

---

# Supplementary Material

---

Anonymous Author(s)

Affiliation

Address

email

## A More Ablation Studies

### A.1 Depth of Manipulation Concept Discovery Components

We present ablation experiments on the depth of neural networks used in Manipulation Concept Discovery (Sec. 3). Specifically, we study how varying the number of transformer blocks in Manipulation Concept Encoder  $\mathcal{E}$  (Eq. 1), Cross-Modal Correlation Network  $\mathcal{C}$  (Eq. 3), and Multi-Horizon Goal Predictor  $\mathcal{F}$  (Eq. 7) affects downstream policy performance.

Table 1: **Ablation experiments on transformer layers in Manipulation Concept Discovery (Sec. 3).** All experiments use diffusion policy (DP), with concepts discovered from and policies trained on LIBERO-90 task demonstrations.

Layers of $\mathcal{E}$	Layers of $\mathcal{C}$	Layers of $\mathcal{F}$	Success Rates (%)
12	4	4	89.6±0.6
12	8	8	89.4±0.6
6	4	4	83.6±0.8

The results in Tab. 1 suggest that the concept encoder  $\mathcal{E}$  depth impacts policy performance more significantly. However, we hypothesize that with scaled-up data and advanced base policies, deeper decoders ( $\mathcal{C}$  and  $\mathcal{F}$ ) may yield clearer benefits.

### A.2 Manipulation Concept Discovery Weighting

Here we present ablation experiments on the concept discovery weights in Eq. 8. Specifically, we study how different weight combinations in Eq. 8 affect downstream policies that are enhanced by manipulation concepts learned with these corresponding weight configurations.

Table 2: **Ablation experiments on the weights in Eq. 8.** We study the effects of different weight combinations in Eq. 8 by evaluating policies enhanced with concepts learned under various weight configurations. All experiments use diffusion policy (DP) as the base model, with concepts discovered from and policies trained on LIBERO-90 task demonstrations.

Weights	Success Rates (%)
$\lambda_{mm} = 1.0, \lambda_{mh} = 1.0$	89.6±0.6
$\lambda_{mm} = 0.1, \lambda_{mh} = 1.0$	88.3±0.5
$\lambda_{mm} = 1.0, \lambda_{mh} = 0.1$	88.2±0.6

The results in Tab. 2 demonstrate that enhanced policy performance remains similar across different weight combinations, indicating robustness in the weight configuration of Eq. 8. This robustness likely stems from the network’s adaptability to different learning rates, as weight adjustments effectively modify the learning rates of different components within the trainable network modules.

## B Analyzing Manipulation Concept Properties

### B.1 Enhanced Cross-Modal Correlation (Sec. 4.3)

We investigate whether the mask-and-predict strategy (Eq. 3) enhances cross-modal correlation learning using real-world data. Following the concept discovery setup from Sec. 4.4 (**Multi-Horizon Goal Prediction Visualization**), we analyze two modalities from BridgeDataV2 [1]: third-person agent view and robot proprioceptive state.

Table 3: **Conditional mutual information between modalities**, conditioned on manipulation concept latents from our method versus the **All** baseline that does not explicitly model cross-modal correlations.  $\mathbf{o}_A$ : agent view,  $\mathbf{o}_P$ : proprioceptive state.

	Ours	All
$\mathbb{I}(\mathbf{o}_P : \mathbf{o}_A \mid \mathbf{z})$	4.5390	2.9327

The results in Tab. 3 show that the conditional mutual information between these modalities (conditioned on manipulation concept latents) is higher when using our mask-and-predict approach compared to the baseline method **All** (Sec. 4.1), which lacks an explicit cross-modal correlation learning design. This confirms that the mask-and-predict strategy effectively enhances cross-modal correlation in manipulation concept latents for real-world scenarios.

### B.2 Alignment with Human Semantic Sub-Goals (Sec. 4.3)

We evaluate whether manipulation concept latents learned from real-world data resemble human-interpretable semantics. Specifically, we assess whether latents assigned to time steps of BridgeDataV2 [1] task demonstrations exhibit higher pairwise similarity when those steps belong to sub-processes pursuing the same human-defined sub-goal. The manipulation concept latents follow the setting of **Multi-Horizon Goal Prediction Visualization** in Sec. 4.4. To analyze the learned representations, we group manipulation concept latents according to human-annotated sub-goals, then quantify similarity between categories by calculating the average cosine similarity between their respective latents (Eq. 10).

Fig. 1 shows results from analyzing manipulation concept latents corresponding to demonstrations randomly selected from two types of tasks in BridgeDataV2: pick-and-place (grasp an object and place it elsewhere) and push-and-sweep (slide an object on a surface to another location). Manipulation concept latents are grouped based on human-defined sub-goals, with similarities between category pairs visualized as heatmaps. Three heatmaps are presented, each using a different granularity of sub-goal annotation:

- Top-1st heatmap** Preserves task-specific distinctions. For pick-and-place tasks, processes include grasp, transition, and release. For push-and-sweep tasks, processes depend on gripper contact: from above (press, sweep, release) or from the side (attach, push, release).
- Top-2nd heatmap** Merges similar manipulation processes in push-and-sweep tasks, combining processes where the gripper presses on top of the object while sliding it with processes where the gripper contacts the side of the object before pushing (represented as an attach-slide-release process).
- Top-3rd heatmap** Further merges differences between pick-and-place and push-and-sweep, generalizing the process to: attaching to the object with the gripper (whether grasping, pressing, or making side contact), transitioning the object to another location, and releasing the object.

In each heatmap, the entry at position  $(i, j)$  represents the average similarity between categories  $i$  and  $j$ . For readability, only the top three similarity values in each row are displayed.

We observe that the highest similarity values consistently appear along the diagonal in each heatmap in Fig. 1, indicating that concept latents from the same category show higher similarity compared to those from different categories. This suggests our model captures meaningful semantic structure in manipulation processes. Moreover, patterns across the three heatmaps with different description granularities reveal that the latents encode semantics at multiple abstraction levels. They capture both generalizable semantics applicable across tasks (e.g., “attach” in the Top-3rd heatmap for all situations

when the gripper contacts an object), while simultaneously preserving fine-grained scene-specific semantics (e.g., “grasp” in the Top-1st and Top-2nd heatmaps for specifically using the gripper to hold an object).

### B.3 Hierarchical Structure of Learned Concepts (Sec. 4.3)

To visualize the multi-horizon nature of our learned concepts, we analyze several demonstrations from LIBERO-90 by applying different coherence thresholds  $\epsilon$  to the demonstrations’ corresponding concept latents in Eq. 4. As shown in Fig. 2, 3, 4, and 5, varying  $\epsilon$  reveals a hierarchical structure in the concept space, with larger values producing coarser sub-processes that correspond to major task phases, and smaller values yielding finer-grained sub-processes that align with primitive manipulation actions. This hierarchical organization closely resembles human conceptualization of manipulation tasks, with actions naturally decomposing into meaningful subtasks at multiple levels of abstraction (as detailed in the captions of the respective figures).

We also conduct similar experiments on BridgeDataV2 [1] using manipulation concepts learned in the **Multi-Horizon Goal Prediction Visualization** experiment (Sec. 4.4). These results are presented in Fig. 6 and 7.

### B.4 Multi-Horizon Goal Prediction Visualization (Sec. 4.4)

To gain insight into the temporal information encoded in our manipulation concepts, we visualize outputs from our Multi-Horizon Goal Predictor (MHGP) network  $\mathcal{F}$  in Eq. 7. Experiments were conducted using the BridgeDataV2 dataset [1], with third-person visual input (preprocessed to 128x128 resolution) and proprioceptive state (7-DoF robot states).

Fig. 8 shows multiple examples of goal states predicted by the trained MHGP when conditioned on the current observations, the corresponding manipulation concept latents, and various coherence thresholds ( $\epsilon$ ). The network generates predictions that capture meaningful multi-horizon task progression, including anticipated object and arm positions at different future time points. These results confirm that our learned concepts encode meaningful temporal structures at multiple horizons, with increasing  $\epsilon$  values corresponding to predictions further into the future.

## C About Video Demonstrations

**Simulation Experiments** We provide demonstrations of ACT and DP (diffusion policy) policies enhanced with our learned manipulation concepts performing LIBERO-90 tasks, including both success and failure cases. These demonstrations are stored as .gif files in the directories `supplementary/rollout_video_samples_gif/ACT` and `supplementary/rollout_video_samples_gif/DP`.

**Real-World Experiments** We provide demonstrations of a policy enhanced with our learned manipulation concepts performing the real-world cleaning cup task described in **Real-world Generalization Study** in Sec. 4.4. These demonstrations are stored as .mp4 files (2x speed) in the directory `supplementary/rollout_video_samples_real`, with subdirectories labeled #1 through #6 corresponding to the six evaluation settings introduced in **Real-world Generalization Study** in Sec. 4.4. We also provide comparative examples with the baseline policy under `supplementary/rollout_video_samples_real/#5-barriers`, demonstrating how the baseline policy fails to adapt to novel scenes at specific sub-goal stages, indicating its lack of awareness of sub-goal completion. In contrast, the policy enhanced with our learned manipulation concepts successfully adapts to these novel situations, overcoming barriers. Additionally, we include interesting replanning cases of our enhanced policy in the directory `supplementary/rollout_video_samples_real/replanning`.

## References

- [1] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023.

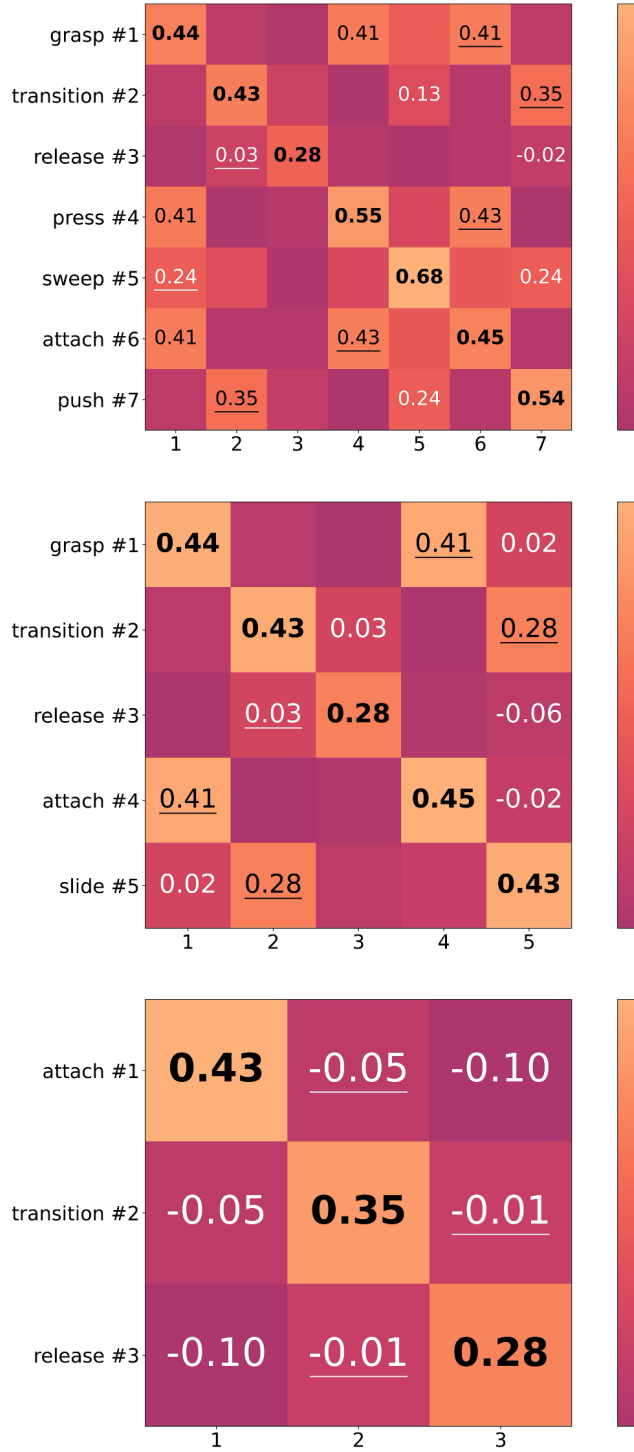


Figure 1: Average cosine similarity between pairs of sub-goal categories (defined by human semantics) computed using manipulation concept latents learned by our method. Each value at position  $(i, j)$  represents the average cosine similarity between latent vectors from the  $i$ -th and  $j$ -th categories. The three heatmaps represent increasingly abstract levels of semantic labeling.





Figure 2: **Multi-granular task decomposition through concept latent clustering.** Visualization of sub-processes derived by clustering manipulation concept latents using Eq. 4 at different coherence thresholds ( $\epsilon$ ) for the task **“Put the chocolate pudding in the top drawer of the cabinet and close it”**. Higher  $\epsilon$  values (top rows) produce coarser decompositions, while lower values (bottom rows) yield finer-grained segmentation. The emergent sub-processes naturally align with semantic task components. For example, the 3rd sub-process of row 3 corresponds to “transitioning the chocolate pudding above the top drawer”; the 4th sub-process of row 4 shows “releasing the chocolate pudding”; the 5th sub-process of row 4 represents “closing the top drawer”; the 5th sub-process of row 5 illustrates “reaching the handle of the top drawer”; the 6th sub-process of row 5 corresponds to “pushing the top drawer closed”.

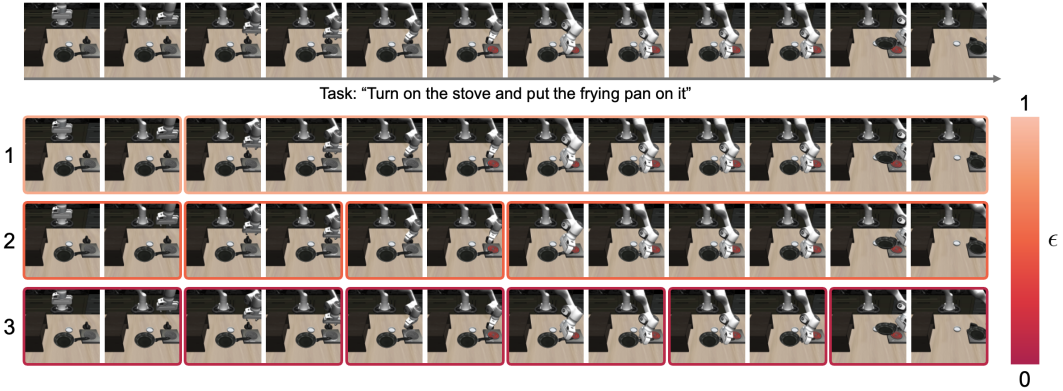


Figure 3: **Multi-granular task decomposition through concept latent clustering.** Visualization of sub-processes derived by clustering manipulation concept latents using Eq. 4 at different coherence thresholds ( $\epsilon$ ) for the task **“Turn on the stove and put the frying pan on it”**. Higher  $\epsilon$  values (top rows) produce coarser decompositions, while lower values (bottom rows) yield finer-grained segmentation. The emergent sub-processes naturally align with semantic task components. For example, the 1st sub-process of row 1 corresponds to “aligning the gripper with the knob of the stove”; the 2nd sub-process of row 2 represents “grasping the knob of the stove”; the 3rd sub-process of row 2 shows “rotating the knob of the stove”; the 4th sub-process of row 2 shows “putting the frying pan on the stove”; the 4th sub-process of row 3 shows “grasping the frying pan”; the 6th sub-process of row 3 shows “transitioning the frying pan onto the stove”.

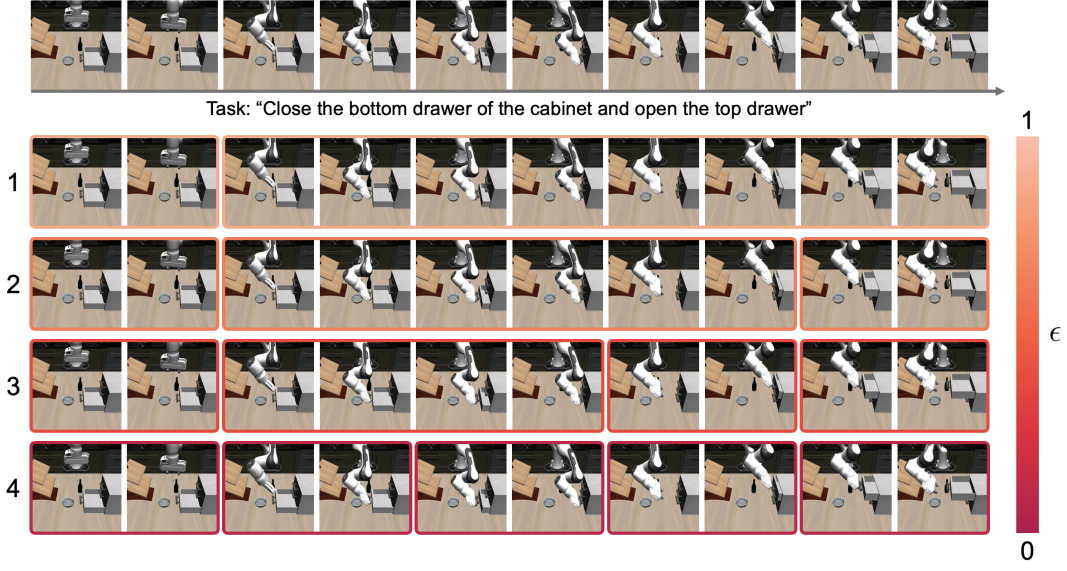


Figure 4: **Multi-granular task decomposition through concept latent clustering.** Visualization of sub-processes derived by clustering manipulation concept latents using Eq. 4 at different coherence thresholds ( $\epsilon$ ) for the task **“Close the bottom drawer of the cabinet and open the top drawer”**. Higher  $\epsilon$  values (top rows) produce coarser decompositions, while lower values (bottom rows) yield finer-grained segmentation. The emergent sub-processes naturally align with semantic task components. For example, the 1st sub-process of row 1 corresponds to “aligning the gripper with the handle of the bottom drawer”; the 2nd sub-process of row 3 represents “closing the bottom drawer”; the 3rd sub-process of row 3 shows “grasping the top drawer”; the 4th sub-process of row 3 shows “pulling the top drawer open”; the 2nd sub-process of row 4 shows “grasping the bottom drawer”; the 3rd sub-process of row 4 shows “pushing the bottom drawer closed”.

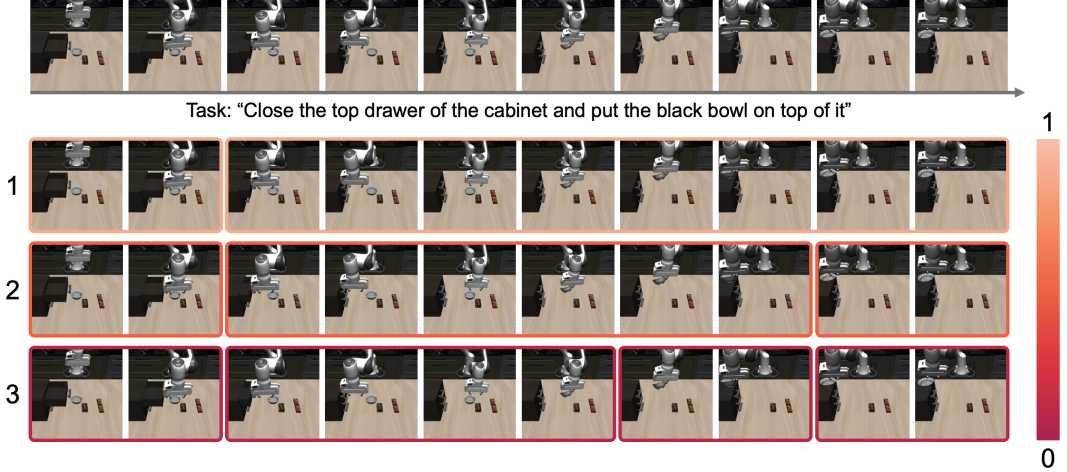


Figure 5: **Multi-granular task decomposition through concept latent clustering.** Visualization of sub-processes derived by clustering manipulation concept latents using Eq. 4 at different coherence thresholds ( $\epsilon$ ) for the task **“Close the top drawer of the cabinet and put the black bowl on top of it”**. Higher  $\epsilon$  values (top rows) produce coarser decompositions, while lower values (bottom rows) yield finer-grained segmentation. The emergent sub-processes naturally align with semantic task components. For example, the 1st sub-process of row 1 corresponds to “reaching the handle of the top drawer”; the 3rd sub-process of row 2 shows “releasing the black bowl”; the 3rd sub-process of row 3 shows “transitioning the black bowl on top of the cabinet”.





Figure 6: **Multi-granular task decomposition through concept latent clustering.** Visualization of sub-processes derived by clustering manipulation concept latents at different coherence thresholds ( $\epsilon$ ) for **pick-and-place** tasks (grasping an object and placing it in another location). Higher  $\epsilon$  values (top rows) produce coarser decompositions, while lower values (bottom rows) yield finer-grained segmentation. The emergent sub-processes naturally align with semantic task components. For the **first pick-and-place-pot** example: the 1st sub-process of row 1 corresponds to “grasping the pot”; the 2nd sub-process of row 1 corresponds to “placing the grasped pot”; the 2nd sub-process of row 2 corresponds to “transitioning the grasped pot”; the 3rd sub-process of row 2 corresponds to “releasing the grasped pot”; the 1st sub-process of row 3 corresponds to “aligning the gripper with the edge of the pot”; the 2nd sub-process of row 3 corresponds to “closing the gripper to grasp the pot”.



Figure 7: **Multi-granular task decomposition through concept latent clustering.** Visualization of sub-processes derived by clustering manipulation concept latents at different coherence thresholds ( $\epsilon$ ) for **push-and-sweep** tasks (attaching an object and sliding it to another location). Higher  $\epsilon$  values (top rows) produce coarser decompositions, while lower values (bottom rows) yield finer-grained segmentation. The emergent sub-processes naturally align with semantic task components. For the **third push-and-sweep-green-cloth** example: the 1st sub-process of row 1 corresponds to “pressing and sliding the cloth”; the 2nd sub-process of row 1 corresponds to “releasing the cloth”; the 1st sub-process of row 2 corresponds to “positioning and pressing on the cloth”; the 2nd sub-process of row 2 corresponds to “sliding the cloth”; the 1st sub-process of row 3 corresponds to “aligning the gripper with the cloth”; the 2nd sub-process of row 3 corresponds to “pressing on the cloth”.



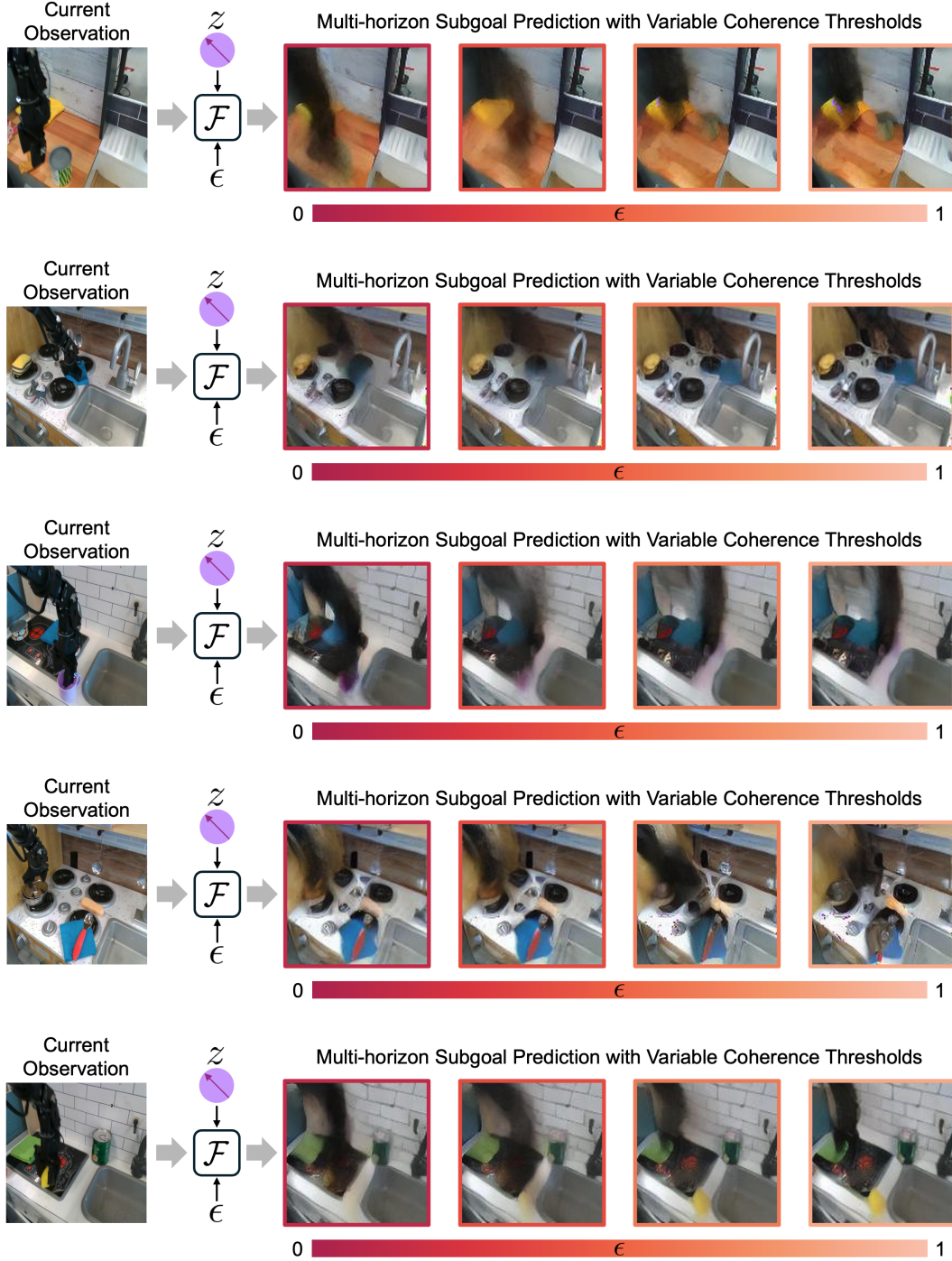


Figure 8: **Multi-horizon goal prediction from learned manipulation concepts.** We visualize the future goal state predictions generated by our Multi-Horizon Goal Predictor (MHGP,  $\mathcal{F}$  in Eq. 7), conditioned on the current observation, learned manipulation concept ( $z$ ), and varying coherence thresholds ( $\epsilon$ ). As  $\epsilon$  increases, predictions extend further into the future, illustrating the multi-horizon temporal abstraction captured by our manipulation concepts. The predicted goal states capture key task structures and functional relationships. For example, in the third row (robot picking and placing a purple cup): the first prediction shows a near-future state resembling current observation; the second prediction depicts robotic arm movement as a subtle shadow; the third and fourth predictions reveal both new arm position (indicated by stripe-like black shadowing) and purple cup trajectory (shown through purple motion blur).