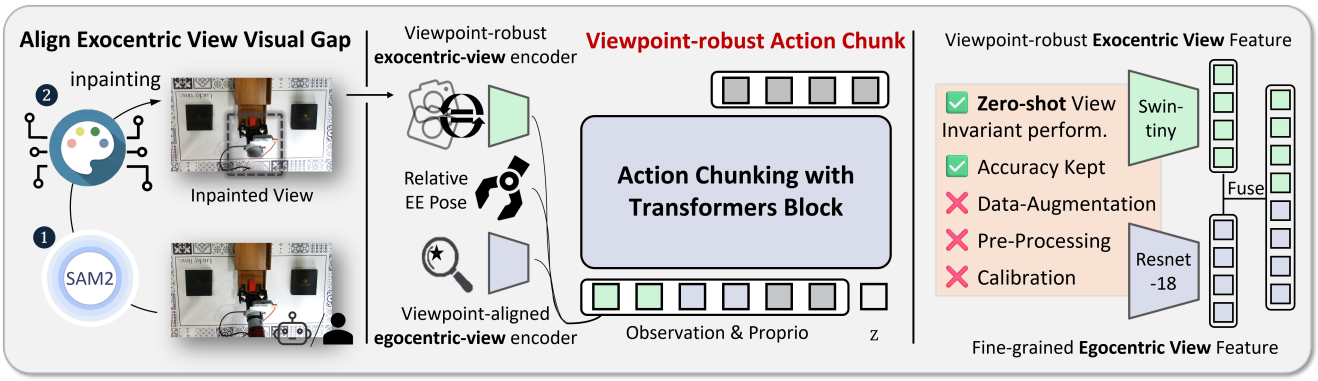

Fig. 2: Algorithmic framework of the proposed zero-shot exocentric viewpoint-robust imitation learning algorithm. Exocentric observations are first processed by a segment-inpainting module (left) to remove manipulator/human components, yielding aligned appearances. Processed exocentric and egocentric views are then passed through a joint feature encoder (right), yielding view-consistent representations. Finally, these features are fed into ACT block to generate viewpoint-robust action.

on novel views without calibration, advancing practical and scalable imitation learning with in-the-wild data collection.

III. Methodology

Zero-shot Exocentric viewpoint-robust Imitation Learning (VIL) is a universal pipeline to integrate handheld data collector with exocentric-egocentric-complementary visuomotor policy learning. It addresses the incompatibility of previous handheld data collectors with exocentric views, thereby enabling the capture of both global context and fine-grained details, and achieving accurate, functional, and stable exocentric viewpoint-robust policy generation. Leveraging the handheld gripper and viewpoint-robust capability, our pipeline is manipulator-independent, can be readily deployed with a Robotiq-2F-85 gripper from arbitrary exocentric viewpoints without fine-tuning.

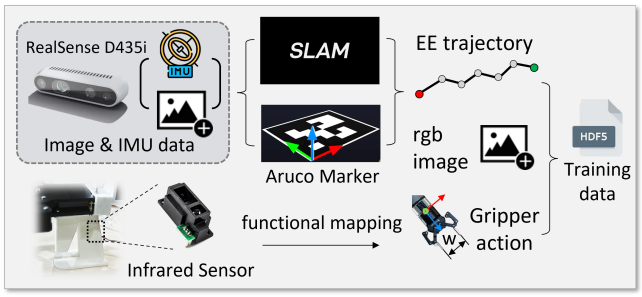

Fig. 3: Data collection pipeline for our handheld gripper.

A. Hardware Modifications

Our handheld gripper pipeline is inspired by UMI, but follows the template of the Robotiq 2F-85, a mature universal gripper that is widely adopted in robotics research. This choice facilitates compatibility with existing large-scale datasets, such as RoboSet [35] and DROID [36], thereby enabling future pretraining or co-training with policies developed on the same gripper hardware. To adapt the design for the Robotiq-like structure and exocentric view integration, we introduce several modifications. First, instead of attaching ArUco markers to the gripper fingertips for width measurement, we embed an infrared sensor in the handle to monitor trigger motion, which is then mapped to gripper action. This preserves gripper structure fidelity while maintaining visual correspondence between the handheld collector and the real gripper. Second, by leveraging third-person view camera to provide global scene context, we eliminate the need for fisheye cameras. Here we employ the Intel RealSense D435i, which offers both reliable synchronization and convenient deployment. More detailed specifications are shown in Appendix.A. The overall Robotiq-like handheld gripper is shown in Fig. 1. It weighs 550 g, with dimensions 300 mm × 155 mm × 255 mm and a finger stroke of 80 mm. Most components are 3D-printed to reduce cost, while the metal fingertips are retained to preserve tactile fidelity. The design is compatible with any robot arm using Robotiq gripper and streams images and gripper action at 30 Hz for efficient visuomotor data collection.

During data collection, we observed that ORB-SLAM3 [37] for trajectory tracking becomes fragile under low-context observations, especially when the gripper approaches surfaces or objects. UMI mitigates this with a fisheye wrist camera, capturing surrounding context to add background cues. However, such cues are unnecessary for policy learning and risk inducing background-dependent bias. To enhance robustness, we further leverage the added exocentric view to track an ArUco marker mounted on the gripper, providing complementary trajectory estimates. Specifically, SLAM performs reliably when the gripper is distant from the target, whereas the ArUco marker is more accurate during close approaches, making the two sources naturally complementary. The detailed data collection pipeline is shown in Fig. 3. Since only relative poses are used for training, exocentric view calibration is unnecessary. In the current setup, the ArUco marker trajectory is used to fill missing parts of SLAM, more advanced fusion strategies could be applied for improvement.

B. Viewpoint-robust Policy Architecture

Leveraging the proposed gripper and dual-view observations, we further propose an imitation learning algorithm that generalizes across exocentric viewpoint shifts. We build the

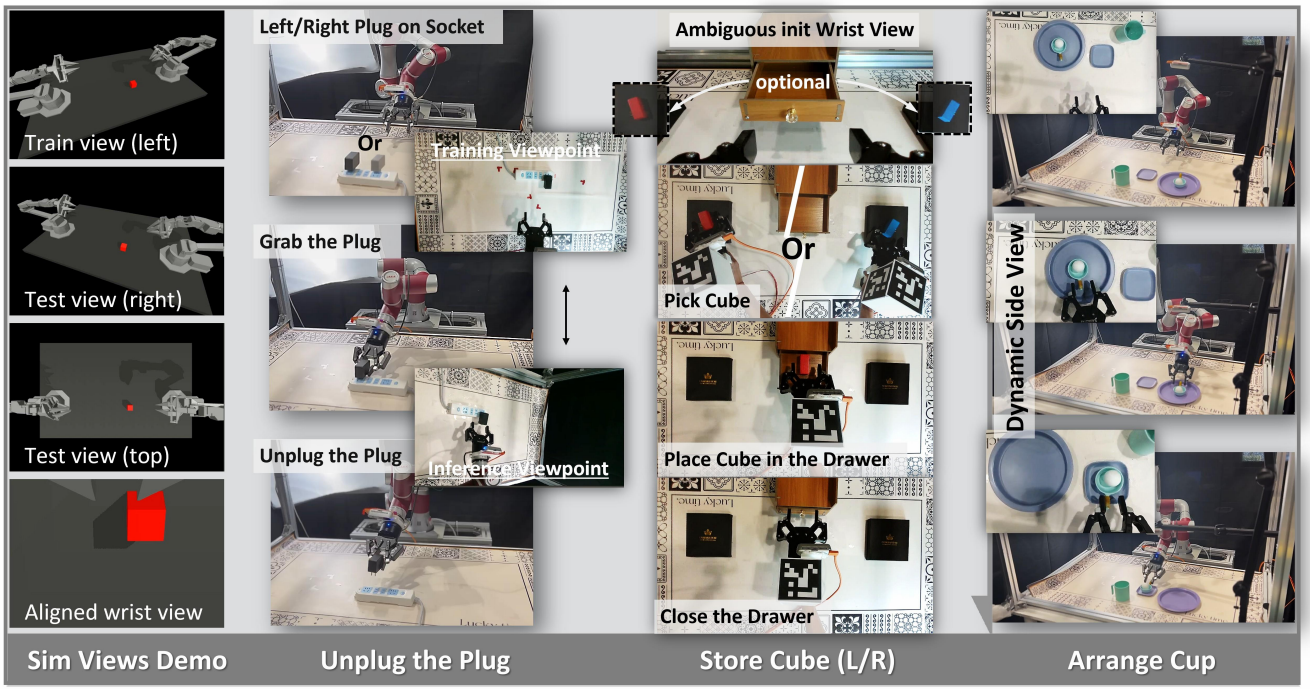

Fig. 4: Experimental settings in simulation and the real-world tasks. Left: different viewpoints used in simulation for viewpoint-robust evaluation. Right: real-world tasks with detailed steps, including unplugging plug under exocentric view changes, storing cube with an ambiguous initial state, and arranging cup under dynamic exocentric views.

algorithm of VIL upon ACT, retaining its strengths to handle multi-view inputs and generate long-horizon planning while substantially improving generalization across viewpoints and ensuring seamless integration with our handheld gripper. VIL contributes two key components: (i) an exocentric–egocentric complementary feature encoder that captures both fine-grained and global cues in a viewpoint-robust manner, and (ii) a visual alignment module that leverages modern computer vision techniques to bridge appearance gaps between human-collected and robot-executed exocentric views.

1) Viewpoint-robust Feature Encoder: Current robotic imitation learning pipelines primarily adopt either ResNet [38] or Vision Transformer (ViT) [39] backbones for image embedding. While both architectures have demonstrated strong representation capacity, they often struggle to generalize across diverse viewpoints, leading to performance degradation when deployment viewpoints deviate from training. After systematic evaluation, we surprisingly find that the Swin Transformer [15] exhibits unexpectedly strong zero-shot viewpoint generalization, significantly outperforming ResNet and ViT in our setting. We attribute this effect to the architectural choices of Swin Transformer. Its use of relative positional bias and shifted-window attention reduces sensitivity to absolute pixel coordinates, enabling the encoder to capture the overall structure of the environment and reason about object state within a shifted scene, rather than focusing solely on local appearance. Furthermore, its hierarchical design integrates local cues with broader global context, potentially reducing its sensitivity to variations in object scale and perspectives. This structure allows the encoder to generalize across the observations at different distances or from shifted viewpoints, thereby enhancing robustness under dynamic view conditions.

While Swin Transformer excels at capturing global scene structure and reasoning across shifted viewpoints, its ability to encode fine-grained details in close-range, egocentric views is limited. To maximize action accuracy for manipulation, we adopt the ResNet-based encoder for the egocentric view, leveraging its strong local feature extraction and low-level inductive bias. This combination (shown in Fig. 2 right) allows our pipeline to benefit from Swin-T's viewpoint robustness for exocentric views while maintaining high-fidelity egocentric representations for precise action generation.

2) Inpainting-based Visual Alignment: Beyond viewpoint variation, a second challenge in leveraging exocentric views for handheld grippers lies in the appearance gap between training and deployment. During data collection, the gripper is manipulated by a human operator while mounted on a robotic arm during inference. This discrepancy in visual appearance introduces the risk of policy instability. Importantly, preserving the fidelity of the gripper itself is critical for policy learning. It provides essential task-related signals for action groundings, whereas the manipulator (human hand or robot arm) contributes little useful information and may even serve as a distractor. Motivated by this insight, we adopt a gripper-centric inpainting module (Fig. 2 left) that removes all non-gripper regions (human hand or robot arm) and replaces them with background. This yields a stable and consistent exocentric view appearance across training and deployment and facilitates cross-manipulator usage.

	ACT(baseline)	DP	ACT(res18&50)	ACT(res&vit)	ACT(all swin)	Ours(res&swin)	ACT(egocentric only)
Sim Transfer Cube (Diverse Exocentric Viewpoints)							
Same	**93.3**	76.7	91.7	90.0	88.3	91.7	83.3
Symmetric	21.7	11.7	33.3	71.7	70.0	**83.3**	N/A
Novel	3.3	18.3	23.3	1.7	76.7	**93.3**	N/A
Sim Insertion (Diverse Exocentric Viewpoints)							
Same	50.0	43.3	36.7	33.3	43.3	**56.7**	46.7
Symmetric	13.3	8.3	28.3	6.7	35.0	**51.7**	N/A
Novel	1.7	13.3	20.0	5.0	33.3	**53.3**	N/A

TABLE I: Success rates (%) in two simulated tasks comparing VIL with benchmark algorithms and design alternatives across three viewpoint generalization stages: aligned, symmetric and novel. Results are averaged over 3 seeds, 20 trials each.

To implement this, we leverage recent advances in foundation and general-purpose inpainting models. Specifically, we employ SAM2 [40] for fine-grained segmentation, prompted by bounding boxes provided by a fine-tuned DEIM model [41], which is used only during the initialization phase to localize manipulator regions. Once a valid manipulator mask is obtained, SAM2 switches to tracking mode for temporal consistency. During tracking, the pipeline periodically re-enters the initialization phase (every 2.5 seconds empirically), where DEIM provides refreshed bounding boxes to reinitialize SAM2, mitigating error accumulation. The resulting masked images are processed by the E2FGVI inpainting model [42] to generate aligned, gripper-only exocentric views for policy learning, while failures in either detection or segmentation simply bypass the inpainting stage. This design ensures generalization without requiring costly manipulator-specific adaptation and operates with low overhead. Tests show that with these modules, the full policy sustains a 10 Hz inference frequency on an RTX 3090.

IV. Evaluations

To evaluate the effectiveness of our proposed zero-shot viewpoint-robust imitation learning pipeline, we design two sets of experiments. First, to isolate algorithm performance from real-world noise and distractions, we adopt two simulation tasks introduced in ACT, which are cube transfer and bimanual insertion. These tasks provide a controlled and fair comparison stage to demonstrate the superior zero-shot generalization of our algorithm under novel viewpoints. For each task, we collect 50 demonstrations with rendered wrist and left-side views for training. During testing, policies are evaluated under three progressively challenging viewpoint conditions: (i) the same viewpoint as training, (ii) the symmetric right-side view, and (iii) a completely novel top-down view (Fig. 4, left). This enables a systematic analysis of viewpoint generalization across other SOTA methods and alternative design choices. Second, we conduct three real-world experiments to evaluate the full pipeline and answer the following key questions: (1) Does the zero-shot exocentric viewpoint generalization observed transfer to handheld-gripper data and noisy real-world environment? (2) Does additional exocentric view observation provides complementary global cues, rather than merely reinforcing redundant static features? (3) What are the limits of VIL—specifically, can it maintain stable behavior under dynamic and continuously changing camera views? The detailed real-world task settings are shown in Fig. 4 right.

A. Simulation Experiments

The simulation tasks results are summarized in Table I. Here we average the performance of each experiment with 3 random seeds and 20 trials each. Compared with state-of-the-art algorithms such as Action Chunking with Transformers (ACT) and Diffusion Policy (DP), our proposed VIL algorithm achieves competitive performance under aligned inference views while significantly outperforming them under viewpoint shifts. Notably, VIL maintains stable performance even under completely novel top views in both the cube transfer (93.3%) and insertion tasks (53.3%), highlighting its ability to achieve zero-shot, viewpoint-robust policy execution. We also compare against an egocentric-only baseline, and the result indicates that without exocentric view input, the policy's success rate degrades to 83.3% and 46.7%, respectively. This demonstrates that view-consistent features captured from the exocentric view provide complementary global information, which strengthens policy learning and improves overall task performance.

To further validate our choice of Swin-T tiny as the exocentric view encoder, we compare it with alternative encoders of similar size, namely ResNet-50 and Vision Transformer-Small (ViT-S), within the ACT baseline. Results show that while ResNet-50 and ViT-S offer limited generalization when test views resemble training data, they completely fail under novel perspectives (all below 25%). In contrast, Swin Transformer achieves consistent performance along viewpoint changes, demonstrating the strongest generalization and thereby justifying this adoption. Finally, we test whether applying Swin-T to the egocentric view encoder can further improve robustness. However, using Swin-T for both encoders results in degraded performance: despite similar training and test losses, grasping accuracy at inference time drops markedly. This suggests that while Swin-T excels at modeling global, viewpoint-robust scene structure, it is less effective at extracting the fine-grained

Method	Viewpoint	Left Hand	Right Hand	Overall
Ours	Same view	10/10	7/10	85%
	Diff view	10/10	6/10	80%
Act	Same view	8/10	8/10	80%
	Diff view	0/10	4/10	20%
DP	Same view	10/10	5/10	75%
	Diff view	7/10	3/10	50%

TABLE II: Socket plug task success rates across methods and views. Left/right columns indicate successes out of 10 trials each and overall shows the total success rate.

task details required for precise egocentric view control. This limitation is marginal with aligned exocentric view where the policy can lean heavily on this anchored perspective for compensation, but becomes critical under exocentric viewpoint changes where accurate egocentric view features become equally indispensable to provide fine corrections, resulting in over 10/20 success decrease. This finding makes the hybrid design—using ResNet for egocentric view fine-grained features and Swin-T for exocentric viewpoint-robust global features—a natural and effective choice.

B. Real-World Evaluation

To further answer the question proposed, we conduct three real-world tasks to verify that our proposed VIL pipeline with added exocentric views generates viewpoint-robust, functional and stable policy in the real world. We use our own designed handheld gripper for data collection, select 50 valid demonstrations each for training and use JAKA-Zu3 robot arm as the manipulator during inference.

1) Unplug the Plug: This task requires unplugging a plug already inserted in a socket, where the plug may be located in either the left or right port. As shown in Fig. 4, we train and evaluate the policy under different exocentric view configurations to assess its viewpoint-robust performance. This setup specifically tests the pipeline's ability to generalize to novel viewpoints in a zero-shot manner, keeping extremely high precision: even slight errors in grasping position can create uneven forces on the plug, causing it to jam tightly in the socket and resulting in task failure.

The results are shown in Table II. For fair comparison, all baselines are provided with the same inpainted visual inputs. While state-of-the-art methods such as ACT and Diffusion Policy achieve comparable performance to VIL under matched exocentric views, their success rates drop sharply when the evaluation viewpoint differs from training, consistent with simulation results. In contrast, our proposed VIL sustains a robust 80% success rate with only marginal degradation, highlighting its effectiveness in reliable exocentric viewpoint-robust policy execution in real-world settings.

2) Store Cube: The robot is tasked with a long-horizon operation: pick up a cube placed on the left or right side, place it into a drawer and close the drawer (Fig 4). To assess the effectiveness of exocentric view input, we intentionally set the egocentric view in an ambiguous initial state where neither cube is visible. Thus, the robot must rely on global information from the exocentric view to determine the cube's location and orient itself accordingly. The challenge arises since, with egocentric view–only input, the policy may be confused by the ambiguous starting state or by visually similar intermediate states (e.g., during transfer or before closing the drawer). In such cases, exocentric view observations are necessary to disambiguate the scene, provide global task context, and guide the policy through the correct task steps.

To better validate the contribution of exocentric view input, we train two models: our vanilla VIL and a variant without exocentric input. Results show that ours achieves 26/30 successes in the pick-and-place step and 14/30 full task completions including drawer closure. In contrast, the exocentric view–free variant degenerates to a shortcut policy, consistently skipping the pick-and-place step and only learning to close the drawer, resulting in an overall 0/30 success rate. The failure suggests that egocentric observations alone introduce state ambiguity, preventing the policy from reliably tracking task progression. These results show that exocentric view observations are not merely auxiliary but provide essential global context, which enables VIL to sustain coherent task execution across stages, underscoring the necessity and effectiveness of incorporating exocentric view input for robust imitation learning.

3) Arrange Cup: In the final task, the robot is required to transfer a cup from the desk to a coaster, with performance evaluated under a dynamically changing exocentric view captured by a human-held tracking camera (see Fig. 4 right). To better visualize disturbances, an additional white ball is placed on the desk to reveal action vibrations. This setting requires the capability to handle continuous viewpoint shifts in a zero-shot manner, which, to our knowledge, has not been reported in prior robot learning pipelines. Despite the challenges of dynamic, noisy, and turbulent viewpoints, our proposed VIL achieves a 7/10 success rate, demonstrating its stability and robustness even under continuously shifting visual conditions. The experiment demonstration can be found in the accompanying video.

C. Ablation Study

This section further validates the effectiveness of two components in our pipeline on the long-horizon store-cube task. Removing the inpainting-based visual alignment module results in a 6.7% (2/30) drop in overall success rate, where failures consistently exhibit jittering gripper motions, confirming its contribution to consistent policy execution. For trajectory tracking, the UMI-style SLAM baseline achieves 48/64 successful demonstrations, whereas incorporating ArUco-marker tracking raises this to 55/64, underscoring its utility for more reliable data acquisition.

V. Conclusion

We present a zero-shot exocentric viewpoint-robust imitation learning framework (VIL), comprising a universal handheld gripper for data collection and an exocentric-egocentric-complementary visuomotor policy that leverages

the viewpoint-robust capability of Swin Transformer. The coordination between these two parts allows us to leverage the exocentric view's essential global context without complicating the real-world deployment for manipulation policy trained with scalable, low-cost data. Skills learned using our handheld gripper can be transferred to manipulators equipped with Robotiq gripper without aligned views or policy fine-tuning, enabling scalable data acquisition and seamless deployment in diverse environmental settings. Future work will explore the integration of our hybrid encoder with other visuomotor policies, such as Diffusion Policy. Additionally, multi-viewpoint mixed training within a single task—though not validated due to time constraints—represents a promising direction for scalability. Moreover, mitigating the effects of egocentric viewpoint variation remains an open challenge, requiring further investigation.

APPENDIX

A. Hardware

1) Hardware and Synchronization: As described in the methodology, we employ two RealSense D435i cameras (exocentric and egocentric) for visual observations and a GP2Y0A51SK0F infrared distance sensor for gripper action signals. Both images and distance signals are recorded at 30 Hz. For synchronization, the RealSense cameras support built-in triggering for hardware-level alignment. Since ArUco marker detection and SLAM rely on image streams, they are naturally synchronized, and the distance sensor signals are interpolated based on image timestamps to construct a synchronized dataset.

2) Temporal Smoothing with Forward Rescaling: During inference, we apply temporal smoothing with forward rescaling to ensure predicted gripper actions comply with hardware limits. The Robotiq-2F-85 gripper supports a maximum jaw speed of 0.1 m/s, whereas our handheld prototype has no such restriction, allowing humans to generate much faster motions. As a result, the learned policy may output commands that exceed the Robotiq's limit. To resolve this, we smooth and rescale the predicted actions in advance, ensuring executed trajectories remain within feasible speed bounds.

3) Inferred Sensing for Adaptive Gripper Width Estimation: The handheld gripper adopts an underactuated, spring-based self-adaptive mechanism that enables both parallel and enveloping grasps, consistent with the original Robotiq design (Fig. 5). While this improves dexterity, it renders direct visual measurement of gripper width unreliable. Instead, we estimate the width from inferred sensing by tracking the trigger's position. Further empirical tests reveal an approximately linear relationship between sensor output and gripper opening, allowing calibration using only fully open and closed states, with linear interpolation for real-time estimation.

B. Training Setups and Implementation Details

In our framework, ResNet18 is used for egocentric view feature extraction and Swin-Tiny for exocentric view. Both are initialized with torchvision checkpoints and frozen to stabilize training. Policies are trained for 4000 epochs in simulation and 8000 epochs in real-world tasks, using 480×640 input images, batch size 32, and learning rate 3×10^{-5} on a single RTX 3090 GPU. To improve inference speed, only a single timestamp image is used.

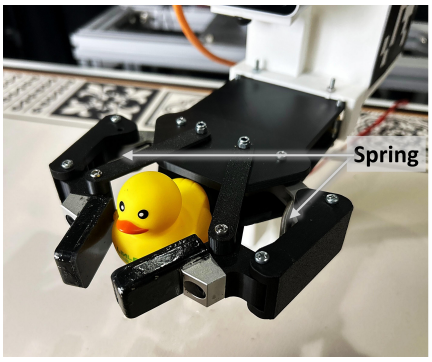

Fig. 5: Hand gripper with self-adaptive mechanism.

For the segmentation and inpainting module, we first train a DEIM model to generate bounding-box prompts for SAM2. We collect 10 videos, apply SAM2 with manual correction, and sample every 10 frames, yielding about 300 training frames. We then fine-tune the DEIM backbone with these frames and integrate it with SAM2 for initialization. Since SAM2 offers both initialization and tracking modes, we run initialization every 2.5 seconds and apply tracking in between to balance accuracy and efficiency. After segmentation, both the mask and original image are downsampled to 240×320 and passed into the E2FGVI inpainting model [14] for fast inference. The inpainted regions are then used to fill the masked areas, providing clean robot-arm-removed images.

C. Experiment Details

For simulation tasks (cube transfer and insertion), datasets are collected with absolute joint-space actions as supervision. Since training and inference share the same visual domain, the inpainting-based visual alignment module is not applied. We adopt the model with a chunk size of 100 and utilize the first 35 steps for execution during inference. The comparison baselines (ACT and Diffusion Policy) are trained with the same joint-space supervision for fair comparison and under the default configuration provided. Notably, Diffusion Policy suffers from noisy gripper actions, which we mitigate via gripper action filtering.

For real-world experiments, we follow UMI in using relative pose supervision and apply the inpainting module for both training and inference time preprocessing. We set the chunk size to 30 and select 20 steps to execute following their timestamps. Although our policy supports inference at 10 Hz, we query actions every 0.5 s to improve consistency. Specifically, we do not apply the inpainting module in the arrange cup task because dynamic viewpoints can lead to unstable inpainting performance.

REFERENCES

[1] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware," *Robotics: Science and Systems*, 2023.

[2] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song, "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," in *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[3] S. Kareer, D. Patel, R. Punamiya, P. Mathur, S. Cheng, C. Wang, J. Hoffman, and D. Xu, "Egomimic: Scaling imitation learning via egocentric video," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[4] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel, "Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024.

[5] J. Wang, C.-C. Chang, J. Duan, D. Fox, and R. Krishna, "Eve: Enabling anyone to train robots using augmented reality," in *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, 2024.

[6] S. Chen, C. Wang, K. Nguyen, L. Fei-Fei, and C. K. Liu, "Arcap: Collecting high-quality human demonstrations for robot learning with augmented reality feedback," *arXiv preprint arXiv:2410.08464*, 2024.

[7] H. Fang, H.-S. Fang, Y. Wang, J. Ren, J. Chen, R. Zhang, W. Wang, and C. Lu, "Low-cost exoskeletons for learning whole-arm manipulation in the wild," in *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition@ CoRL2023*, 2023.

[8] H. Etukuru, N. Naka, Z. Hu, S. Lee, J. Mehu, A. Edsinger, C. Paxton, S. Chintala, L. Pinto, and N. M. M. Shafiullah, "Robot utility models: General policies for zero-shot deployment in new environments," *arXiv preprint arXiv:2409.05865*, 2024.

[9] M. Seo, H. A. Park, S. Yuan, Y. Zhu, and L. Sentis, "Legato: Cross-embodiment imitation using a grasping tool," *IEEE Robotics and Automation Letters*, 2025.

[10] H. Ha, Y. Gao, Z. Fu, J. Tan, and S. Song, "Umi on legs: Making manipulation policies mobile with manipulation-centric whole-body controllers," *arXiv preprint arXiv:2407.10353*, 2024.

[11] 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.

[12] 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, 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*, 2024.

[13] S. Xia, H. Fang, C. Lu, and H.-S. Fang, "Cage: Causal attention enables data-efficient generalizable robotic manipulation," *arXiv preprint arXiv:2410.14974*, 2024.

[14] L. Y. Chen, C. Xu, K. Dharmarajan, M. Z. Irshad, R. Cheng, K. Keutzer, M. Tomizuka, Q. Vuong, and K. Goldberg, "Roviaug: Robot and viewpoint augmentation for cross-embodiment robot learning," in *Conference on Robot Learning (CoRL)*, 2024.

[15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

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

[17] J. Duan, Y. R. Wang, M. Shridhar, D. Fox, and R. Krishna, "Ar2-d2: Training a robot without a robot," in *Proceedings of The 7th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, J. Tan, M. Toussaint, and K. Darvish, Eds., 2023.

[18] N. Nechyporenko, R. Hoque, C. Webb, M. Sivapurapu, and J. Zhang, "Armada: Augmented reality for robot manipulation and robot-free data acquisition," *arXiv preprint arXiv:2412.10631*, 2024.

[19] Y. Park, J. S. Bhatia, L. Ankile, and P. Agrawal, "DexHub and DART: Towards Internet Scale Robot Data Collection," 2024.

[20] H. Fang, C. Wang, Y. Wang, J. Chen, S. Xia, J. Lv, Z. He, X. Yi, Y. Guo, X. Zhan, *et al.*, "Airexo-2: Scaling up generalizable robotic imitation learning with low-cost exoskeletons," *arXiv preprint arXiv:2503.03081*, 2025.

[21] H. Kim, Y. Ohmura, A. Nagakubo, and Y. Kuniyoshi, "Training robots without robots: deep imitation learning for master-to-robot policy transfer," *IEEE Robotics and Automation Letters*, 2023.

[22] N. M. M. Shafiullah, A. Rai, H. Etukuru, Y. Liu, I. Misra, S. Chintala, and L. Pinto, "On bringing robots home," *arXiv preprint arXiv:2311.16098*, 2023.

[23] S. Young, D. Gandhi, S. Tulsiani, A. Gupta, P. Abbeel, and L. Pinto, "Visual imitation made easy," in *Conference on Robot learning*, 2021.

[24] K. Doshi, Y. Huang, and S. Coros, "On hand-held grippers and the morphological gap in human manipulation demonstration," *arXiv preprint arXiv:2311.01832*, 2023.

[25] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. K. Liu, "Dexcap: Scalable and portable mocap data collection system for dexterous manipulation," *arXiv preprint arXiv:2403.07788*, 2024.

[26] A. Lee, I. Chuang, L.-Y. Chen, and I. Soltani, "Interact: Inter-dependency aware action chunking with hierarchical attention transformers for bimanual manipulation," *arXiv preprint arXiv:2409.07914*, 2024.

[27] H. Xiong, X. Xu, J. Wu, Y. Hou, J. Bohg, and S. Song, "Vision in action: Learning active perception from human demonstrations," *arXiv preprint arXiv:2506.15666*, 2025.

[28] I. Chuang, A. Lee, D. Gao, M.-M. Naddaf-Sh, and I. Soltani, "Active vision might be all you need: Exploring active vision in bimanual robotic manipulation," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 7952–7959.

[29] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," in *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

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

[31] Z. Qian, M. You, H. Zhou, X. Xu, and B. He, "Robot learning from human demonstrations with inconsistent contexts," *Robotics and Autonomous Systems*, 2023.

[32] A. Zhou, M. J. Kim, L. Wang, P. Florence, and C. 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*, 2023.

[33] Y. Seo, J. Kim, S. James, K. Lee, J. Shin, and P. Abbeel, "Multi-view masked world models for visual robotic manipulation," in *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[34] X. Zhang, M. Chang, P. Kumar, and S. Gupta, "Diffusion meets dagger: Supercharging eye-in-hand imitation learning," *arXiv preprint arXiv:2402.17768*, 2024.

[35] H. Bharadhwaj, J. Vakil, M. Sharma, A. Gupta, S. Tulsiani, and V. Kumar, "RoboAgent: Generalization and Efficiency in Robot Manipulation via Semantic Augmentations and Action Chunking," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024.

[36] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," *arXiv preprint arXiv:2403.12945*, 2024.

[37] C. Campos, R. Elvira, J. J. Gomez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM," *IEEE Transactions on Robotics*, 2021.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[39] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[40] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[41] S. Huang, Z. Lu, X. Cun, Y. Yu, X. Zhou, and X. Shen, "Deim: Detr with improved matching for fast convergence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[42] Z. Li, C.-Z. Lu, J. Qin, C.-L. Guo, and M.-M. Cheng, "Towards an end-to-end framework for flow-guided video inpainting," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.