Adapting by Analogy: OOD Generalization of Visuomotor Policies via Functional Correspondence

Author Names Omitted for Anonymous Review. Paper-ID [add your ID here]

***Abstract*—End-to-end visuomotor policies trained using behavior cloning have shown a remarkable ability to generate complex, multi-modal low-level robot behaviors. However, at deployment time, these policies still struggle to act reliably when faced with out-of-distribution (OOD) visuals induced by objects, backgrounds, or environment changes. Prior works in interactive imitation learning solicit corrective expert demonstrations under the OOD conditions—but this can be costly and inefficient. We observe that task success under OOD conditions does not always warrant novel robot behaviors. In-distribution (ID) behaviors can directly be transferred to OOD conditions that share functional similarities with ID conditions. For example, behaviors trained to interact with in-distribution (ID) pens can apply to interacting with a visually-OOD pencil. The key challenge lies in disambiguating which ID observations functionally correspond to the OOD observation for the task at hand. We propose that an expert can provide this OOD-to-ID functional correspondence. Thus, instead of collecting new demonstrations and re-training at every OOD encounter, our method: (1) detects the need for feedback by checking if current observations are OOD and the most similar training observations show divergent behaviors, (2) solicits functional correspondence feedback to disambiguate between those behaviors, and (3) intervenes on the OOD observations with the functionally corresponding ID observations to perform deployment-time generalization. We validate our method across diverse real-world robotic manipulation tasks with a Franka Panda robotic manipulator. Our results show that test-time functional correspondences can improve the generalization of a vision-based diffusion policy to OOD objects and environment conditions with low feedback. More details on our project page: https://anon-corl2025.github.io/project-page/**

I. Introduction

A central goal in robot learning is to enable robots to generalize: to successfully perform tasks in environments they have never seen before. Imagine a robot encountering a pencil for the first time. With just an RGB image, it should be able to reason about the scene and delicately place the object into a nearby cup, as shown in Figure 1. One popular approach towards this is imitation-based visuomotor policy learning. However, while internet-scale datasets have powered generalization breakthroughs in vision and language, robotics still lacks access to the same data scale [8, 1, 4, 5, 19], and collecting expert demonstration data remains expensive and time-consuming. This results in robots failing in unintuitive ways when faced with out-of-distribution (OOD) environments (lower left, Figure 1). Nevertheless, even with modest expert demonstration datasets, recent advances in policy architectures and training algorithms have enabled robots to learn complex visuomotor skills—such as grasping thin tools or folding clothes and operating articulated objects—that work well in-distribution (ID) [2, 22, 9, 11, 17]. This raises the central question of our work: *How can we reuse robot behaviors learned in in-distribution settings to succeed in out-of-distribution scenarios?*

Our key insight is that behavior generalization may not always require more demonstration data: it may just need a better correspondence between the training and test conditions. For example, in Figure 1, even though the robot has never seen pencils before, it has seen similarly-thin pens and thicker markers (top row). Thus, if it understood that the pencil is functionally equivalent to the pen in this task, it could "imagine" that the pencil is a pen and reuse the pen pickup behavior to successfully complete the task. Based on this insight, we present ***Adapting by Analogy (ABA)***: a method which establishes *functional correspondences* between in-distribution and out-of-distribution scenes to steer a visuomotor policy through OOD conditions. A key aspect of our method is to leverage expert human knowledge—in the form of a textual description—to interactively learn high-level functional correspondences relevant to the task at hand. The textual description is decoded into a functional correspondence feature space that matches corresponding semantic segments of the scene to retrieve ID behaviors that are "relevant" for the current OOD scene. To measure whether the functional correspondence is well-specified, the robot estimates its uncertainty over the retrieved behavior modes and continues to ask for correspondence refinement until it is certain in the mapping.

We instantiate ***Adapting by Analogy*** on hardware with a Franka Research 3 manipulator acting with a diffusion-based visuomotor policy [2]. By controlling the training and test environments, we study i) how functional correspondences can improve the task success rate in increasingly OOD environments, ii) if our method seeks expert feedback efficiently, and iii) we verify how critical functional correspondences are for OOD generalization. We find that even a relatively small number of expert-guided functional correspondences can significantly improve the generalization capabilities of a visuomotor policy interacting with OOD objects from new semantic categories.

II. Related Works

Test-time Policy Interventions. Runtime policy interventions are a policy failure mitigation, where-in the policy's execution is intervened and new knowledge is supplied in-order to help mitigate the failure. For instance, a line of work directly proposes interventions on the policy's behavior space,

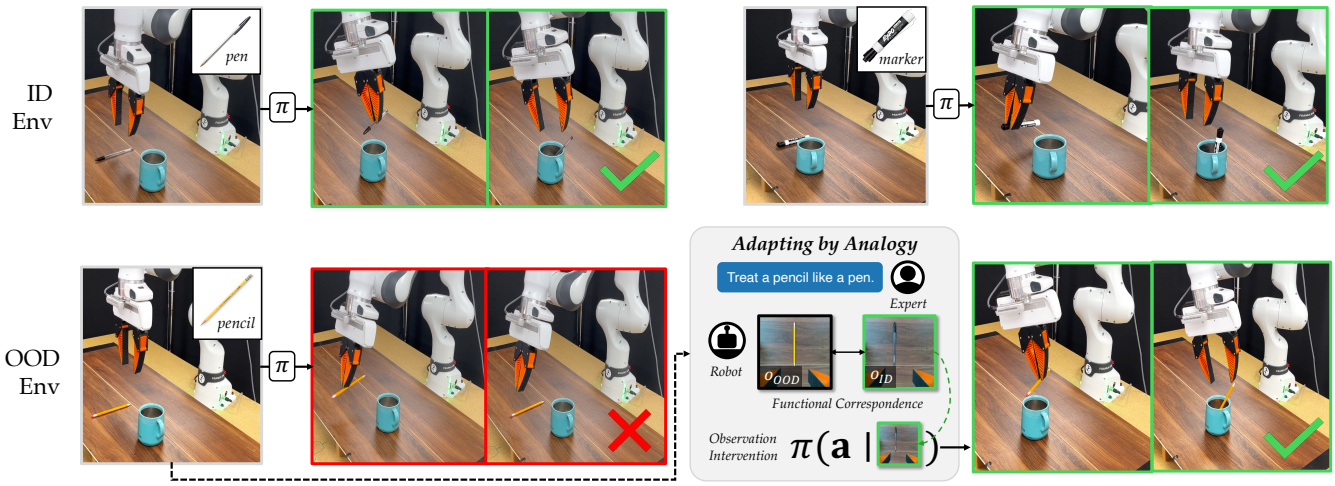

Fig. 1: We present ***Adapting by Analogy***, a test-time method that uses functional correspondences between deployment and training conditions to improve a policy's performance in OOD conditions.

steering the policy into desired modes either through human feedback [20], through Q functions optimized on large scale offline datasets [13], or through predictive modeling [21, 15]. Another line of work proposes intervention directly on the policy's observations, with synthesized observations to remedy known causes of failure [6]. Our work also proposes policy observations with functionally similar ID observations to generalize to novel out-of-distribution conditions.

Functional Correspondence for Behavior Transfer. The ability to transfer behaviors from one set of objects to an unseen set of objects hold the potential to unlock robot generalization in the wild. This problem has been studied through functional correspondences [10]. Prior work have leveraged functional correspondences to directly transfer behaviors across objects from a single demonstration to novel objects in a one shot manner, or in a zero shot manner by leveraging an affordance dataset [18, 12, 7]. Here, functional correspondences are typically established through keypoint based reasoning. In our work, we establish functional correspondences through a correspondence description provided by an expert. Furthermore, instead of directly adapting retrieved behavior, we intervene using the functionally similar training observation

III. Problem Formulation

Environment, Observation, & Action Models. We model the robot's environment $E \in \mathbb{E}$ as broadly consisting of factors external to the robot such as the objects in the scene, the background, camera configurations, etc. In a particular environment E, the robot senses its proprioceptive states $q \in \mathcal{Q}$ (e.g., end-effector pose, gripper state) and uses a sensor $\sigma : \mathcal{Q} \times E \to \mathcal{I}$ to obtain high-dimensional RGB image observations of the scene. At any time t, let the stacked image-proprioception observations be, $o_t \in \mathcal{O} := \mathcal{I} \times \mathcal{Q}$. Finally, let $a \in \mathcal{A}$ be the robot's action (e.g., end-effector positions and rotations and gripper action).

Training Data. For training the visuomotor policy, we assume access to a dataset of observation-action tuples, $\mathcal{D}_{\text{ID}} := \{(o_t^i, a_t^i)\}_{i=1}^N$, drawn from a set of M environments we treat as "in distribution": $E_{\text{ID}} := \{E_{\text{ID}}^1, E_{\text{ID}}^2, \ldots, E_{\text{ID}}^M\}$. For example, a training distribution of environments could consist of M unique objects and their configurations in the environment.

Visuomotor Policy. Let the robot's policy be a multi-modal imitative action generation model [2, 11] denoted by $\pi(\mathbf{a}_t \mid \mathbf{o}_t)$. Here $\mathbf{a}_t := a_{t:t+T}$ is a T-step action plan and $\mathbf{o}_t := o_{t:t-H}$ is an H-step history of observations. We assume that the policy network first encodes any observation into a corresponding latent state, $z_t = \mathcal{E}(o_t)$, via an encoder. Let $\mathbf{z}_t = \mathcal{E}(o_{t:t-H})$ be a sequence of latent state embeddings. The policy is pre-trained via an imitation learning loss on the in-distribution dataset of observation-action pairs from $\mathcal{D}_{\text{ID}}$.

OOD-to-ID Generalization via Functional Correspondances. Given a visuomotor policy $\pi(\mathbf{a}_t \mid \mathbf{o}_t)$ pre-trained on behaviors from in-distribution environments E_{ID}, we seek to generalize the robot's task performance to *out-of-distribution* (OOD) environments, E_{OOD}. Since the general problem of OOD generalization is an extremely challenging open problem, in this paper we assume that (1) E_{ID} and E_{OOD} differ only by the objects present in the scene and background color (but the environment geometry remains the same), (2) the training observations $\mathcal{O}_{\text{ID}}$ and deployment time observations $\mathcal{O}_{\text{OOD}}$ are obtained on the same robot embodiment, and (3) we have access to the training data, $\mathcal{D}_{\text{ID}}$.

Our key idea is to identify *functional correspondences* between the test-time OOD scene—in which the base policy would fail to act correctly—and training-time ID scenes, in which the policy can generate high-quality behaviors. Functional correspondences identify parts of the image observations with similar affordances for the task at hand. Intuitively, learned robot behaviors should be transferable across observations whose affordance maps are aligned, i.e., observations where regions that have similar affordances overlap. Thus, we

aim to retrieve ID observations whose functional correspondences are aligned with the test-time OOD observation.

Problem Formulation: Expert-Guided Functional Correspondences. The core challenge lies in identifying the functional correspondences across the OOD image observations and the ID image observations. Humans possess the ability to infer object affordances, and generalize them to novel objects. Thus, we propose to leverage experts feedback in the form of natural language to acquire these functional correspondences between the OOD and the ID image observations.

Formally, let the **functional correspondance map** be denoted by $\Phi : \mathcal{I} \times \mathcal{I} \times \mathcal{L} \to \mathcal{P}(\Omega \times \Omega)$. Given two images $i, \hat{i} \in \mathcal{I}$ and a natural language description $l \in \mathcal{L}$ provided by the expert, this mapping returns all pairs of functionally corresponding image segments $(\omega, \hat{\omega})$ where $\omega \in \Omega$ are image segments from image i and $\hat{\omega} \in \hat{\Omega}$ are image segments from image $\hat{i}$. Here, $\mathcal{P}(\Omega \times \Omega)$ is the powerset of all paired image segments. Let K be the number of corresponding image segments. Thus, the functional correspondence map is defined as:

$$\Phi(i, \hat{i}, l) := \{(\omega_j, \hat{\omega}_j) \mid j \in \{0, 1, \ldots, K\}\} \tag{1}$$

We measure the **functional correspondence alignment** of any two images via $f : \mathcal{P}(\Omega \times \Omega) \to \mathbb{R}$. In this work, we model f as the total Intersection over Union (IoU) between functionally corresponding regions of the images returned by the functional correspondance map, Φ:

$$f(\Phi(i, \hat{i}, l)) = \sum_{i=0}^{K} \text{IoU}(\omega_j, \hat{\omega}_j) \tag{2}$$

Finally, given any OOD observation $\hat{o} = (\hat{q}, \hat{i})$ observed by the robot at deployment time and an expert language input l describing the functional correspondences, we retrieve the ordered set of in-distribution observations $\mathcal{O}_f = (o_1, o_2, \ldots, o_k) \subseteq \mathcal{O}_{\text{ID}}$ ranked by their functional alignment from Eq. (2). For intervention, we use the behaviors extracted from the top-M observations in $\mathcal{O}_f$.

IV. Method: Adapting by Analogy

A. Detecting Out-of-Distribution Observations

At each timestep, our method first detects if the robot's observations $\hat{o}$ are anomalous via a fast OOD detector. We measure the cosine similarity between the encoded observation $\hat{z} = \mathcal{E}(\hat{o})$ and embeddings of in-distribution observations $z \in \mathcal{E}(o_i),\ o \in \mathcal{O}_{\text{ID}}$ via $\text{IDScore}(\hat{o}, \mathcal{O}_{\text{ID}}) := \min_{o \in \mathcal{O}_{\text{ID}}} \frac{\mathcal{E}(\hat{o}) \cdot \mathcal{E}(o)}{\|\mathcal{E}(\hat{o})\| \|\mathcal{E}(o)\|}$. If the IDScore is above a threshold λ, then we deem the observation to be nominal and directly execute the action $\hat{a} \sim \pi(\cdot \mid \hat{o})$. Otherwise, we deem the observation to be OOD, ask the expert for an initial language instruction l describing relevant functional correspondences, and use those to intervene on the observation before action generation.

B. Establishing OOD-to-ID Functional Correspondences

Given an OOD observation $\hat{o} = (\hat{q}, \hat{i})$ identified via our fast anomaly detector and the expert's language description l, we want to intervene on the policy by reusing learned behaviors from functionally similar ID observations. This requires computing the functional alignment from Eq. 2 between $\hat{o}$ and every ID observation $o \in \mathcal{O}_{\text{ID}}$. However, implementing this matching is challenging in practice for two reasons: first, it is computationally expensive (requiring image segmentation and IoU computation over all corresponding segments), and second, matching correspondances between 2D image segments does not directly reveal correspondences in the high-dimensional robot state. Thus, we filter the demonstration dataset $\mathcal{D}_{\text{ID}}$ consisting of N observation-action trajectories τ to retrieve one observation $o \in \tau$ per trajectory which contain similar proprioceptive states q to the test-time robot state $\hat{q}$. Mathematically, for a distance threshold $\lambda_q \in \mathbb{R}^+$ and the current configuration $\hat{q}$, let the filtered observation dataset $\mathcal{O}_q \subset \mathcal{O}_{\text{ID}}$ be:

$$\mathcal{O}_q = \left\{ o \;\middle|\; o = \arg \min_{(o,a) \in \tau} \left(\|q - \hat{q}\|_2 - \lambda_q \right), \quad \forall \tau \in \mathcal{D}_{\text{ID}} \right\} \tag{3}$$

Using this filtered dataset, we can now compute our **functional correspondence map** from Eq. 1 via two internal models: one which converts the expert's language feedback $l \in \mathcal{L}$ into a functional feature set (denoted by ϕ_l) and another which semantically segments each ID image $i \in \mathcal{O}_q$ and the current OOD image observations $\hat{i}$ to generate the set of image masks and semantic labels denoted by $\hat{\Omega}$ and Ω respectively. We use Grounded Segment Anything [16] for semantic segmentation.

Next, the expert's language input l is decoded into a set of correspondence features ϕ_l that can be applied to the semantic segmentations $\Omega, \hat{\Omega}$ to return a set of K functionally corresponding image segments $(\omega^j, \hat{\omega}^j), j \in \{0, \ldots, K\}$. For example, the ϕ_l that is decoded from l =*"Match pencils with pens"* lifts pixels corresponding to the segmentation label *'pencil'* in the OOD image, and pairs it with pixels corresponding to the label *'pen'* in the ID images. In this work, we use a templated l, but future work could explore the use of LLMs as an interface.

Ultimately, after Φ extracts the set of functionally corresponding image segments, we measure their alignment using Eq. 2. Each ID observation $o \in \mathcal{O}_q$ is ranked based on its functional alignment with the OOD observation $\hat{o}$ to obtain the ordered set of functionally corresponding ID observations $\mathcal{O}_f \subseteq \mathcal{O}_q$ used during intervention (Sec. IV-D).

C. Refining Functional Correspondences Until Confident

Thus far, we have assumed that the initial expert description l of functional correspondences was sufficient for the entire task. However, correspondences may evolve during task execution. For example, consider the task of picking up trash and sorting it into organic and recycling. The functional correspondence between types of trash (organic and recycling)

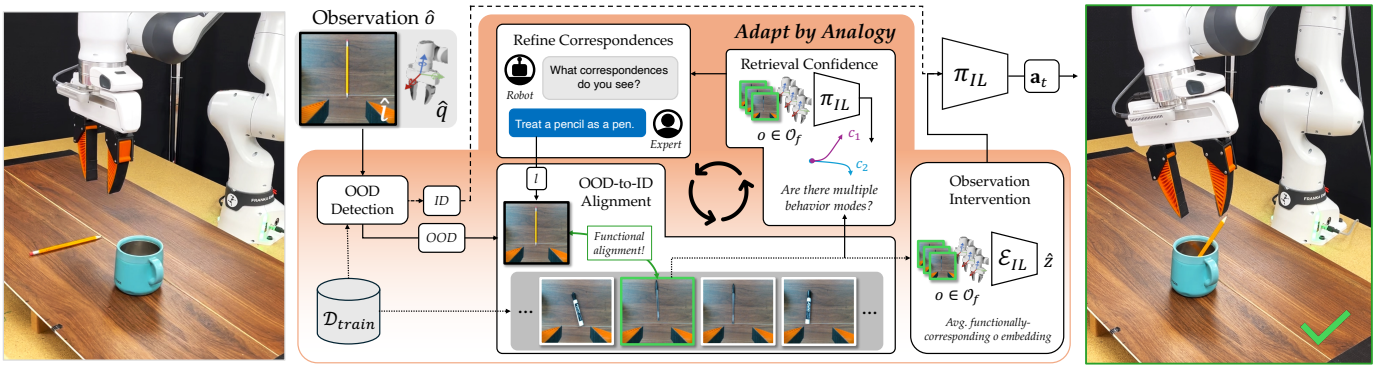

Fig. 2: **Adapting by Analogy** consists of four key phases. (left) First, we run a fast OOD detector by checking the cosine similarity between the current observation $\hat{o}$ and the training observations. (center, top-left) Given a correspondence description l, we establish OOD-to-ID functional correspondences to retrieve corresponding ID observations (center, bottom). We refine the correspondances with the expert as long as there is ambiguity in the predicted behavior mode (center, top-right). Once finalized, we intervene on the observations and execute the planned actions (right).

does not matter initially when the robot is planning a grasp, but becomes relevant once the item has been picked up and needs to be sorted. Thus, our method interactively refines the functional correspondence description until the robot is confident in the behavior it has retrieved.

Intuitively, a well-established functional correspondence will reduce the diversity in robot action plans, focusing on the "correct" behavior mode. To quantify the relevant behavior modes before functional alignment, we obtain a set of action plans $\mathcal{A}_q := \{\mathbf{a} \sim \pi(\cdot \mid o) \mid o \in \mathcal{Q}_q\}$ for all observations with the same proprioceptive state via a forward pass through the policy. Behavior mode labels are obtained by fitting n_c clusters to $\mathcal{A}_q$ via K-means clustering. Since the current functionally-aligned observations are a subset $\mathcal{Q}_f \subseteq \mathcal{Q}_q$, we can obtain labels for all functionally-aligned *action plans* $\mathcal{A}_f \subseteq \mathcal{A}_q$ and measure the reduction in behavior modes via the entropy over the action plan labels. As long as the entropy in the retrieved actions is high, the robot keeps asking the expert to refine their functional correspondence description l by showing them the current observations and their behaviors, then re-doing the OOD-to-ID matching from Sec. IV-B.

D. Intervening on Observations to Generate Functionally-Corresponding Behavior

Once the correspondence description l is complete and the retrieved action mode uncertainty is sufficiently low, the robot intervenes on its observations to generate functionally "correct" behavior. Specifically, observations in the final refined $\mathcal{O}_f$ are ranked based on their functional alignment as measured by Eq. 2. To smooth out action prediction, we generate the final executed action plan by interpolating the embeddings of M-highest ranked corresponding observations: $\hat{z} := \frac{1}{M} \sum_{o \in \mathcal{O}_f} \mathcal{E}(o)$ before passing the average embedding to the policy network.

V. Hardware Experiments

We conduct a series of experiments in robot hardware to study: (1) How much does *Adapting by Analogy* improve the visuomotor policy's closed-loop performance on OOD environments induced by novel objects and backgrounds conditions?, (2) What kind of features (e.g., base policy's embedding, DINOv2 [14], or functional correspondences) maximally help observation interventions?, (3) How efficient is our method at seeking expert feedback for adaptation in OOD environments?, (4) When intervention schemes succeed, are they retrieving functionally-aligned observations?

Real Robot Setup. We use a Franka Research 3 robotic manipulator equipped with a 3D printed UMI gripper [3] for our real-world experiments. The RGB image observations $i \in \mathcal{I}$ come from a wrist mounted RealSense D435 camera and a third-person Zed mini 2i camera overlooking the workspace. The overall robot observation $o := (i, q)$ consists of the concatenated images and the robot proprioception. More details about our setup can be found in the supplementary.

ID Environments & Tasks. We train two visuomotor policies on two different real-world manipulation tasks. The first task is **sweep-trash**, wherein robot must sweep trash towards different goals, based on whether the trash is organic and recycling. The next task is **object-in-cup**, where-in a robot arm is tasked with picking up a object such as a marker or a pen and dropping it in a mug. Pens—which are grasped above their center-of-mass—need to be dropped into the mug from the bottom, and markers—which are grasped below their center-of-mass—need to be dropped from the front. We divide the task in 3 sub-goals (A) grasping the object, (B) picking the correct behavior mode based on the grasp, and (C) dropping object into the cup.

Visuomotor Policy Training. We use a diffusion policy [2] as the base visuomotor policy $\pi(\mathbf{a} \mid \mathbf{o})$. It takes as input o and predicts a T-step action plan, where $T = 16$. For **sweep-trash**, the training dataset $|\mathcal{D}_{\text{ID}}| = 100$ consists of 50 demonstrations cleaning up crumpled paper (recycling trash) and 50 demonstrations cleaning up M&Ms (organic trash). For **object-in-cup**, the policy is trained on $|\mathcal{D}_{\text{ID}}| = 200$ demonstrations with 100 placing a marker (dropped from the back) and 100 placing a pen (dropped from the top).

OOD Environments. We test on two in-distribution environments and five OOD environments for each task. In addition to pens and markers, we e evaluate sweep trash with one background variation (workspace covered with black cloth), and three novel instances of trash (doritos, crumpled napkin, thumb tacks). For the object-in-cup, we test with in-distribution objects, one novel background (workspace covered with black cloth), and three instances of novel objects varying in shapes and sizes (pencil, battery, jenga block).

Baselines. We compare our method, **ABA**, with three baselines. **Vanilla** is the base visuomotor policy without any intervention mechanism. **PolicyEmbed** intervenes on the observations with a similar mechanism to ours, but it retrieves ID observations using cosine similarity in the base policy's learned embedding space, $\mathcal{E}(o) \in \mathcal{Z}$. It does not use any expert feedback. **DINOEmbed** also intervenes on the base policy, but it retrieves ID samples using cosine similarity in the DinoV2 [14] feature space of the OOD and ID observations. We use this to test if powerful pre-trained vision foundation models can implicitly capture functional correspondences beyond semantic object categories. We use both the class token features and the patch features. For **ABA**, we generate correspondence features ϕ via decoding l into a pre-templated set of features. The choices of the features are (1) match *'ood object semantic label'* with *'id object semantic label'*, (2) overlap segments of *'ood object'* with *'id object'* (3) Align left/right edge of segments, (4) align top/base of segments, and (5) "Pass", which meant that the expert does not want to refine the set of correspondence features. All intervention methods perform matching in the refined set of ID observations based on the robot's current proprioception $\mathcal{O}_q$, as described in Sec. IV-B.

Evaluation Procedure. All methods are evaluated via the same procedure and in the same conditions. For each ID and OOD environmental condition, we perform 10 rollouts of each method, placing the object of interest uniformly at random within a 15 cm horizontal range on the table. With a total of 14 environmental conditions, we collect a total of 140 rollouts for evaluation.

A. How much does ABA improve the policy's closed-loop performance?

In this section, we compare the overall task success rate of **ABA** with **Vanilla**. As shown in Fig. 3, **ABA** improves the **Vanilla** policy even in in-distribution environments by 15% on the sweep-trash task and 25% on the object-in-cup task. While the vanilla policy was robust to the novel background for the sweep-trash task, it completely degraded when faced with the novel background in the more challenging object-in-cup task. By reasoning about functional correspondences, **ABA** improved over the vanilla policy by 20% on the sweep trash task and by 90% on the object in cup task, staying robust to task-irrelevant changes to the background. Finally, we observe strong OOD generalization with **ABA** when evaluated under OOD objects where it improves over the vanilla policy by 76% on both tasks, showcasing that learned behaviors can be transferred to OOD objects from different semantic categories by reasoning about functional correspondences.

B. What kind of features maximally help observation interventions?

In this section, we compare how the features used for retrieving ID observations affect policy performance. For in-distribution environments, on the sweep trash task both **PolicyEmbed** and **DINOEmbed** perform on par with **ABA**. However, **ABA** outperforms by 15% on the object in cup task. Similar to **ABA**, both the **DINOEmbed** and **PolicyEmbed** are also robust to the novel background. Interestingly, **DINOEmbed**'s performance *improves* under the novel background. We hypothesize that this is due to the exceptional capabilities of the dino features at dense correspondence matching across objects within the same semantic category [14]. When tested on OOD objects, both **DINOEmbed** and **PolicyEmbed** struggle, achieving only 36.67% and 33.34% success rate respectively on sweep trash. On the object in cup task both baselines failed to successfully complete the task. Taking a closer look at the performance with specific OOD objects revealed common failures at the grasping stage and at picking the correct behavior mode. More analysis in supplementary.

C. How efficient is ABA at seeking expert feedback in OOD environments?

Next we study how often does **ABA** request feedback from the expert at test-time. Fig. 4 shows the number of times **ABA** requested feedback on average across 10 rollouts (with standard error bars) for both the sweep trash and the object in cup task, in all the three experiment settings (ID, OOD-Bg, OOD-Object).

For the **sweep-trash** task, **ABA** asks for feedback 3.8 ± 1.66 times for ID crumpled paper. In OOD backgrounds,feedback requests increased, e.g., to 4.5 ± 1.2 times per rollout for crumpled paper. **ABA** requested feedback the most for OOD objects, with the highest number of requests for doritos with an average of 5.2 ± 1.16 times per rollout. Note that each rollout for sweep trash ran for 80 timesteps, so this corresponds to asking for feedback 6% of the rollout. For the **object-in-cup** task, feedback was requested 2.1 ± 0.7 times for the ID pen, and 1.3 ± 1.0 times for the ID marker. Similar to sweep-trash, the feedback requests increased with OOD backgrounds: e.g., feedback about the pend was requested 4.3 ± 1.73 times per rollout. Finally, amongst the OOD objects, feedback was requested the most for battery at 3.4 ± 1.62 times per rollout. Note that each rollout in the object in cup task ran for 120 timesteps.

VI. Conclusion

In this work, we present *Adapting by Analogy*, a method for enabling deployment-time generalization of visuomotor policies by leveraging functional correspondences between out-of-distribution (OOD) and in-distribution (ID) observations. Rather than requiring new expert demonstrations for each novel scenario, our approach uses expert-provided functional

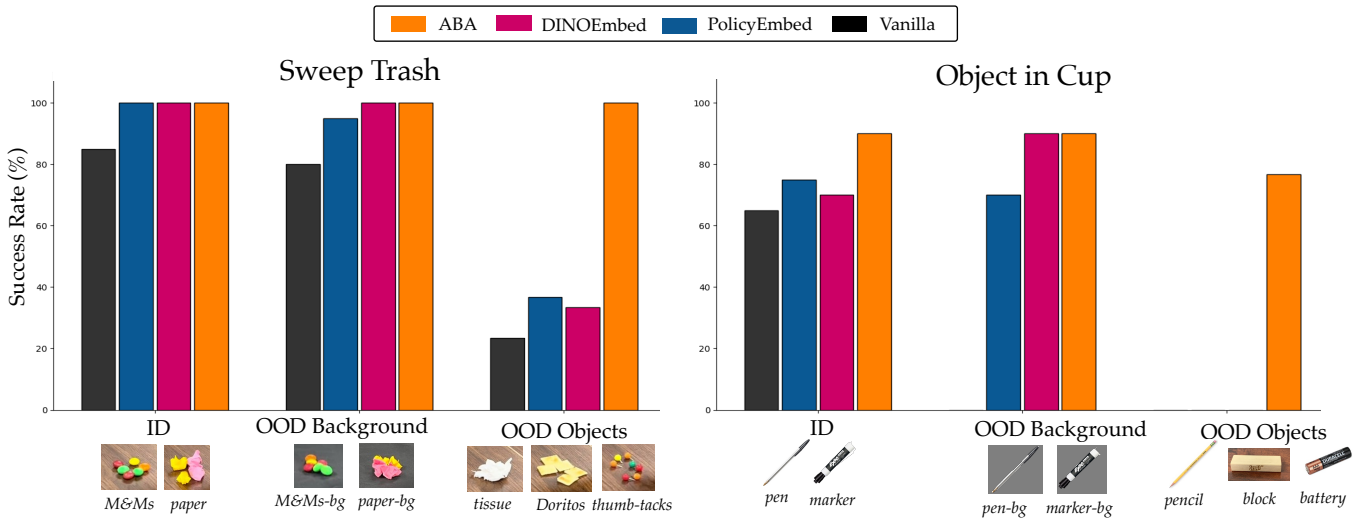

Fig. 3: **Task Success in ID and OOD Environments.** We report the task success rate averaged across 10 rollouts (per each ID and OOD conditions) and averaged across ID, OOD background, or OOD object conditions. For both the sweep-trash and the object-in-cup tasks, we see that **ABA** consistently achieves the highest task success rate compared to baselines.

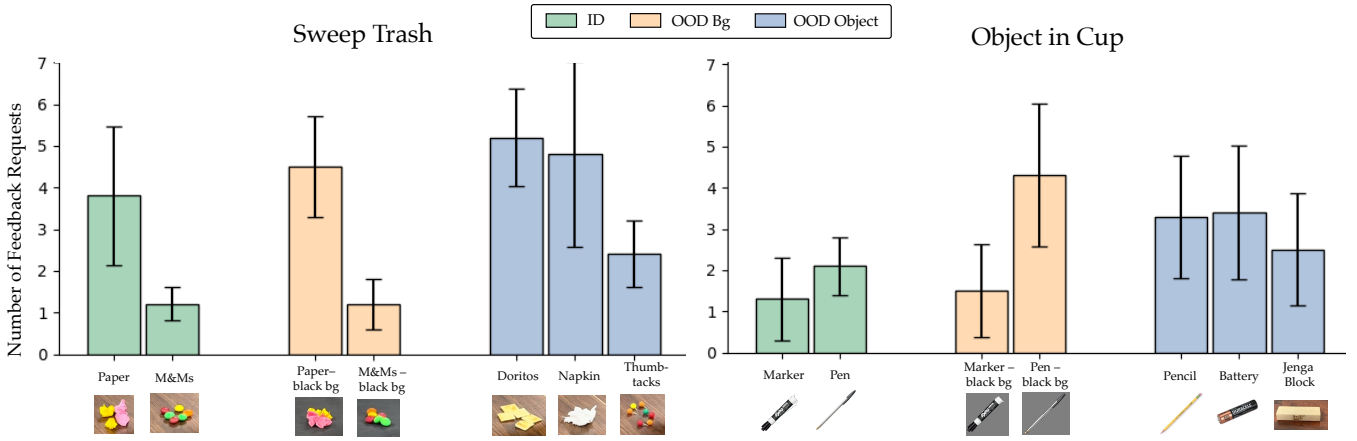

Fig. 4: **Expert Feedback Requested by ABA.** We show mean and standard error for the number of feedback requests across 10 rollouts per each environment. We find that **ABA** infrequently queries the expert for correspondances, given that sweep-trash has 70 timesteps and object-in-cup has 120.

features—which are interactively refined to represent during task execution—to repurpose existing ID policy behaviors in OOD environments. Empirical results across two real-world manipulation tasks with ten OOD environments demonstrate that establishing functional correspondences can improve a diffusion policy's success rate by 76% to new objects and backgrounds with minimal human intervention.

References

[1] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alex Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *arXiv preprint arXiv:2307.15818*, 2023.

[2] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 2024.

[3] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-

the-wild robot teaching without in-the-wild robots. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[4] AgiBot World Colosseum contributors. Agibot world colosseum. https://github.com/OpenDriveLab/AgiBot-World, 2024.

[5] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A robotic dataset for learning diverse skills in one-shot. In *RSS 2023 Workshop on Learning for Task and Motion Planning*, 2023.

[6] Asher J Hancock, Allen Z Ren, and Anirudha Majumdar. Run-time observation interventions make vision-language-action models more visually robust. *arXiv preprint arXiv:2410.01971*, 2024.

[7] Yuanchen Ju, Kaizhe Hu, Guowei Zhang, Gu Zhang, Mingrun Jiang, and Huazhe Xu. Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation. In *European Conference on Computer Vision*, pages 222–239. Springer, 2024.

[8] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

[9] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. *Conference on robot learning (CoRL)*, 2024.

[10] Zihang Lai, Senthil Purushwalkam, and Abhinav Gupta. The functional correspondence problem. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15772–15781, 2021.

[11] Seungjae Lee, Yibin Wang, Haritheja Etukuru, H Jin Kim, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. Behavior generation with latent actions. In *Forty-first International Conference on Machine Learning*, 2024.

[12] Yuyao Liu, Jiayuan Mao, Joshua Tenenbaum, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. One-shot manipulation strategy learning by making contact analogies. *arXiv preprint arXiv:2411.09627*, 2024.

[13] Mitsuhiko Nakamoto, Oier Mees, Aviral Kumar, and Sergey Levine. Steering your generalists: Improving robotic foundation models via value guidance. *Conference on Robot Learning (CoRL)*, 2024.

[14] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023.

[15] Han Qi, Haocheng Yin, Yilun Du, and Heng Yang. Strengthening generative robot policies through predictive world modeling. *arXiv preprint arXiv:2502.00622*, 2025.

[16] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[17] Moritz Reuss, Ömer Erdinç Yağmurlu, Fabian Wenzel, and Rudolf Lioutikov. Multimodal diffusion transformer: Learning versatile behavior from multimodal goals. In *Robotics: Science and Systems*, 2024.

[18] Chao Tang, Anxing Xiao, Yuhong Deng, Tianrun Hu, Wenlong Dong, Hanbo Zhang, David Hsu, and Hong Zhang. Functo: Function-centric one-shot imitation learning for tool manipulation. *arXiv preprint arXiv:2502.11744*, 2025.

[19] Open X-Embodiment Team. Open X-Embodiment: Robotic learning datasets and RT-X models. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[20] Yanwei Wang, Lirui Wang, Yilun Du, Balakumar Sundaralingam, Xuning Yang, Yu-Wei Chao, Claudia Perez-D'Arpino, Dieter Fox, and Julie Shah. Inference-time policy steering through human interactions. *arXiv preprint arXiv:2411.16627*, 2024.

[21] Yilin Wu, Ran Tian, Gokul Swamy, and Andrea Bajcsy. From foresight to forethought: Vlm-in-the-loop policy steering via latent alignment. *arXiv preprint arXiv:2502.01828*, 2025.

[22] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware, 2023. URL https://arxiv.org/abs/2304.13705.