A APPENDIX

A.1 MOTIVATING QUESTIONS

Inspired by the appendix of (Karamcheti et al., 2023), in this section, we list some motivating questions that may arise from reading the main paper.

Q1. *The experiments all consider preferences beyond task progress. If the end-user's preference is only progress, can RAPL achieve comparable performance compared to the SOTA TCC-based visual reward (Zakka et al., 2022)?*

To investigate this, we return to the X-Magical **grouping** task (middle plots in Figure 2) where a short stick robot needs to push two objects to goal. We removed the grouping preference so the ground truth task reward is consistent with the original benchmark in (Zakka et al., 2022). We trained RAPL with 150 preference queries and compare it with the TCC reward model trained using 500 demonstrations. In Figure 10, we show the policy evaluation success rate during policy learning. We see that **RAPL** has comparable final success rate compared to **TCC** and has a more stable policy training, showing that it can learn a superset of preferences when compared to TCC.

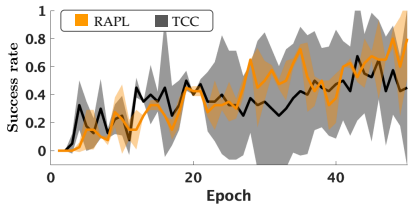

Figure 10: **X-Magical.** Progress-only reward success rate.

Q2. *What makes RAPL different from prior robot learning works that use optimal transport (OT) based visual rewards?*

Indeed, OT-based visual rewards have become increasingly popular for learning robot manipulation (Haldar et al., 2023b;a; Guzey et al., 2023). However, key to making the OT-based visual reward successful in Haldar et al. (2023b) is fine-tuning the representation model via behavior cloning tasks. This helps the model to capture some task-relevant information at the cost of requiring action labels. Furthermore, by relying on action labels, it is unclear if the learned reward can generalize to a different embodiment. Instead, our approach learns the representation only using preference queries (no action labels) and can generalize to embodiments.

Q2. *Why do **MVP-OT** and **TCC-OT** achieve near 0 success rate in the robot manipulation experiments in Figure 6 and Figure 9?*

Recall that both **MVP-OT** and **TCC-OT** use optimal transport to match the embedding distribution of the robot and the expert, but they vary which visual representation they use to obtain the embedding.

The MVP encoder is trained via masked autoencoding (He et al., 2022) to reconstruct heavily masked video frames. As such, it captures representations amenable to per-pixel reconstruction. Prior work (Karamcheti et al., 2023) has demonstrated that this representation struggles with higher-level problems (e.g., language-based imitation). We hypothesize this is why **MVP-OT** struggles to capture preference-relevant features and does not lead to aligned robot behaviors. Our results are also consistent with the experiments in (Haldar et al., 2023b) where an OT-based visual reward with a pre-trained MVP representation model gives near 0 success rate for manipulation tasks.

The TCC encoder is trained via temporal cycle-consistency constraints, and as such captures representations that encode solely task progress (e.g., distance to the goal image). Such a representation works well when goal reaching is the only preference of the end user. In our tabletop grouping task, the end user cares about goal reaching, but they also prefer moving the two objects together to goal region over moving the objects one-by-one. Thus if the robot happens to push one object towards the goal during policy learning, **TCC-OT** will reward this behavior (since this image is getting "closer" to the goal image) even though this is not preferred by the user.

A.2 EXTENDED RELATED WORK

Visual robot rewards promise to capture task preferences directly from videos. Self-supervised approaches leverage task progress inherent in video demonstrations to learn how "far" the robot

is from completing the task (Zakka et al., 2022; Kumar et al., 2023; Ma et al., 2023) while other approaches identify task segments and measure distance to these subgoals (Sermanet et al., 2016; Tanwani et al., 2020; Shao et al., 2020; Chen et al., 2021). However, these approaches fail to model preferences *during* task execution that go beyond progress (e.g., spatial regions to avoid during movement). Fundamental work in IRL uses feature matching between the expert and the learner in terms of the expected state visitation distribution to infer rewards (Abbeel & Ng, 2004; Ziebart et al., 2008), and recent work in optimal transport has shown how to scale this matching to high dimensional state spaces (Xiao et al., 2019; Dadashi et al., 2021; Papagiannis & Li, 2022; Luo et al., 2023). However, key to making this matching work from high-dimensional visual input spaces is a good visual embedding. Previous works used proxy tasks, such as behavior cloning (Haldar et al., 2023a;b) or temporal cycle-consistency learning (Dadashi et al., 2021), to train the robot's visual representation. In contrast to prior works that rely on hard-to-obtain action labels or using only self-supervised signal, we propose an OT-based visual reward that is trained purely on videos (no action labels needed) that ranked by the *end-user's* preferences.

Preference-based learning. While demonstrations have been the data of choice for reward learning in the past, an increasingly popular approach is to use preference-based learning (Christiano et al., 2017; Sadigh et al., 2017; Biyik & Sadigh, 2018; Wirth et al., 2017; Brown et al., 2019; Shin et al., 2023; Stiennon et al., 2020). Here the human is asked to compare two (or more) trajectories, and then the robot infers a map from ranked trajectories to a scalar reward. This feedback is often easier for people to give than kinesthetic teaching or fine-grained feedback (Shin et al., 2023). At the same time, prior works and our experiments show that directly predicting the reward from preference queries and high-dimensional input suffers from high sample inefficiency and causal confusion (Bobu et al., 2023b; Tien et al., 2022). To mitigate this issue, (Brown et al., 2020) augments multiple self-supervised objectives like inverse dynamics prediction or enforcing temporal cycle-consistency with the preference learning loss, but this requires additional signals like actions and the additional self-supervised objective may bias the learned rewards towards capturing spurious correlations.

Representation alignment in robot learning. Representation alignment studies the agreement between the representations of two learning agents. As robots will ultimately operate in service of people, representation alignment is becoming increasingly important for robots to interpret the world in the same way as we do. Previous work has leveraged user feedback, such as human-driven feature selection (Bullard et al., 2018; Luu-Duc & Miura, 2019), interactive feature construction (Bobu et al., 2021; Katz et al., 2021), or similarity-implicit representation learning (Bobu et al., 2023a), to learn aligned representations for robot behavior learning. But they either operate on a manually defined feature set or learning features in state space settings. In visual domain, (Zhang et al., 2020) directly uses per-image reward signal to align image representation with the human preference encoded in the reward signal, while directly accessing such a preference signal is not practical. Our work utilizes human preference feedback that naturally contains human preference to align robot's visual representations with the end user.

A.3 Optimal Transport Based Reward

Setup. Let $\mathbf{o} = \{o^t\}_{t=1}^{t=T}$ be a trajectory of observations, where T is the trajectory length. Let $\mathcal{D}_+ \subset \mathcal{S}_{\phi_\mathrm{H}}$ be a dataset of preferred videos from the preference video dataset and $\mathcal{D}_\mathrm{R}$ be the set of videos induced by a given robot policy π_R. We denote $\phi : \mathbb{R}^{h,w,3} \to \mathbb{R}^{n_e}$ as an observation encoder that maps a $h \times w$ RGB image to a n_e dimensional embedding. For any video $\mathbf{o}$, let the induced empirical embedding distribution be $\rho = \frac{1}{T}\sum_{t=0}^{T} \delta_{\phi_\mathrm{R}(o^t)}$, where $\delta_{\phi_\mathrm{R}(o^t)}$ is a Dirac distribution centered on $\phi_\mathrm{R}(o^t)$.

Background. Optimal transport finds the optimal coupling $\mu^* \in \mathbb{R}^{T\times T}$ that transports the robot embedding distribution, ρ_R, of a robot video $\mathbf{o}_\mathrm{R} \in \mathcal{D}_\mathrm{R}$ to the expert video embedding distribution, ρ_+ for $\mathbf{o}_+ \in \mathcal{D}_+$, with minimal cost (as measured by a distance function, e.g. cosine distance). This comes down to an optimization problem that minimizes the Wasserstein distance between the two distributions:

$$\mu^* = \underset{\mu\in\mathcal{M}(\rho_\mathrm{R},\rho_+)}{\arg\min} \sum_{t=1}^{T}\sum_{t'=1}^{T} c\big(\phi(o_\mathrm{R}^t), \phi(o_+^{t'})\big)\mu_{t,t'}. \tag{8}$$

where $\mathcal{M}(\rho_\mathrm{R}, \rho_+) = \{\mu \in \mathbb{R}^{T\times T} : \mu\mathbf{1} = \rho_\mathrm{R}, \mu^T\mathbf{1} = \rho_+\}$ is the set of coupling matrices and $c : \mathbb{R}^{n_\mathrm{R}} \times \mathbb{R}^{n_+} \to \mathbb{R}$ is a cost function defined in the embedding space (e.g., cosine distance).

The optimal transport plan gives rise to the following reward signal that incentivizes the robot to stay within the expert demonstration distribution by explicitly minimizing the distance between the observation distribution and expert distribution:

$$r(o_{\mathrm{R}}^t; \phi_{\mathrm{R}}) = -\sum_{t'=1}^{T} c\big(\phi_{\mathrm{R}}(o_{\mathrm{R}}^t), \phi_{\mathrm{R}}(o_{+}^{t'})\big)\mu_{t,t'}^{*}. \tag{9}$$

Regularized optimal transport. Solving the above optimization in Equation 8 exactly is generally intractable for high dimensional distributions. In practice, we solve a entropy regularized version of the problem following the Sinkhorn algorithm (Peyré et al., 2019) which is convex in μ and amenable to fast optimization:

$$\mu^* = \underset{\mu \in \mathcal{M}(\rho_{\mathrm{R}}, \rho_{+})}{\arg\min} \sum_{t=1}^{T}\sum_{t'=1}^{T} c\big(\phi(o_{\mathrm{R}}^t), \phi(o_{+}^t)\big)\mu_{t,t'} - \epsilon\mathcal{H}(\mu), \tag{10}$$

where $\mathcal{H}$ denotes the entropy term that regularizes the optimization and ϵ is the associated weight.

Choosing an $\mathbf{o}_+$ to match with π_{R}. The reward (9) requires matching the robot to an expert observation video. To choose this expert observation, we follow the approach from (Haldar et al., 2023a). During policy optimization, given a robot's trajectory's observation $\mathbf{o}_{\mathrm{R}}$ induced by the robot policy π_{R}, we select the the "closest" expert demonstration $\mathbf{o}_{+}^{*} \in \mathcal{D}_{+}$ to match the robot behavior with. This demonstration selection happens via:

$$\mathbf{o}_{+}^{*} = \underset{\mathbf{o}+\in\mathcal{D}_{+}}{\arg\min} \min_{\mu \in \mathcal{M}(\rho_{\mathrm{R}}, \rho_{+})} \sum_{t=1}^{T}\sum_{t'=1}^{T} c\big(\phi(o_{\mathrm{R}}^t), \phi(o_{+}^{t'})\big)\mu_{t,t'}. \tag{11}$$

A.4 Attention Map for RAPL and RLHF

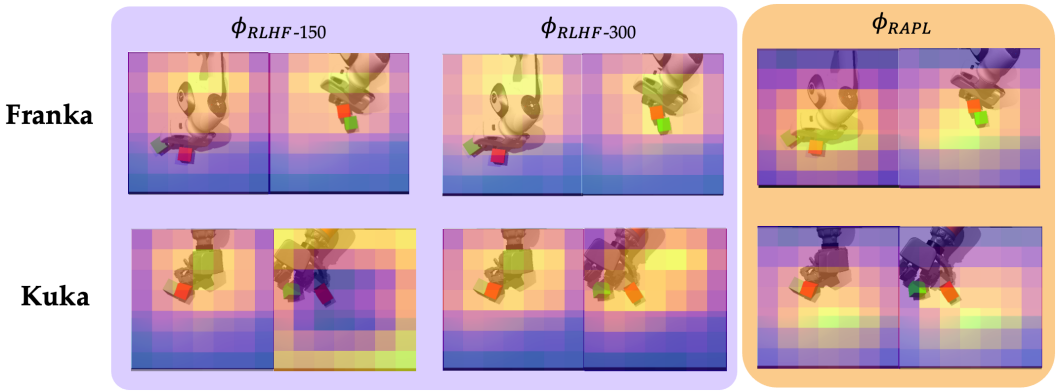

Figure 11: **Manipulation: Attention Map.** Visualization of attention map for RLHF-150 demos, RLHF-300 demos, and RAPL with 150 demos for both Kuka and Franka (cross-embodiment) images. Each entry of the figure shows two image snapshots from the relevant demonstration set with the attention map overlaid. Bright yellow areas indicate image patches that contribute most to the final embedding; darker purple patches indicate less contribution.

A.5 Additional RLHF results

In Sec.5.2, it's surprising that RLHF fails to learn a robot policy in a more realistic environment since its objective is similar to ours, but without explicitly considering representation alignment. To further investigate this, we apply a linear probe on the final embedding and visualize the image heatmap of what RAPL's (our representation model trained with 150 training samples), RLHF-150's (RLHF trained with 150 samples), and RLHF-300's (RLHF trained with 300 samples samples) final embedding pays attention to in Figure 11.

We see that ϕ_{RAPL} learns to focus on the objects, the contact region, and the goal region while paying less attention to the robot arm; $\phi_{RLHF-300}$ is biased towards irrelevant areas that easily induce "spurious" correlations such as the robot arm and background are; $\phi_{RLHF-300}$'s attention is slightly shifted to objects while still pays high attention to the robot embodiment.

When deploying $\phi_{RLHF-300}$ in Franka manipulation policy learning, we observe that policy performance is slightly improved (indicating that with more feedback data, preference-based reward prediction could yield to an aligned policy), but **RAPL** still outperforms **RLHF** by 75% with 50% less training data, supporting the hypothesis: ***RAPL*** *outperforms* ***RLHF*** *with lower amounts of human preference queries*.

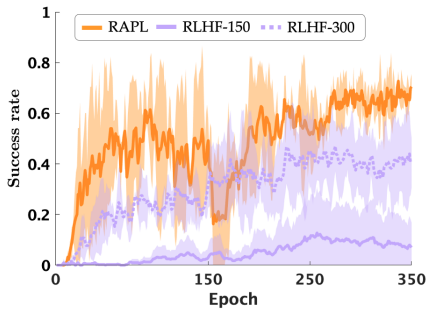

Figure 12: **Manipulation.** **RAPL** outperforms **RLHF** by 75% with 50% less training data.

A.6 Additional Cross-Embodiment Learning Results

Figure 13 shows the rewards over time for the three cross-embodiment video observations (marked as preferred by the end- user's ground-truth reward or disliked) in the avoid (left) and group task (right). Across all examples, **RAPL** 's rewards are highly correlated with the **GT** rewards even when deployed on a cross-embodiment robot.

Figure 14 shows the rewards over time for the two cross-embodiment video observations (marked as preferred by the end- user's ground-truth reward or disliked). Across all examples, **RAPL** 's rewards are highly correlated with the **GT** rewards even when deployed on a cross-embodiment robot.

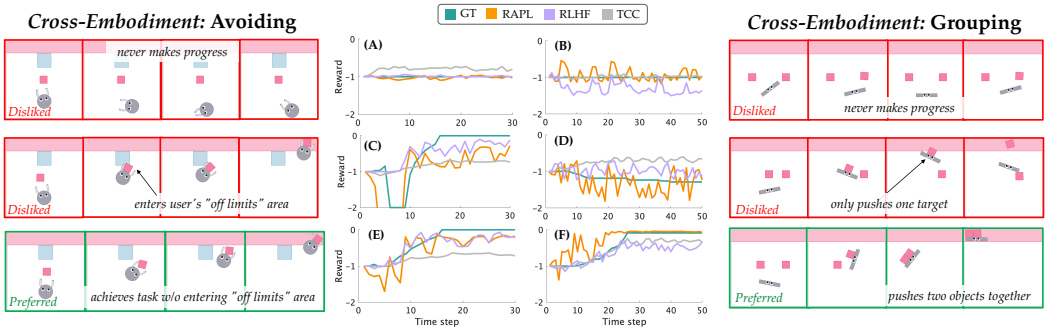

Figure 13: **Cross-Embodiment: X-Magical.** **RAPL** can distinguish preferred and disliked videos in the cross-embodiment setting.

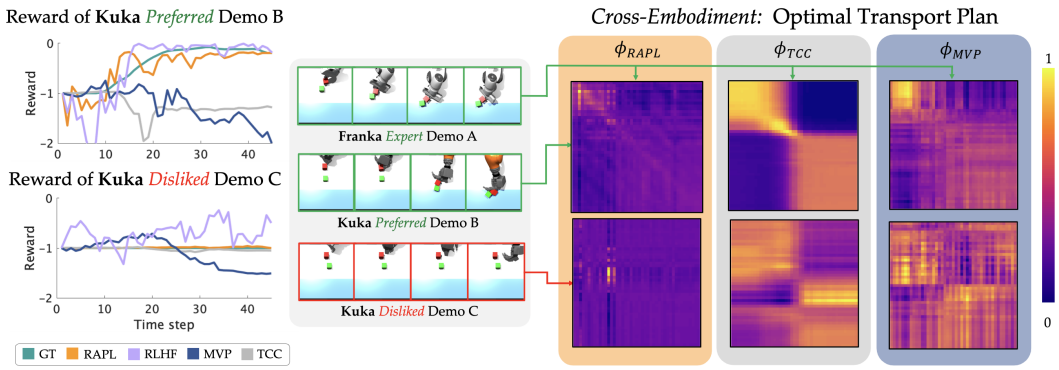

Figure 14: **Cross-Embodiment: Manipulation.** (left) Reward over time for a Kuka preferred and disliked video. (center) Expert video on *Franka* robot, preferred video on *Kuka*, and disliked *Kuka* video demo. (right) OT plan for each representation. Columns are embedded frames of expert demo. Rows of top matrices are embedded frames of preferred demo; rows of bottom matrices are embedded frames of disliked demo. Peaks exactly along the diagonal indicate that the frames of the two videos are aligned in the latent space; uniform values in the matrix indicate that the two videos cannot be aligned (i.e., all frames are equally "similar" to the next). RAPL matches this structure: diagonal peaks for expert-and-preferred and uniform for expert-and-disliked, while baselines show diffused values no matter the videos being compared.