Touch begins where vision ends: Generalizable policies for contact-rich manipulation

Zifan Zhao[1] **Siddhant Haldar**[2] **Jinda Cui**[3] **Lerrel Pinto**[2] **Raunaq Bhirangi**[2*]

[1]New York University Shanghai [2]New York University [3]Honda Research

vitalprecise.github.io

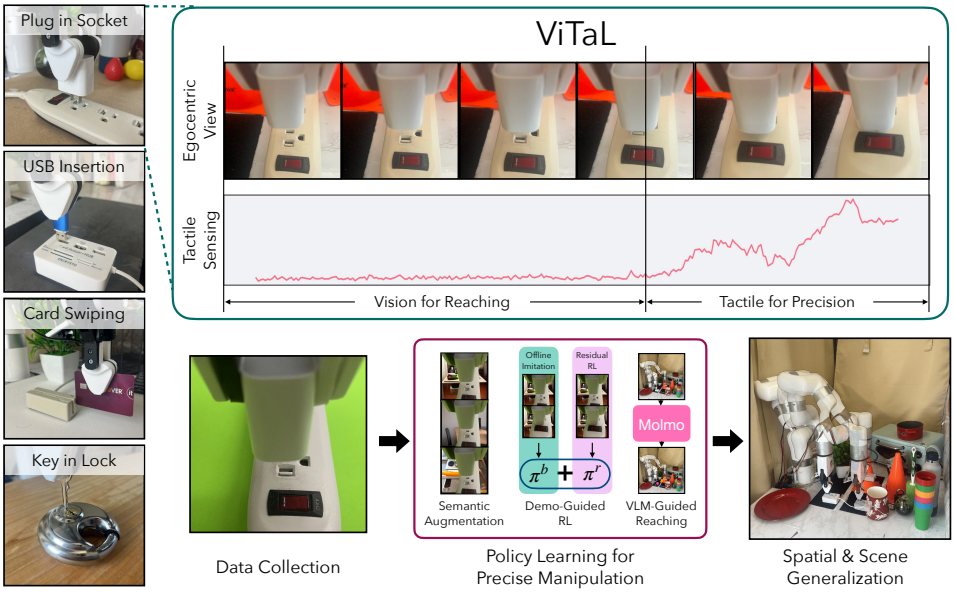

Figure 1: We present VITAL, a framework that combines tactile sensing, vision foundation models, and residual reinforcement learning to enable learning policies that can operate with millimeter-level precision while generalizing to large spatial variations and significant environmental perturbations.

Abstract:

Data-driven approaches struggle with precise manipulation: imitation learning requires many hard-to-obtain demonstrations, while reinforcement learning yields brittle, non-generalizable policies. We introduce VisuoTactile Local (VITAL) policy learning, a framework that solves fine-grained manipulation tasks by decomposing them into two phases: a *reaching phase*, where a vision-language model (VLM) enables scene-level reasoning to localize the object of interest, and a local *interaction phase*, where a reusable, scene-agnostic VITAL policy performs contact-rich manipulation using egocentric vision and tactile sensing. This approach is motivated by the observation that while scene context varies, the low-level interaction remains consistent across task instances. By training local policies once in a canonical setting, they can generalize via a localize-then-execute strategy. VITAL achieves ∼90% success on contact-rich tasks in unseen environments and is robust to distractors. VITAL's effectiveness stems from three key insights: (1) foundation models for segmentation enable training robust visual encoders via behavior cloning; (2) these encoders improve the generalizability of policies learned using residual RL; and (3) tactile sensing significantly boosts

*Correspondence to: raunaqbhirangi@nyu.edu

Workshop on Safe and Robust Robot Learning for Operation in the Real World at CoRL 2025.

performance in contact-rich tasks. Ablation studies validate each of these insights, and we demonstrate that VITAL integrates well with high-level VLMs, enabling robust, reusable low-level skills. Results and videos are available at vitalprecise.github.io.

Keywords: visuotactile, residual learning, local policy

1 Introduction

Imitation learning for sensorimotor skills has made significant strides in recent years, propelled by the increasing scale and diversity of robotic datasets. From controlled tabletop environments to open-world household settings, large-scale data has been shown to improve the generalization of robotic policies [1, 2] – mirroring advances in vision [3, 4, 5] and language [6, 7, 8]. However, precise, contact-rich manipulation poses a significant challenge to this data-centric approach. Fine-grained tasks such as inserting USBs and swiping credit cards have low error tolerance (millimeter to sub-millimeter), and the high fidelity required makes demonstration collection time-consuming, brittle, and difficult to scale. Deep reinforcement learning (RL) provides an alternative by learning directly through online interaction, but often sacrifices generalization in favor of narrowly tuned policies sensitive to training-specific cues like scene layout or background distractors.

In this work, we propose VisuoTactile Local (VITAL), a policy learning framework that bridges this gap by enabling robust, precise manipulation while maintaining generalizability. VITAL decomposes manipulation into two phases: a global *reaching phase*, where a vision-language model (VLM) performs scene-level reasoning to identify and localize the object of interest, and a local *interaction phase*, where a reusable, scene-agnostic policy performs fine-grained, contact-rich manipulation using egocentric vision and tactile sensing. This decomposition is motivated by the observation that while the environmental context for a task may vary drastically, the low-level physical interactions required for manipulation remain consistent. Our work focuses on capturing this invariant local policy: training it once in a canonical setting allows it to generalize across environments via a simple localize-then-execute strategy. With just 32 demonstrations and 45 minutes of online reinforcement learning per task, VITAL achieves the precision necessary for real-world deployment while maintaining adaptability across scenes.

A core design motivation behind VITAL is the deliberate pairing of sensing modalities with complementary strengths. Tactile sensing is indispensable during contact-rich phases of manipulation, providing direct, localized feedback about forces and slip, that cannot be captured by vision. It is inherently robust to lighting, background clutter, and occlusion, but lacks the spatial awareness necessary for planning and coarse alignment in the pre-contact phase. Egocentric vision fills this gap by offering a consistent, robot-centered perspective that captures the relative pose of the end-effector and surrounding objects. Unlike third-person or fixed external cameras, egocentric views are naturally aligned with the robot's actions and are easy to replicate across different environments without introducing viewpoint-specific biases that can severely hinder learned policy transfer.

While visuotactile design is not novel in itself, existing works typically fail to use it effectively. Imitation learning methods require large, diverse datasets [9, 10] to handle spatial and scene variation, making them expensive and difficult to scale, especially for precise manipulation. Reinforcement learning is capable of refining policies through interaction, but tends to overfit to training environments [11, 12]. A key reason for this is that learning from raw RGB inputs in constrained settings lacks the visual diversity needed for generalization. Without sufficient variation in appearance, background, and lighting, policies trained via RL become brittle and environment-specific.

VITAL addresses this limitation with a key insight: task success depends primarily on the visual features of task-relevant objects, which remain relatively stable across environmental changes. To exploit this invariance, we introduce a semantic, task-aware data augmentation pipeline powered by vision foundation models. These augmentations introduce altering distractors, backgrounds, and lighting, while preserving object and robot identity. This allows visual encoders to learn more

general representations from the same amount of demonstration data, eliminating the need for costly scene variations in data collection.

Finally, to further improve performance and address the inevitable imperfections in teleoperated demonstrations, we fine-tune our policies using offset-based reinforcement learning. Rather than learning policies from scratch, we apply DrQ-v2 [13] to refine behavior-cloned policies by predicting small corrective actions, or offsets, relative to the predicted actions. Crucially, this refinement is done without discarding the visual generalization learned during imitation, as we continue to apply semantic augmentations during online training. This final phase boosts precision and robustness while preserving the broad generalization enabled by our visuotactile design and augmentation strategy.

Our key findings can be summarized as follows:

1. VITAL learns generalizable, contact-rich manipulation policies with a 90% success rate from just 32 demonstrations and 45 minutes of interaction, outperforming the best baseline by 40% on average across four challenging precise manipulation tasks in unseen environments.
2. Tactile sensing is essential for precision and reliability: removing tactile input reduces success rates by an average of 40%, underscoring its critical role in contact-rich task phases where vision alone is insufficient.
3. VITAL extends the benefits of semantic visual augmentation beyond imitation learning by combining it with residual RL, enabling policy fine-tuning without sacrificing generalization.

All of our datasets, and training and evaluation code have been made publicly available. Videos of our trained policies can be seen here: vitalprecise.github.io.

2 VITAL

The core insight behind VITAL is that vision offers task-level spatial awareness for scene generalization, while tactile sensing is essential for millimeter-scale precision during physical contact. By leveraging the strength of each modality, our method enables policies to be trained in localized settings and deployed across diverse spatial variations and background configurations. VITAL operates in three phases: (1) Visuotactile behavior cloning learns a generalizable base policy using visual semantic augmentations; (2) Residual RL enhances downstream performance by optimizing policy refinements while maintaining vision-driven robustness; (3) VLM-based reaching facilitates zero-shot adaptation to novel spatial configurations by identifying actionable regions and decoupling task dynamics from environment configuration. Our pipeline has been illustrated in Figure 2.

2.1 Generalizable behavior cloning through semantic augmentations

Our method starts by collecting visuotactile robot demonstrations using a virtual reality (VR) based teleoperation framework [14]. All the tasks presented in this paper consist of a target object on the table that the robot interacts with, and a grasped object that is held within the robot gripper. We hypothesize that for most precise manipulation tasks, the core interaction dynamics remain consistent across task instances, despite variations in the broader environment, ie., the dynamics of plugging your charger in the kitchen are consistent with the dynamics of plugging your charger in the bedroom. To focus data collection on these invariant interactions, we fix the target object position, and collect successful demonstrations with the robot randomly initialized in the vicinity of the target object. We ensure that observations and actions are grounded in the robot's end-effector frame to enable transfer to novel spatial configurations during inference. This is achieved by using the wrist camera image and tactile readings as input observations, and computing relative actions in the robot's end effector frame. By constraining spatial variability and focusing on local interaction patterns, our method achieves robust policy learning with only 32 demonstrations per task.

To maintain policy performance across variations in the visual environment, we implement semantic augmentations targeting visual regions irrelevant to the task. Our collected demonstrations use

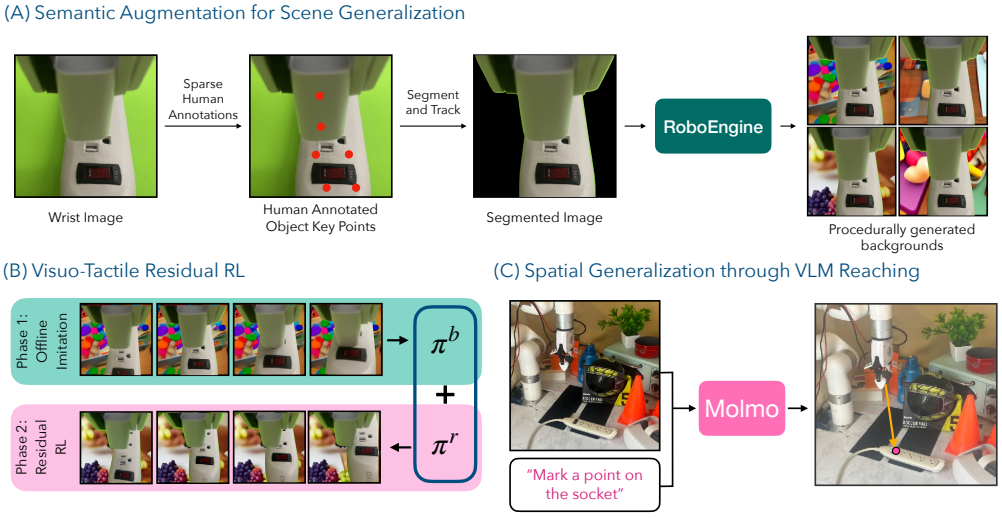

Figure 2: Overview of VITAL. (A) VITAL utilizes vision foundation models to enhance task data with procedurally generated backgrounds, improving visual diversity. (B) This data is then used to train a generalizable visuo-tactile policy, which is later refined through online residual reinforcement learning (RL) for precision. (C) Finally, VLM-guided reaching enables zero-shot deployment in novel spatial configurations, despite policies being trained on fixed object positions.

a green screen background to facilitate background replacement through procedural scene generation [15] using RoboEngine [16] during policy learning. In our initial experiments, we observed that naive color-key based background filtering performs poorly, which prompted our multi-stage segmentation pipeline: First, a human annotator marks key points on task-relevant objects in a *single* reference demonstration frame. This often requires only a few seconds. Next, DIFT-based correspondence matching [17] propagates these annotations to the first frame of all demonstrations, followed by Segment-Anything 2 [5] for instance segmentation. Finally, XMem [18] tracks the segmented masks temporally along trajectories, separating the relevant task elements from augmentable background regions (Fig. 2). This allows targeted background transformations while preserving contact-relevant visual features critical for tactile coordination.

The demonstration data is then used to train a base visuotactile policy using behavior cloning. The augmented visual data is encoded using a randomly-initialized ResNet-18 [19] encoder, and tactile reading from AnySkin [20] is encoded using a multilayer perception (MLP). The encoded observations are fed into a visuo-tactile transformer policy π^b for action prediction [21]. The policy is trained with action chunking [22] using a mean squared error loss between predicted and ground truth action chunks. By jointly enforcing spatial and visual invariance through semantic augmentations and sensory observations grounded in the end-effector frame, the policy develops robust task understanding decoupled from environmental context.

2.2 Fine-tuning with Demonstration-guided Reinforcement Learning

While the pretrained base policy π^b enables generalizable visuo-tactile policies, we observe that the policy only achieves a modest success rate. To improve the performance of π^b, we employ residual reinforcement learning (RL) to train a residual policy π^r on top of the base policy. In residual RL [23], given a base policy $\pi^b : \mathcal{Z} \to \mathcal{A}$ with encoded representations $z \in \mathcal{Z}$ and action $a \in \mathcal{A}$, we learn a residual policy $\pi^r : \mathcal{Z} \times \mathcal{A} \to \mathcal{A}$ such that an action sampled from the final policy π is the sum of the base action $a^b \sim \pi^b(z)$ and the residual offset $a^r \sim \pi^r(z, a^b)$. Following prior work [24, 25], we use n-step DDPG [26] as our RL optimizer, a deterministic actor-critic based method that provides high performance in continuous control [13].

During online learning, the encoders (ResNet-18 for vision, MLP for tactile) trained for behavior cloning are fixed, and feed compressed representations z^i (image) and z^t (tactile) to both the frozen base policy π_b and the residual actor network, π_r, which takes as input (z^i, z^t, a^b) to predict a^r. Similarly, the residual critic Q^r evaluates (z^i, z^t, a^b, a^r) pairs using layer normalization and high update-to-data (UTD) ratios for sample-efficient Q-learning [27]. Crucially, we observe that adding L2 weight regularization for the actor network improves policy training, resulting in better performance. For RL training, our reward is simply a sum of a binary success reward provided by a human at the end of the trajectory and a dense L1 distance from the goal for the task. The RL training objective is as follows:

$$\pi^r = \underset{\pi^r}{\operatorname{argmax}}\, \mathbb{E}_{(z^i, z^t, a^b, a^r)\sim\mathcal{D}_\beta}\left[Q(z^i, z^t, a^b, a^r)\right] \tag{1}$$

where $\mathcal{D}_\beta$ contains rollouts enriched with the same semantic visual augmentations from the behavior cloning phase to maintain generalization. The executed action a is a sum of a^b and a^r. This approach of combining fixed pretrained features with adaptive residuals improves policy performance while preserving cross-environment robustness through augmentations. Details about hyperparameters and network architectures used in our experiments have been included in Appendix A1.

2.3 Inference

Our framework achieves spatial and scene generalization through a hierarchical inference strategy: global semantic navigation by a high-level agent followed by localized visuotactile control for precise low-level execution. By combining offline behavior cloning and online residual adaptation, the policy operates within a constrained task space while maintaining robustness to environmental perturbations. For global positioning, we employ Molmo [28], a vision-language model (VLM) pretrained on web-scale data, to coarsely localize target objects specified via natural language.

Given an external RGB-D observation, Molmo predicts a 2D coordinate for the target object, which is projected to 3D workspace coordinates using depth data and camera calibration parameters. The robot then samples an initial end-effector pose within a pre-defined region of the target coordinate. For example, for USB insertion, the target point for the robot is sampled at a height of 10cm above the predicted coordinate. Empirically, we observe that this coarse initialization falls within the pretrained policy's operational envelope, ensuring target visibility in the wrist camera feed. Upon reaching the target position, the learned visuotactile local policy is deployed to complete the task. Our results in Section 3 demonstrate the potential of combining general-purpose VLMs for coarse robotic navigation, with localized visuo-tactile policies handling the precise parts of a task.

3 Experiments

Our experiments seek to answer the following questions: (1) How does VITAL perform in an in-domain setting? (2) How does VITAL perform under environmental perturbations? (3) What are the important design choices for VITAL? (4) How well does the VLM navigation work with VITAL?

3.1 Experimental Setup

Our experiments are conducted using a UFACTORY xArm 7 robot equipped with a two-fingered xArm Gripper. For tactile sensing, we integrate the AnySkin [20] magnetic tactile sensor into the gripper. The observations for policy learning include 128×128 RGB images captured by a fisheye camera mounted on the robot's wrist, and 15-dimensional tactile readings from the AnySkin sensor. For coarse navigation via the VLM, we use a calibrated third-person Intel RealSense RGB-D camera. For each task, demonstrations are collected using a VR-based teleoperation system [14] operating at 30 Hz. The collected data is subsampled to 6Hz for policy training, and the learned policies are deployed at 6Hz during real-world execution.

Table 1: Policy performance of VITAL in an in-domain setting.

Method	Plug in Socket	Insert USB	Card Swiping	Key in Lock	Pick Bread
BAKU [29]	4/10	4/10	1/10	5/10	**10/10**
ViSK [21]	7/10	3/10	4/10	5/10	**10/10**
RLPD [30]	3/10	0/10	2/10	1/10	**10/10**
VITAL-BC	7/10	4/10	5/10	5/10	**10/10**
VITAL **(Ours)**	**9/10**	**9/10**	**10/10**	**9/10**	**10/10**

Table 2: Study of spatial and scene generalization in VITAL.

Method	Spatial Generalization				Scene Generalization			
	Plug in Socket	Insert USB	Card Swiping	Key in Lock	Plug in Socket	Insert USB	Card Swiping	Key in Lock
BAKU [29]	15/30	5/30	7/30	11/30	0/30	0/30	0/30	0/30
ViSK [21]	19/30	6/30	15/30	14/30	0/30	0/30	0/30	0/30
RLPD [30]	5/30	0/30	3/30	2/30	0/30	0/30	0/30	0/30
VITAL-BC	21/30	10/30	16/30	12/30	24/30	8/30	13/30	17/30
VITAL	**28/30**	**24/30**	**28/30**	**22/30**	**27/30**	**23/30**	**25/30**	**24/30**

3.2 Task Descriptions

We demonstrate the versatility of our framework by evaluating VITAL on four precise, contact-rich manipulation tasks and a pick bread task. We collect 32 demonstrations for each task while fixing the target object and randomly initializing the robot in a predefined area around it. Detailed task descriptions can be found in Appendix A2.

3.3 Baselines

We demonstrate the versatility of our framework by evaluating VITAL on a pick bread task and four precise, contact-rich manipulation tasks. We compare VITAL with four primary baselines: **BAKU [29]:** Transformer policy for behavior cloning that maps RGB images to robot actions; **ViSk [21]:** BAKU augmented with both RGB images and tactile readings as input; **RLPD [30]:** an RL approach trained from scratch on a 1:1 mix of expert and RL replay buffers; and **VITAL-BC**: our visuotactile base policy employing semantic augmentation within the ViSk architecture. Further details on baseline implementations can be found in Appendix A3.

3.4 How does VITAL perform in an in-domain setting?

Table 1 evaluates VITAL's performance in a controlled in-domain setting, where both the background (green screen) and object positions are fixed. For each method, we conduct 10 trials per task, with the robot randomly initialized within a predefined area around the target object (Section 3.2). Both VITAL and RLPD receive identical visual and tactile observations and are trained online for 45 minutes. While VITAL incorporates semantic augmentations in its RL replay buffer, we find that such augmentations degrade performance for RLPD; therefore, RLPD results in Table 1 do not use semantic augmentations. Our results demonstrate that VITAL significantly outperforms all baselines, achieving an absolute improvement of 40% over the strongest alternative. Notably, ViSk outperforms BAKU, highlighting the importance of tactile sensing for precise manipulation. Further, VITAL surpasses RLPD, emphasizing the value of offline policy pretraining for sample-efficient online learning. Overall, these findings illustrate that visuotactile behavior cloning and residual RL scaffolded by semantic augmentations enables robust, high-precision manipulation.

Table 3: Study of VITAL's robustness to combined spatial and scene perturbations.

Method	Plug in Socket	Insert USB	Card Swiping	Key in Lock
BAKU [29]	0/30	0/30	0/30	0/30
ViSK [21]	0/30	0/30	0/30	0/30
RLPD [30]	0/30	0/30	0/30	0/30
VITAL-BC	24/30	9/30	19/30	13/10
VITAL **(Ours)**	**29/30**	**25/30**	**27/30**	**27/30**

3.5 How does VITAL perform under environmental perturbations?

Spatial Generalization Table 2 evaluates VITAL's spatial generalization by testing three novel target object positions outside the training distribution, with the green screen background retained to isolate spatial variations from scene-level changes. Across 10 trials per position, each initializing the robot within a predefined workspace around the target object, results show comparable performance to in-domain settings, confirming that localized end-effector frame observations effectively enable spatial generalization. Notably, BAKU and ViSk admit a performance decline when target objects approach the edges of the green screen, resulting in background elements entering into the fisheye wrist camera's field of view, inducing visual distribution shifts relative to training data.

Scene Generalization Table 2 assesses VITAL's scene generalization by testing on three novel, cluttered scene configurations (see Appendix A4 for examples) while keeping the target object position fixed and identical to training. For each configuration, we run 10 trials with the robot randomly initialized within a predefined area around the target. The results demonstrate VITAL's robustness to unstructured scene variations, significantly outperforming all baselines. The strong performance of both VITAL and VITAL-BC highlights the critical role of semantic augmentations in enabling policies to disentangle task-relevant visual cues from environmental noise. Moreover, VITAL's improvement over VITAL-BC illustrates how residual RL combined with semantic augmentations substantially enhances performance while preserving VITAL-BC's generality. Table 3 extends this evaluation to scenarios varying both target spatial positions and background appearances. To decouple policy performance from VLM navigation effects, we manually initialize the robot near the target object and conduct 10 trials per position. The results revealing a consistent pattern: VITAL and VITAL-BC outperform baselines, with VITAL maintaining a clear advantage. Overall, the use of localized observation spaces alongside semantic augmentations during training endows VITAL with strong spatial and scene generalization capabilities.

3.6 What are the important design choices for VITAL?

VITAL is an amalgam of several techniques that enable learning generalizable visuo-tactile policies. Here, we systematically ablate several design choices in VITAL and justify their importance.

Tactile sensing Table 4 investigates tactile sensing's role in enabling millimeter-scale precision, with experiments conducted under controlled conditions (fixed object positions, green screen background) to isolate sensory effects. Comparing visual (BAKU) and visuo-tactile (ViSk) BC, both with and without residual RL, reveals a consistent performance advantage with tactile inputs. While visual BC with residual RL is competent on two tasks, utilizing tactile inputs further improves performance. Qualitatively, this improvement stems from visual occlusion challenges: as the end effector approaches the target, the object held by the gripper obstructs the egocentric camera's view of the goal, rendering visual feedback unreliable and causing hesitation or blind actions. Tactile sensing proves indispensable in tasks like *Card Swiping*, where the card occludes the machine and the policy has to heavily rely on tactile sensing for task completion. The results confirm that tactile sensing compensates for dynamic visual obstructions while enabling finer contact-driven adjustments.

Table 4: Study of important design choices for VITAL.

Method	Plug in Socket	USB Insertion	Card Swiping
Tactile Ablations			
Visual BC	4/10	4/10	1/10
Visuo-Tactile BC	7/10	3/10	4/10
Visual BC + Res. RL	**9/10**	7/10	0/10
Semantic Augmentation Ablations			
Visual BC (BAKU)	0/10	0/10	0/10
Visual BC + Aug.	4.7/10	2.3/10	0.7/10
Visuo-Tactile BC (ViSk)	0/10	0/10	0/10
Visuo-Tactile BC + Aug.	8.3/10	3/10	6.3/10
VITAL	**9/10**	**9/10**	**10/10**

Semantic Augmentation Table 4 studies the importance of semantic augmentations for novel scene generalization. We average the performance of visual (BAKU) and visuo-tactile (ViSK) BC – with and without semantic augmentations – across three unseen object positions with background distractors. Our results demonstrate that semantic augmentations enable both approaches to adapt to new spatial and visual conditions, with visuotactile BC achieving superior performance than its visual counterpart.

3.7 How well does the VLM navigation work with VITAL?

Table 5: VLM Navigation for spatial generalization.

Method	Plug in Socket	USB Insertion	Card Swiping
ViSk [21]	0/25	0/25	0/25
VITAL-BC	16/25	9/25	13/25
VITAL **(Ours)**	21/25	16/25	19/25

Table 5 evaluates VLM-based coarse navigation across five novel object positions, conducting five trials per position while including background distractors to test robustness to environmental perturbations. Compared to the strongest baseline, VITAL-BC, we observe that both methods generalize to unseen object positions and maintain consistent performance in cluttered scenes, despite being trained on fixed configurations. This highlights the utility of VLM navigation for imparting spatial robustness to visuotactile policies.

4 Conclusion and Limitations

This work introduces VITAL, a framework that integrates local observation spaces and semantic augmentations for visuotactile policy learning with VLM-guided coarse navigation to achieve millimeter-precision manipulation across diverse scenes and object configurations. We recognize a few limitations of this work: (1) Our spatial variation experiments focus on horizontal surfaces – extending the method to arbitrary 3D configurations (e.g., vertical or angled placements) would be an interesting future direction. (2) Our VLM navigation assumes obstacle-free paths. Enhancing spatial reasoning to handle cluttered environments through advanced VLMs with obstacle-aware path planning could broaden the applicability of the method. (3) Our current evaluations use controlled lab conditions. Testing in real-world home environments with dynamic lighting, occlusions, and unstructured layouts would better validate robustness.

References

[1] 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.

[2] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, J. Tanner, Q. Vuong, A. Walling, H. Wang, and U. Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2024. URL https://arxiv.org/abs/2410.24164.

[3] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[4] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494, 2022.

[5] 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. URL https://arxiv.org/abs/2408.00714.

[6] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[7] M. Reid, N. Savinov, D. Teplyashin, D. Lepikhin, T. Lillicrap, J.-b. Alayrac, R. Soricut, A. Lazaridou, O. Firat, J. Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[8] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[9] S. Levine, C. Finn, T. Darrell, and P. Abbeel. End-to-end training of deep visuomotor policies, 2016. URL https://arxiv.org/abs/1504.00702.

[10] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T.-W. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *arXiv preprint arXiv:2307.15818*, 2023.

[11] K. Cobbe, O. Klimov, C. Hesse, T. Kim, and J. Schulman. Quantifying generalization in reinforcement learning, 2019. URL https://arxiv.org/abs/1812.02341.

[12] C. Zhang, O. Vinyals, R. Munos, and S. Bengio. A study on overfitting in deep reinforcement learning, 2018. URL https://arxiv.org/abs/1804.06893.

[13] D. Yarats, R. Fergus, A. Lazaric, and L. Pinto. Mastering visual continuous control: Improved data-augmented reinforcement learning. *arXiv preprint arXiv:2107.09645*, 2021.

[14] A. Iyer, Z. Peng, Y. Dai, I. Guzey, S. Haldar, S. Chintala, and L. Pinto. Open teach: A versatile teleoperation system for robotic manipulation. *arXiv preprint arXiv:2403.07870*, 2024.

[15] E. Teoh, S. Patidar, X. Ma, and S. James. Green screen augmentation enables scene generalisation in robotic manipulation. *arXiv preprint arXiv:2407.07868*, 2024.

[16] C. Yuan, S. Joshi, S. Zhu, H. Su, H. Zhao, and Y. Gao. Roboengine: Plug-and-play robot data augmentation with semantic robot segmentation and background generation. *arXiv preprint arXiv:2503.18738*, 2025.

[17] L. Tang, M. Jia, Q. Wang, C. P. Phoo, and B. Hariharan. Emergent correspondence from image diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[18] H. K. Cheng and A. G. Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022.

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[20] R. Bhirangi, V. Pattabiraman, E. Erciyes, Y. Cao, T. Hellebrekers, and L. Pinto. Anyskin: Plug-and-play skin sensing for robotic touch. *arXiv preprint arXiv:2409.08276*, 2024.

[21] V. Pattabiraman, Y. Cao, S. Haldar, L. Pinto, and R. Bhirangi. Learning precise, contact-rich manipulation through uncalibrated tactile skins. *arXiv preprint arXiv:2410.17246*, 2024.

[22] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi:10.15607/RSS.2023.XIX.016.

[23] T. Silver, K. Allen, J. Tenenbaum, and L. Kaelbling. Residual policy learning. *arXiv preprint arXiv:1812.06298*, 2018.

[24] S. Haldar, J. Pari, A. Rai, and L. Pinto. Teach a Robot to FISH: Versatile Imitation from One Minute of Demonstrations. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi:10.15607/RSS.2023.XIX.009.

[25] I. Guzey, Y. Dai, B. Evans, S. Chintala, and L. Pinto. See to touch: Learning tactile dexterity through visual incentives. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13825–13832, 2024. doi:10.1109/ICRA57147.2024.10611407.

[26] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[27] L. Smith, I. Kostrikov, and S. Levine. A walk in the park: Learning to walk in 20 minutes with model-free reinforcement learning. *arXiv preprint arXiv:2208.07860*, 2022.

[28] M. Deitke, C. Clark, S. Lee, R. Tripathi, Y. Yang, J. S. Park, M. Salehi, N. Muennighoff, K. Lo, L. Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024.

[29] S. Haldar, Z. Peng, and L. Pinto. Baku: An efficient transformer for multi-task policy learning. *arXiv preprint arXiv:2406.07539*, 2024.

[30] P. J. Ball, L. Smith, I. Kostrikov, and S. Levine. Efficient online reinforcement learning with offline data. In *International Conference on Machine Learning*, pages 1577–1594. PMLR, 2023.

[31] A. Karpathy. mingpt: A minimal pytorch re-implementation of the openai gpt. `https://github.com/karpathy/minGPT`, 2021.

Appendix

A1 Hyperparameters and Network Architecture

A1.1 Hyperparameters

The complete list of hyperparameters is provided in Table 6. We collect 32 demonstrations for each task using a VR-based teleoperation framework [14] operating at 30Hz. The collected data is subsampled to 6Hz for policy training, and the learned policies are deployed at 6Hz during real-world execution. All training is done on a local desktop with an NVIDIA RTX 3080 GPU with 8GB VRAM. For BAKU [29], ViSk [21], and VITAL-BC, training for 20k iterations with a training time of around 30 minutes provided the best results. For online RL (RLPD [30], VITAL), each method is trained online for 45 minutes at a 6Hz frequency(around 16k environment steps). For both offline and online training, the image observations are augmented with random cropping and color jitter. For tactile observations from the AnySkin [20] sensor, we subtract a baseline measurement from each tactile reading to account for sensor drift [21].

A1.2 Network Architecture

We use a randomly initialized ResNet-18 [19] as our image encoder and a 2-layer MLP for encoding the 15-dimensional tactile reading from AnySkin. For offline training, we use a transformer-based policy [29, 21], using the minGPT [31] architecture for the transformer. For offline training, all baselines and VITAL use an MLP action head, predicting a chunk of 10 future actions. Instead of using only the current action prediction, we use a temporal ensemble to combine all the past chunked action predictions. This temporal ensemble performs a weighted average over these predictions with an exponential weighing scheme $w_i = exp(-m * i)$, where w_0 is the weight for the oldest action. The speed for incorporating a new observation is governed by m, where a smaller m means faster incorporation. It must be noted that this ensembling incurs no additional training cost, only extra inference-time computation. In our experiments, similar to prior work [29, 21], we find both action chunking and temporal ensembling to be important for producing precise and smooth motion. During online residual RL, an offset is learned on top of the temporally smoothed offline action. We set m to 0.01 for all our experiments.

During online residual RL, the offset scale is chosen to balance exploration capacity about the base action and the convergence speed of online training. Since we focus on precise tasks demanding sub-millimeter level accuracy, we set to maximum offset magnitude to be 20% of the maximum action observed in the training data. Further, we observed that exploration plays an important role in preventing early collapse during online training. Thus, we employ a linearly decaying standard-deviation schedule – starting with high noise during the initial RL phase to ensure flexibility, then gradually reducing it to guarantee stable convergence. For residual RL, the actor is a 1-layer MLP while the critic comprises 2-MLP layers. For RLPD, which trains the RL policy from scratch, we use a 4-layer MLP for the actor network.

VITAL-BC has a total of 7.6M parameters, while VITAL has an additional 2.46M from the residual RL phase, resulting in a total of 9.06M parameters.

A2 Task Descriptions

Plug in Socket The robot arm holds a plug within the gripper and is tasked with inserting the plug into a socket. The robot's initial position is randomly sampled in a 6cm×6cm area around the socket, 10cm above the socket.

USB Insertion The robot arm holds a USB stick within the gripper and is tasked with going down and inserting the USB stick into a USB socket. The robot's initial position is randomly sampled in a 6cm×6cm area around the socket, 10cm above the socket.

Table 6: List of hyperparameters.

Parameter	Value
Learning rate	$1e^{-4}$
Image size	128×128
Batch size	256
Optimizer	Adam
Hidden dim	256
Observation history length	1
Action head	MLP
Action chunk length	10
Residual RL Offset scale	20% of max absolute action
Exploration schedule for RL	linear(0.25,0.1,5000)
Update-to-data ratio (UTD)	16

Card Swiping The robot arm holds a credit card within the gripper and is tasked with swiping the card through a card machine. The robot's initial position is sampled in a 4cm×4cm×2cm area in front of the card machine.

Key in Lock The robot arm holds a key within the gripper and is tasked with inserting the key into a lock. The robot's initial position is sampled in a 6cm×6cm area around the key hole, 10 cm above the socket.

Pick block The robot arm is tasked with picking up a block placed at a fixed position on the table. The robot's initial position is sampled in a 6cm×6cm area around the block, 10 cm above the block.

A3 Baseline Implementations

BAKU [29] This is a visual behavior cloning baseline using a transformer architecture and a deterministic MLP action head. We follow the hyperparameters and network architectures described in Appendix A1 for BAKU.

ViSk [21] This is a visuo-tactile behavior cloning baseline using a transformer architecture and a deterministic MLP action head. We follow the hyperparameters and network architectures described in Appendix A1 for ViSk.

RLPD [30] This involves collecting a few expert demonstrations and training an RL policy from scratch, where the data during RL training is sampled 1:1 between the expert and RL replay buffer. RLPD employs a high update-to-data ratio (UTD) and layer normalization in the critic to enable sample-efficient online learning. We observe that during the initial phase of training, since the actor is randomly initialized, the RLPD policy outputs unsafe actions, making the rollout jerky. This highlights the importance of pretraining for stable online learning.

A4 Spatial and Scene Generalization

Figure 3 demonstrates VITAL performing four precise manipulation tasks in the same position as during training, but with a novel background. Figure 4 further highlights VITAL's spatial and scene generalization capabilities: spatial generalization is achieved through VLM-guided reaching in conjunction with a localized observation and action space, while semantic augmentations support scene

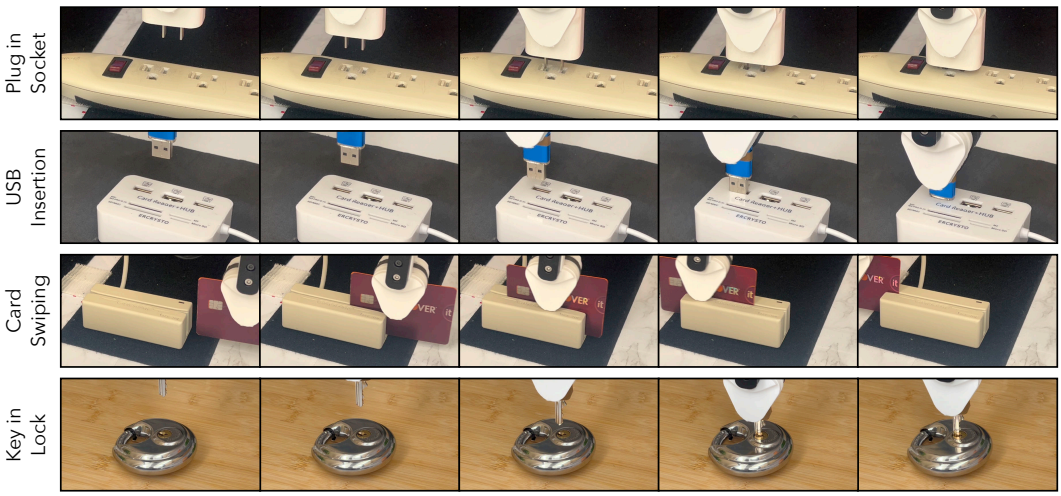

Figure 3: Real-world rollouts showing VITAL's ability on four precise manipulation tasks.

generalization during training, which promote invariance to changes in background, textures, lighting, and clutter. Together, this allows VITAL to decouple spatial reasoning from scene understanding, enabling reliable, precise manipulation across a wide range of previously unseen environments without retraining.

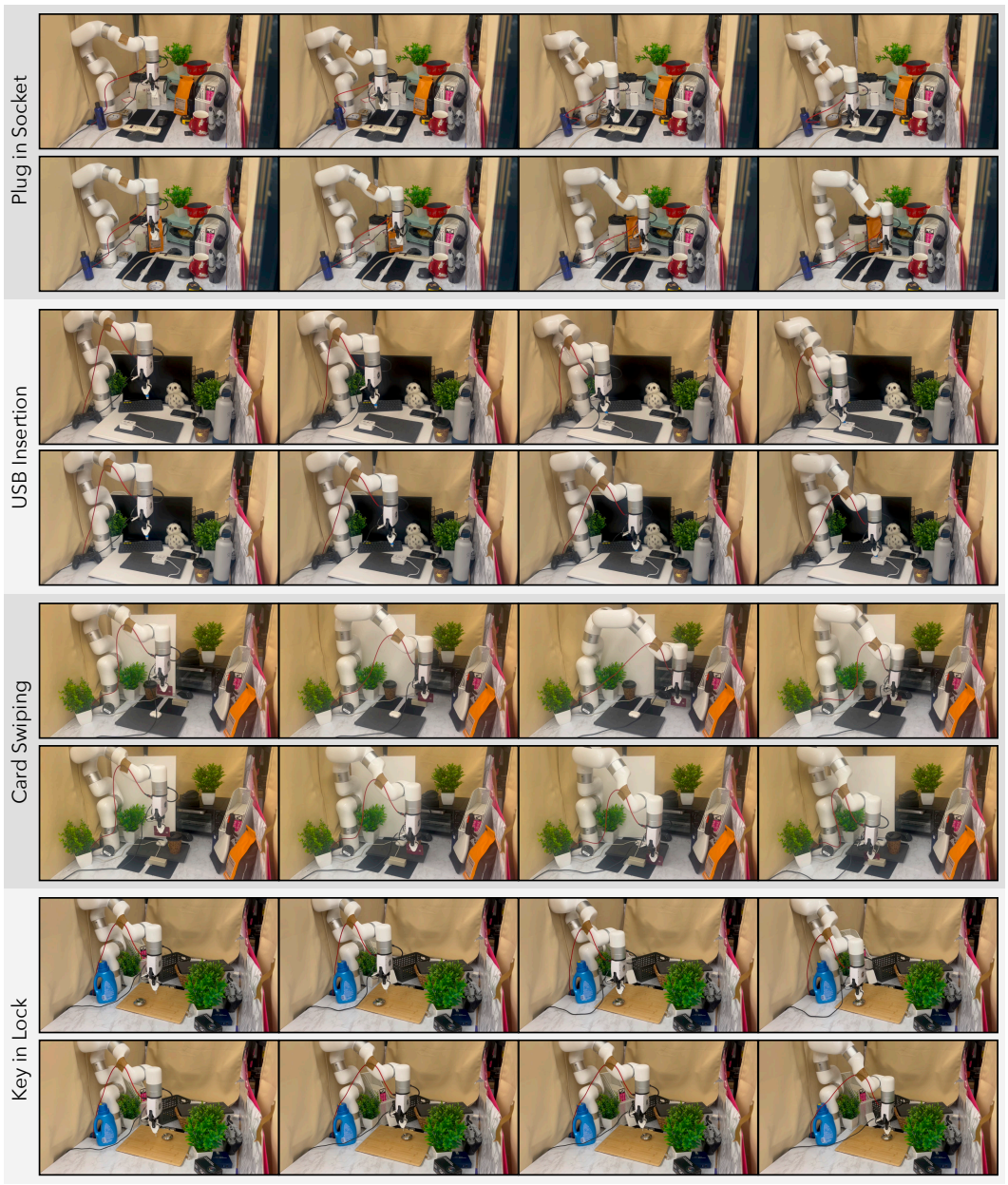

Figure 4: Real-world rollouts showing that VITAL generalizes to spatial variations and background distractor objects.