Supplementary

This is the supplementary material to the paper, Adapting by Analogy: OOD Generalization of Visuomotor Policies via Functional Correspondence. More details can be found on our project page https://anon-corl2025.github.io/project-page/

1 Hardware Experiment Setup

Figure 1: Our hardware experiment setup, we use a Franka Research 3 robot, with a UMI gripper. The RealSense D435 wrist camera, and Zed mini 2i third person camera are placed as shown.

2 Tasks

We conduct our experiments on two real-world tasks. The first task is **sweep-trash**, wherein robot must sweep trash towards different goals, based on whether the trash is organic and recycling. For evaluation, we divide the task in two sub-goals (A) properly aligning the wiper with the trash, (B) sweeping to the correct location. 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. Markers—which are grasped above their center-of-mass—need to be dropped into the mug from the bottom, and pens—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. Fig. 2 demonstrates the various modes and the sub-goals for the two tasks.

We evaluate both tasks on the in-distribution conditions and OOD conditions induced by background and novel objects. The OOD environments are shown in Fig. 3 For **sweep-trash** our ID environments are $E_{\text{ID-trash}} := \{E_{\text{ID}}^{paper}, E_{\text{ID}}^{M\&Ms}\}$. The OOD environments are $E_{\text{OOD-Trash}} := \{E_{\text{OOD}}^{\text{doritos}}, E_{\text{OOD}}^{\text{napkin}}, E_{\text{OOD}}^{\text{thumb-tack}}, E_{\text{OOD}}^{\text{paper-bg}}, E_{\text{OOD}}^{\text{M\&M-bg}}\}$.

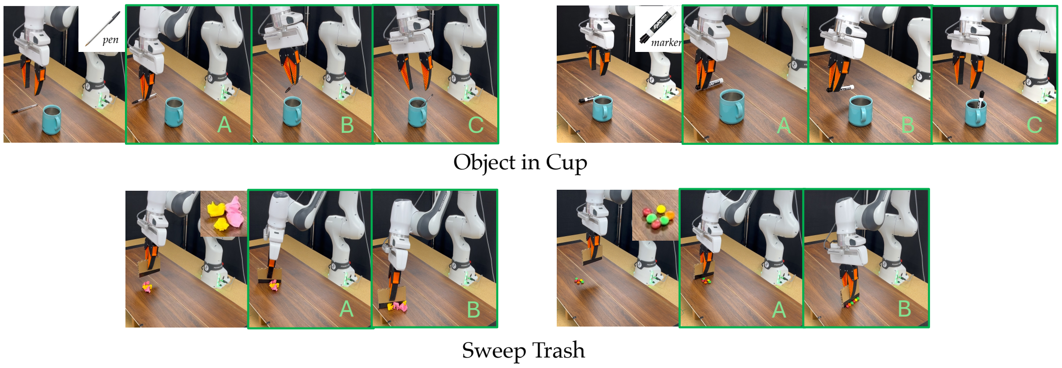

Figure 2: The training demonstrations for our two tasks, with their sub-goals(A, B, C). For the object in cup task, the pen is grasped below the center-of-mass, and is dropped into the mug from the front. The marker is grasped above the center-of-mass and is dropped into the mug from the bottom. For the sweep trash task, paper (i.e., recycling) is swept up, and M$Ms (i.e., organic) is swept down.

For **object-in-cup** our ID environments are $E_{\text{ID-object}} := \{E_{\text{ID}}^{marker}, E_{\text{ID}}^{pen}\}$. The OOD environments are $E_{\text{OOD-object}} := \{E_{\text{OOD}}^{\text{pencil}}, E_{\text{OOD}}^{\text{battery}}, E_{\text{OOD}}^{\text{block}}, E_{\text{OOD}}^{\text{marker-bg}}, E_{\text{OOD}}^{\text{pen-bg}}\}$.

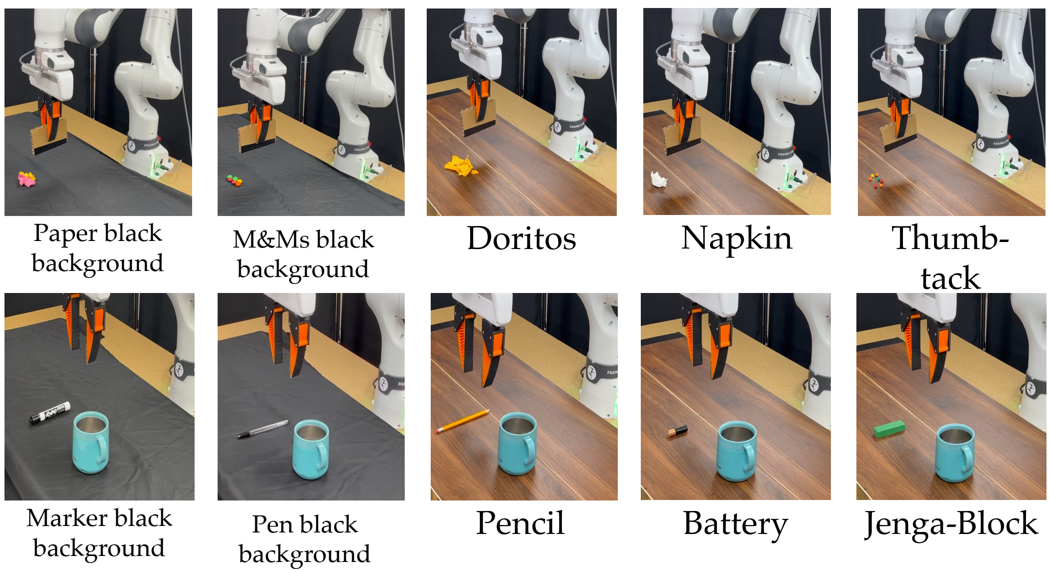

Figure 3: Our OOD environments for both the sweep-trash and object-in-cup task

3 Additional Results

3.1 How much does ABA improve the policy's sub-goal level closed-loop performance?

Fig. 4 shows that on the sweep-trash task, both **ABA** and **Vanilla** are able to successfully accomplish both subgoals on $E_{\text{ID}}^{M\&Ms}$. However, in E_{ID}^{paper}, **Vanilla** fails at aligning the wiper with the paper trash (subgoal A) 10% of the times, and sweeps paper incorrectly (subgoal B) 30% of the times. **ABA** maintains 100% performance on E_{ID}^{paper}.

Showing a similar trend, both **Vanilla** and **ABA** show 100% success rate on the E_{ID}^{marker} for the object-in-cup task. On the E_{ID}^{pen}, while both **Vanilla** and **ABA** are able to grasp the pen (subgoal A) 100% of the times, **ABA** improves over **Vanilla** by 40% at picking the right mode (subgoal B), showing a 100% success rate. Finally, since dropping the pen into the cup (subgoal C) is the most fine-grained aspect of the task, both **ABA** and **Vanilla** struggle but **ABA** still improves over **Vanilla** by 50%.

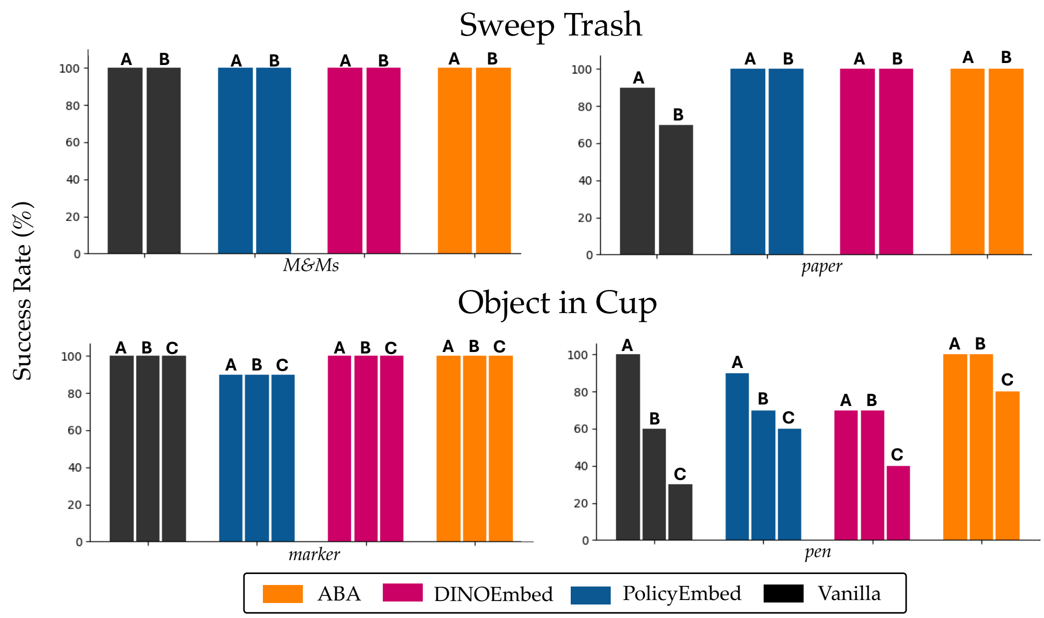

Figure 4: Subgoal Success in each ID Environment. We report the subgoal level task success rate averaged across 10 rollouts. For both the sweep-trash and the object-in-cup tasks, we see that **ABA** consistently achieves the highest task success rate compared to baselines.

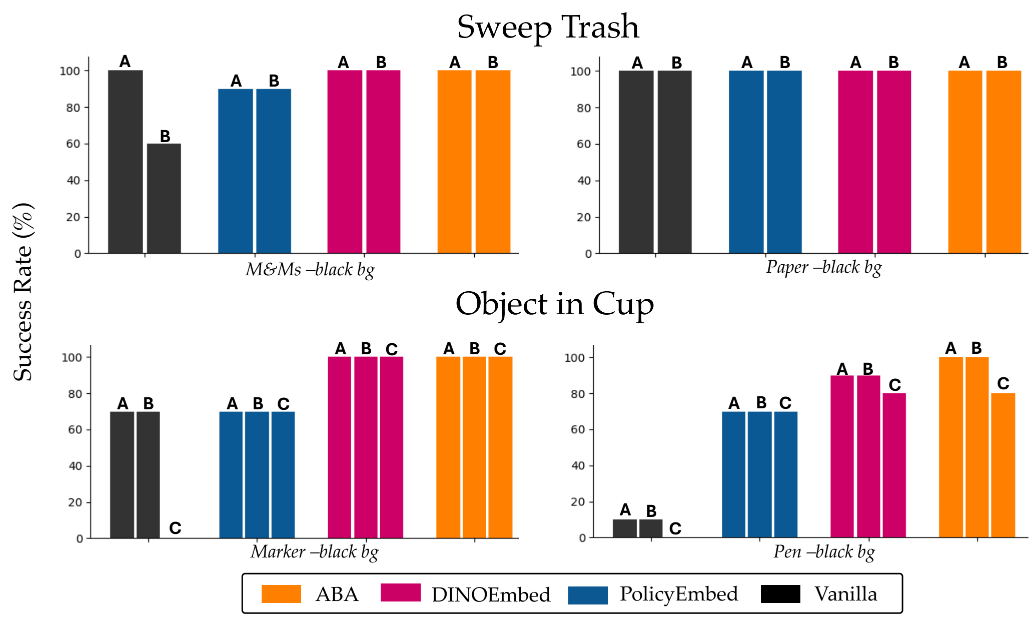

Figure 5: Subgoal Success in each OOD Environment, induced by changing the background. The success rate is averaged across 10 rollouts. **ABA** again consistently achieves the highest task success rate compared to baselines.

Next, we compare **ABA** and **Vanilla** across OOD environments, induced using a novel background ($E^{\text{paper-bg}}_{\text{OOD}}$, $E^{\text{M\&M-bg}}_{\text{OOD}}$, $E^{\text{pen-bg}}_{\text{OOD}}$, $E^{\text{marker-bg}}_{\text{OOD}}$).

Fig. 5 shows that on the sweep-trash task the performance trends are similar to the ID environments, although interestingly instead of $E^{paper-bg}_{\text{OOD}}$, **Vanilla** now shows poorer performance on subgoal B of $E^{\text{M\&Ms-bg}}_{\text{OOD}}$. **ABA** shows 100% success rate on both goals of both $E^{\text{M\&Ms-bg}}_{\text{OOD}}$ and $E^{\text{paper-bg}}_{\text{OOD}}$. On the object-in-cup task, **Vanilla** struggles on all subgoals of both $E^{\text{pen-bg}}_{\text{OOD}}$, $E^{\text{marker-bg}}_{\text{OOD}}$ environments. **ABA**

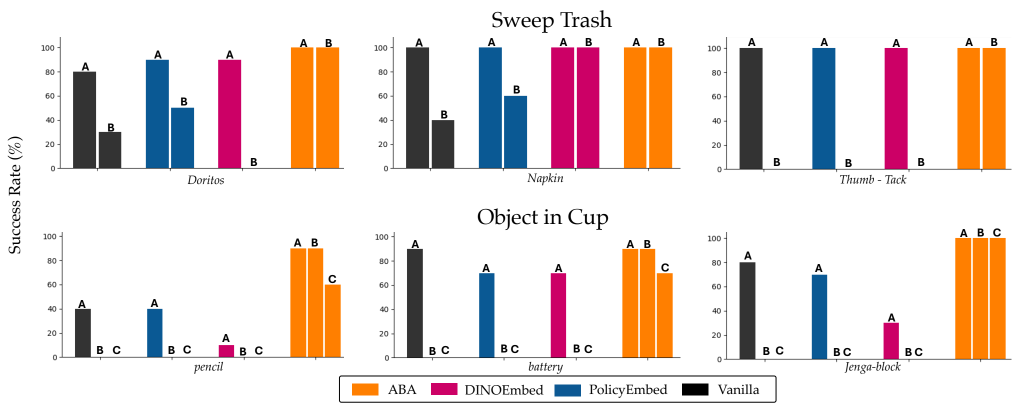

Figure 6: Subgoal Success in each OOD Environment with 3 novel objects for both sweep-trash and object-in-cup task. The success rate is averaged across 10 rollouts. **ABA** again consistently achieves the highest task success rate compared to baselines.

improves over **Vanilla** on both environments showing a 100% performance on all subgoals of both environments, except subgoal C of the $E_{\text{OOD}}^{\text{pen-bg}}$, where it shows an 80% success rate. This shows that a with novel background, **Vanilla** fails to even grasp the objects, however interventions with ID observations ignores the OOD conditions induced by the novel background, allowing **ABA** to uphold closed loop performance under the OOD environments.

Finally, on OOD environments induced by novel object categories for the sweep-trash task we observe from Fig. 6 that while **Vanilla** is able to align the wiper with the trash, it fails to pick the correct direction for sweeping the trash (subgoal B), as visual features are not enough to decide whether the trash is organic or recycling. **ABA** is able to successfully accomplish both subgoals for all novel objects as the relevant features for deciding the trash type are supplied by the expert as functional correspondences.

Since the object-in-cup task is more challenging, **Vanilla** is only performant at grasping (subgoal A). It is able to grasp the pencil with 40%, the battery with 90%, and the jenga-block with 80% success-rate. However, the sizes of the objects are such that they can only be dropped into the mug from the top (subgoal B), however **Vanilla** is not able to infer these features solely from the training data and hence fails at subgoal B and C. With **ABA**, the expert language feedback helps establish the correct functional correspondences, leading to an improvement in the performance across all subgoals.

3.2 What kind of features maximally improve the sub-goal level performance for observation interventions based methods?

As shown in Fig. 4, all intervention based method demonstrate a 100% task success on all subgoals of the sweep-trash task, in the ID environments. On the object-in-cup task intervention based methods again perform comparably on the E_{ID}^{marker}, however on the E_{ID}^{pen} both **PolicyEmbed** and **DINOEmbed** perform worse as compared to **ABA** on all subgoals.

As shown in Fig. 5, under a novel background, intervention based methods perform comparably on the sweep-trash task. On the object-in-cup task, **PolicyEmbed** performs worse compared to both **DINOEmbed** and **ABA**, whereas **DINOEmbed** performs comparably with **ABA**. This can be attributed to the ability of dino features to perform dense correspondence matching, specially across objects in the same semantic class.

Fig. 6 shows that under novel objects both **PolicyEmbed** and **DINOEmbed** struggle. Since **PolicyEmbed** relies on the policy embeddings, under 'doritos' and 'napkin' it sweeps them in either direction. **DINOEmbed** matches the visual features and since napkin closely resembles paper, it is able to correctly sweep napkin as recycling, and fails on other objects.

For the object-in-cup task, because policy embeddings and visual features alone are not enough to match the objects with the ID sample that lead to the desired behavior mode, both **PolicyEmbed** and **DINOEmbed** perform worse as compared to **ABA**.