2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Hierarchical Human-to-Robot Imitation Learning for Long-Horizon Tasks via Cross-Domain Skill Alignment

Zhenyang Lin[1], Yurou Chen[1] and Zhiyong Liu[1], *Senior Member, IEEE*

***Abstract*— For a general-purpose robot, it is desirable to imitate human demonstration videos that can effectively solve long-horizon tasks and perform novel ones. Recent advances in skill-based imitation learning have shown that extracting skill embedding from raw human videos is a promising paradigm to enable robots to cope with long-horizon tasks. However, generalization to unseen tasks in a different domain with a human prompt video poses a significant challenge due to the big embodiment and environment difference. To this end, we present Hierarchical Human-to-Robot Imitation Learning (H2RIL) that learns the mapping of cross-domain sensorimotor skills and utilizes it to generalize to unseen tasks given a human video in a different environment. To allow for generalizing zero-shot across environments and embodiments, H2RIL leverages task-agnostic play data for low-level policy training and paired human-robot data for both semantic and temporal skill embedding alignment. Extensive experiments in a simulated kitchen environment demonstrate that H2RIL significantly outperforms other prior baselines and is capable of generalizing to composable new tasks and adapting to Out-of-Distribution (OOD) tasks.**

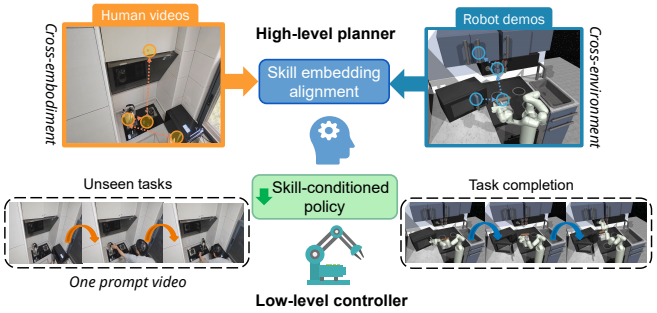


Fig. 1. **Hierarchical Human-to-Robot Imitation learning.** We present H2RIL, a new skill-based imitation learning framework that leverages cross-domain human videos to effectively generalize to unseen long-horizon tasks. First, we extract task-relevant skill embeddings that capture the behavior patterns and jointly learn the low-level skill-conditioned policy from on-robot trajectories. Given paired human-to-robot data, H2RIL further trains a robust cross-domain skill mapping policy, with objectives to ensure both semantic and temporal skill embedding alignment.

I. INTRODUCTION

A long-standing challenge in robot learning is to imitate human behaviors from easy-to-collect human videos to effectively solve long-horizon tasks. The crux to this challenge lies in not only comprehending task representations semantically but also mapping them to skill memories from prior experience. Imitation Learning (IL) has recently fueled great progress towards enabling agents to tackle a variety of skills [1], [2]. Despite this promise, these IL algorithms have been confined to relatively short-horizon tasks due to their data-hungry nature. Realizing this limitation, a stream of work resorts to Hierarchical Imitation Learning (HIL) which leverages the skill priors [3], [4] or retrieval-based data augmentation [5] to improve sample efficiency. Yet, long-horizon robot demonstrations of new tasks can also be difficult and expensive to obtain especially in tricky real-world environments. These considerations motivate our research problem: *how can we encapsulate skill memories into cross-domain human videos in order to generalize to unseen long-horizon tasks in a user-friendly manner?*

To scale up imitation learning to general long-horizon manipulation, recent advances have shown that visually imitating humans is a promising alternative to solving novel tasks. By explicitly extracting human priors from human videos [6], [7] or implicitly aligning the representation of human videos with robot demonstrations [8], the continuous skill embedding of task specification enables agents to perform goal-directed control and generalize few-shot to new tasks by giving human prompt videos of the new task at test time. However, the generalization capability of visual imitation learning is limited due to the issue of domain mismatch including different embodiments and environments. Achieving this goal is required not just to identify inherent semantic structure from cross-domain task specifications, but also to implicitly distill fine-grained knowledge of temporally extended skills into target human videos.

In this paper, we develop **H**ierarchical **H**uman-to-**R**obot **I**mitation **L**earning (**H2RIL**), a new skill-based imitation learning algorithm using cross-domain demonstrations that enables improvement in generalization to unseen long-horizon tasks given a target human video. We consider a setting with significant embodiment mismatches and environmental variations between human videos and on-robot demonstrations. The focus of H2RIL is the setup depicted in Fig. 1. To acquire a skill representation that is suitable for capturing behavior modes, H2RIL first learns uni-modal skill distributions conditioned on behavior prior via using sub-trajectories of in-domain task-agnostic play data. Conditioned on the predictable skill embedding, low-level policies can in principle enable robots to perform guided motion generation and learn generalizable long-horizon manipulations. To bridge the domain gap between human videos

[1]Zhenyang Lin, Yurou Chen and Zhiyong Liu are with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China. Corresponding author: zhiyong.liu@ia.ac.cn

[1]Project website can be found at https://tobbylincasia.github.io/H2RIL/

and robot demonstrations, we propose a temporal sequence contrastive learning objective for cross-modal paired skill sequences to learn the fine-grained mapping of the temporal structure from human videos to on-robot demonstrations after a coarse semantic alignment of human gaze-based subtask decomposition. Using this mapping, H2RIL learns a policy conditioned on clips of human videos as skill specifications to perform novel long-horizon tasks.

We demonstrate on several challenging long-horizon tasks in a kitchen environment that the proposed H2RIL algorithm clearly improves the generalization ability to a new task in a cross-domain scenario compared with relevant baselines in imitation learning with human videos. Through comprehensive analysis, we highlight the role that our temporal sequence contrastive learning objective plays in the robust alignment of human videos and heterogeneous robot demonstrations, which facilitates both cross-domain skill transfer and composition for long-horizon tasks.

To summarize, our contributions are threefold:

- We present H2RIL, a new paradigm for hierarchical imitation learning with human videos that leverages task-agnostic play data for skill embedding acquirement and paired cross-domain data for skill alignment. Our H2RIL is capable of generalizing to new long-horizon tasks given a demonstration across substantially differing domains.
- We introduce a temporal sequence contrastive learning objective to learn a robust mapping between the human domain and task-relevant skill embedding.
- Extensive experiments on several challenging manipulation tasks show that our method shows stronger generalization than prior skill-based imitation learning with human videos.

II. Related Work

A. Imitation Learning

IL methods can be formulated as a distribution-matching problem with offline robot demonstrations generated by an expert agent, which enables robots to successfully accomplish a myriad of manipulation tasks [9]–[14]. Two typical branches of IL are behavioral cloning (BC) [15], [16] and generative adversarial imitation learning (GAIL) [17]. While promising, both BC and GAIL methods fall short of tackling long-horizon manipulation tasks since a robust policy requires considerable sources of supervision for each task and is prone to drifting away from downstream demonstrations. To address these considerations, several lines of work seek to leverage large "in-the-wild" datasets to facilitate downstream task learning. The form of datasets can be human videos [18], or a combination of natural language and human videos [19]. However, the generalization ability can be hindered when there is a substantial domain shift between pre-trained data and target tasks. In this paper, we aim to bridge the domain gap by learning a cross-domain skill mapping policy through a handful of paired human videos and robot demonstrations, which can in principle alleviate the requirement for on-robot data when generalizing to new long-horizon tasks.

B. Skill-based Imitation Learning

Skill-based imitation learning is commonly formulated as a hierarchy of policies with low-level controllers and high-level skills. A number of recent research have focused on hard-coding prior knowledge into discrete high-level skills [20] and learning latent skill representation of sub-trajectories segmented either in an unsupervised fashion [21]–[23] or relying on additional supervision to enable more efficient imitation learning [24], [25]. Another stream of work embeds fixed-length sub-trajectories via variational autoencoder-based methods [4], [5], [26]–[28], where the skill embedding captures the full range and fidelity of behaviors and exhibits flexibility. While promising, most prior methods require on-robot demonstrations of target tasks when generalizing to new long-horizon tasks. In contrast, we resort to a cost-effective human video in lieu of retrieving from in-domain demonstrations to perform zero-shot generalization to a new task.

C. Learning from Human Videos

A large field of robot learning has paid significant attention to learning robotic policies from human videos since they are easy to collect at scale. One main thread of methods is to explicitly extract human priors such as human hand trajectories [6], [7], [29], [30], thus obtaining transferable trajectory-level plans for long-horizon task execution. On the other hand, implicitly aligning human videos and robot demonstrations has also been explored for acquiring a robust mapping tackling with cross-domain scenarios. One such method [31] is to learn to convert a demonstration from one context to another context, while others attempt to learn a shared embedding space through skill prototypes [32] and generative adversarial training [33]. However, such methods mainly focus on overcoming embodiment differences and relying on environments being similar. To reduce the domain gap, [34] learns a reward function using "in-the-wild" human videos and thus enables generalization to new tasks, while [8] performs cross-domain state matching to enable high-level semantic skill transfer. Similar to our work, [35] learns a video embedding conditioned policy that leverages the task semantic similarity by contrastive representation learning to retrieve the skill embedding from the robot domain. Unlike these prior works which perform coarse semantic alignment, H2RIL introduces a temporal sequence contrastive learning objective to capture fine-grained knowledge beyond task semantics which facilitates generalization to a different domain.

III. METHOD

Our goal is to encapsulate extracted skill memories into cross-domain human videos to endow robots with the capability of video-guided generalization to new long-horizon tasks. We aim to leverage a provided human video in a different environment to follow the performed skills distilled in on-robot demonstrations. To that end, we decompose the problem of hierarchical human-to-robot imitation learning into two sub-stages: (1) extracting skill embeddings that

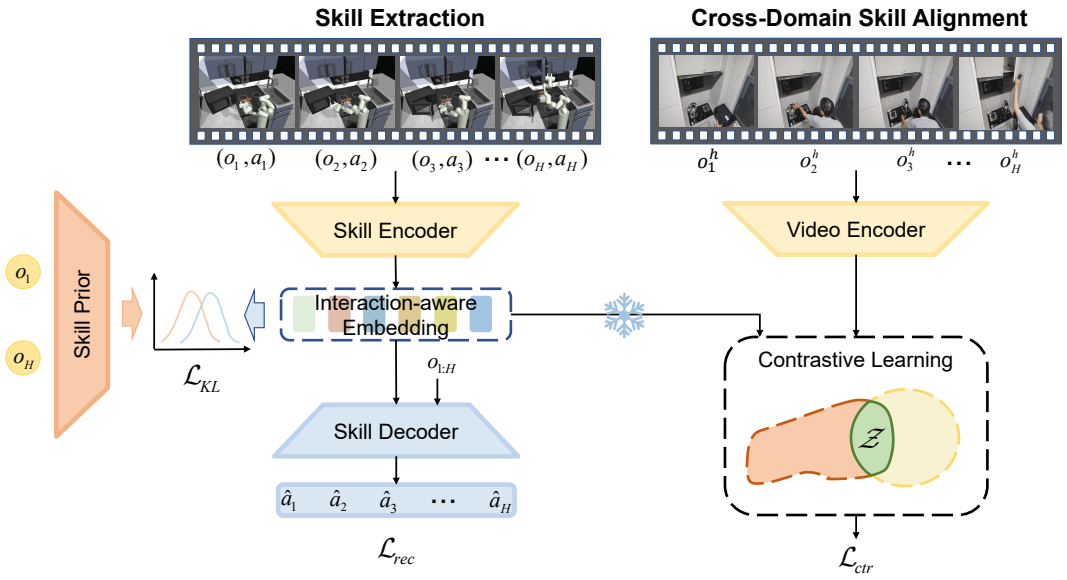


Fig. 2. **Model Architecture.** H2RIL is composed of a skill extraction (left) and cross-domain skill alignment (right) phase. In the skill extraction phase, we learn a trackable latent skill representation and jointly train a low-level policy via a variational autoencoder. In the cross-domain skill alignment phase, a cross-domain skill mapping policy modeled as a transformer-based video encoder is trained to align the cross-domain data both semantically and temporally with the proposed temporal sequence contrastive learning objective implemented.

contain task-relevant information from on-robot data, and (2) finding a robust mapping between human videos and learned skill embeddings with paired data sources.

A. Problem Formulation

Formally, the target task learning problem is formulated as a Markov Decision Process (MDP) specified by the tuple $\{\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma\}$: the state space, action space, transition distribution, reward function, and discount factor respectively. Our objective is to learn a human video-conditioned policy $\pi(a|o, v^h)$ that maximizes the discounted sum of rewards, where $a \in \mathcal{A}$ denotes actions, $o \in \mathcal{O}$ denotes the observations, and $v^h \in \mathcal{V}$ represents the human video of target task in a different environment. In the phase of skill extraction, we assume access to a task-agnostic robot teleoperation dataset $\mathcal{D}^r$ of previous robot interactions with the environment in the form of N variable-length trajectories $\{\{o_0, a_0, o_1, \ldots, o_{T_i}, a_{T_i}\}\}_{i=1}^N$. We aim to leverage the reward-free data to discover the skill embedding space $\mathcal{Z}$ which encapsulates the semantically meaningful fixed-length sub-trajectories $\tau_t = \{o_t, a_t, \ldots, o_{t+H-1}, a_{t+H-1}\}$ and learn the skill-conditioned low-level policy $\pi(a|o, z)$, where skill embeddings $z \in \mathcal{Z}$ capture versatile behavior patterns and are composable to perform new tasks.

In the phase of cross-domain skill alignment, we are given two small paired task-specific demonstration dataset $\mathcal{D}^r_{demo}$ and human video dataset $\mathcal{D}^h_{demo}$, both of which are labeled with subtasks. Such labels can be obtained in a number of approaches, e.g. via human supervision, and action recognition [36]. Our goal is to leverage the paired cross-modal data to learn the cross-domain skill mapping π_{map}: $\mathcal{M}(\mathcal{V}) \rightarrow \mathcal{Z}$, where $\mathcal{M}$ indicates that is the set of both semantic and temporal alignment operations on human videos. At inference, H2RIL takes as input a human video v^h_{prompt} for a new target task in a different environment. Based on this video serving as the task specification, the approach first maps sub-clips of the video to skill memories, and subsequently performs the skill using the learned skill-conditioned policy $\pi(a|o, z)$.

B. Interaction-aware Skill Embedding Extraction

A key to efficiently encapsulating skill memories into human videos is sufficient expressiveness of the skill representation distilled from task-agnostic on-robot data. Off-the-shelf visual representation learning has difficulties in capturing object-agent interactions since it lays more emphasis on the appearance and texture of the agent and background. To obtain interaction-aware skill embeddings, we adopt a continuous latent variable model to encode fixed-length sub-trajectories, which flexibly decomposes skills to enable long-horizon task completion.

Fig. 2 left depicts the training setup for skill-conditioned policy $\pi(a|o, z)$. This skill extraction model is composed of two modules: the skill encoder q_ϕ and the low-level skill decoder π_θ. Given a sub-trajectory $\tau_t = \{o_t, a_t, \ldots, o_{t+H-1}, a_{t+H-1}\}$ with fixed length H, we use a long short-term memory (LSTM) to embed the sub-trajectory consisting of interaction features and predict parameters of a Gaussian distribution $q_\phi(z|\tau_t)$. Conditioned on the skill embedding z, our skill decoder is modeled as an MLP-based network, which takes the corresponding observations o_t as inputs and aims to reconstruct the action $\hat{a}_t$. Inspired by [27], [37], we also implement the learned conditional prior with parameters ψ to extract the distribution of initial and goal states from the same sub-trajectory τ_t. Overall, our skill

Algorithm 1 H2RIL Algorithm

1: Pre-train the skill encoder $q(z|\tau)$.
2: Pre-train the skill-conditioned policy $\pi(a|o,z)$.
3: Semantically match human videos to on-robot demonstrations.
4: **for** each training iteration **do**
5: Sample $\tau_x = \{(o_t, a_t)\}_{t=x}^{x+H-1} \sim \mathcal{D}^r_{demo}$
6: Sample aligned human video clip $v_x, v_x^+, v_x^- \sim \mathcal{D}^h_{demo}$
7: Update the cross-domain skill mapping policy $\pi_{map}(v_x)$ with Eq. (2) and (3)
8: **end for**
9: **return** cross-domain skill mapping policy $\pi_{map}(\cdot)$

extraction objective is to minimize the loss:

$$\mathcal{L}_{skill} = -\mathbb{E}_{z\sim q_\phi(z|\tau_k)} \underbrace{\left[\sum_{t=k}^{k+H-1} \log \pi_\theta(a_t|o_t,z)\right]}_{\mathcal{L}_{rec}} + \beta \cdot \underbrace{KL(q_\phi(z|\tau_k)\|p_\psi(z|o_k, o_{k+H-1}))}_{\mathcal{L}_{KL}} \quad (1)$$

Where the $KL(\cdot)$ represents the Kullback-Leibler (KL) divergence and β is a weight factor for the regularization term. In this way, our skill extraction approach is capable of capturing the continuous interaction-aware skill embedding, while discarding the irrelevant features of appearance and texture. This form of skill representation is tailored for cross-domain alignment, allowing us to use this embedding as an anchor in subsequent alignment operations.

C. Cross-Domain Skill Alignment

Our goal is to leverage the learned skills for generalizing to new long-horizon tasks guided by a human video. Crucially, we introduce a cross-domain skill mapping policy beyond task semantics between human videos and on-robot demonstrations, which is required to be independent of the domain and focus on features of object-agent interactions. Thus, we can perform a coarse alignment of subtask decomposition and further a fine-grained temporal alignment over the skill embedding space $\mathcal{Z}$.

Many choices of video alignment approaches [38]–[40] can be used to match paired videos. However, these methods are not suitable for our task since there is no synchronization information (multiple cameras recording the same event). Following this intuition, we first perform a coarse semantic pre-alignment to match the paired datasets in sequence length and enable alignment of significant variations within a subtask category, which is optional for event understanding [41]. Specifically, for a given human video $v \in \mathcal{D}^h_{demo}$ with recorded subtask labels s_j and according number of frames m_j, we perform uniform subsampling of sequences for each subtask s_j, with the length n_j of the sub-sequences matching those in the corresponding subtasks of the paired on-robot demonstration from $\mathcal{D}^r_{demo}$.

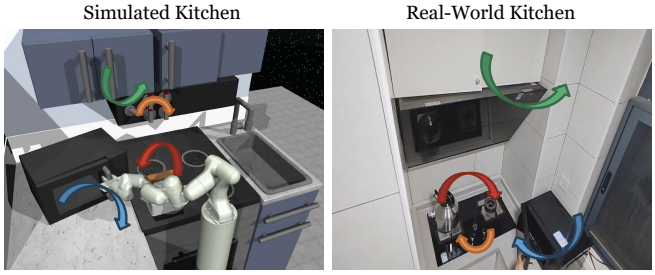


Fig. 3. **Cross-Domain Scenarios and Evaluation Tasks.** (**Left**) Franka Kitchen: a complete task consists of a permutation of four subtasks. We consider six sub-tasks: open the microwave (M), move the kettle (K), turn on the top burner (T), turn on the bottom burner (B), turn on the light switch (S), and open the hinge cabinet (H). (**Right**) Real-World Kitchen: there are similar objects to interact with, but the arrangement and the embodiment significantly differ from the simulated kitchen.

Inspired by [42], we introduce a temporal sequence contrastive learning objective to discover the fine-grained mapping of the temporal structure between domains. As is illustrated in Fig. 2 right, the embedding of a human video clip v_x from time x in video v is represented as $f_\omega(v_x)$, where $f(\cdot)$ is complemented by a transformer-based video encoder [43]. The objective ensures that the embedding of an anchor clip v_x is similar to the positive clips v_x^+ sampled within the positive window w_+, while far apart from the negative samples v_x^- sampled beyond the negative window w_- in the skill embedding space. Besides, we align the human video clip and pairwise on-robot sub-trajectory τ_x in the pre-trained skill embedding space $\mathcal{Z}$ that encodes many different skills. We learn the cross-domain skill mapping policy π_{map} by minimizing the temporal contrastive learning objective:

$$\mathcal{L}_{ctr} = KL\left(f_\omega(v_x)\|q(\tau_x)\right) + max\left[0, \alpha + z_{metric}\right] \quad (2)$$

$$z_{metric} = \left\|\mu(f_\omega(v_x)) - \mu(f_\omega(v_x^+))\right\|_2^2 - \left\|\mu(f_\omega(v_x)) - \mu(f_\omega(v_x^-))\right\|_2^2 \quad (3)$$

Here, $\mu(\cdot)$ denotes taking the mean of the distribution and α determines the margin that acts on the positive and negative samples. By mapping cross-domain human videos to on-robot skill embeddings via $f_\omega(\cdot)$, we create a proxy dataset of target on-robot trajectories, which is used to generalize to new downstream tasks with skill-conditioned policy $\pi(a|s,z)$. We summarize the skill embedding extraction and cross-domain skill alignment steps in Algorithm 1.

IV. EXPERIMENTS

Through our experiments, we aim to answer the following questions:

- Can the H2RIL algorithm improve performance and generalization compared to prior skill-based imitation methods in long-horizon manipulation tasks?
- How does the task-relevant skill embedding in H2RIL align with human videos?
- Do our semantic alignment and the temporal sequence contrastive learning objective lead to a better mapping policy?

TABLE I

QUANTITATIVE EVALUATION RESULTS IN THE KITCHEN ENVIRONMENT IN TERMS OF THE AVERAGE REWARD.

Task	**No Switch**			**No Topknob**		
	seen	composable	unseen (fine tune)	seen	composable	unseen (fine tune)
	KBTH	MKBT	MKSH	KBSH	MKBS	MKTH
FIST (aligned)	2.9 ± 0.62	1.0 ± 0.8	2.2 ± 0.56	1.2 ± 0.3	0.8 ± 0.5	2.2 ± 0.56
DVD	3.6 ± 0.5	2.1 ± 0.23	3.6 ± 0.6	3.2 ± 0.74	2.2 ± 0.74	3.2 ± 0.66
H2RIL-R3M	3.7 ± 0.48	3.0 ± 0.0	1.8 ± 0.3	2.8 ± 0.66	2.0 ± 0.47	1.3 ± 0.35
H2RIL-3DCNN	3.8 ± 0.45	3.0 ± 0.48	3.2 ± 0.74	3.2 ± 0.45	2.8 ± 0.3	3.2 ± 0.74
H2RIL-ViViT	$\mathbf{3.8 \pm 0.3}$	$\mathbf{3.1 \pm 0.41}$	$\mathbf{3.8 \pm 0.45}$	$\mathbf{3.7 \pm 0.48}$	$\mathbf{2.9 \pm 0.23}$	$\mathbf{3.3 \pm 0.48}$

A. Experimental Setup

1) Environment: We evaluate our H2RIL method on a simulated kitchen environment [44] that involves 7 subtasks and comes with approximately 600 on-robot trajectories. We consider two task-agnostic robot datasets $\mathcal{D}^r$ excluded interactions with the switch (**No Switch**) and top burner (**No Topknob**), which have 91 and 62 demonstrations respectively. To obtain cross-domain demonstrations, we collect two sets of 15 task-specific demonstrations $\mathcal{D}^r_{demo}$ and an equal number of human videos $\mathcal{D}^h_{demo}$ demonstrating the same permutation of subtasks in a different real kitchen environment as illustrated in Fig. 3. During the execution, the robot is required to complete a new composition of subtasks after giving a human prompt video with subtask labels as a task specification and achieve a new task including the excluded subtask after fine-tuning with paired cross-domain data.

2) Baselines: We compare our H2RIL against skill-based imitation leaning methods:

- **FIST (aligned)** Few-shot Imitation with Skill Transition Models, an in-domain few-shot imitation method leverages a semi-parametric policy $p(z|o_t, o_g)$ to select skills that will take the agent to the desired future state [4]. We perform the alignment by replacing the current image o_t and goal image o_g with the corresponding frames in the paired human video.
- **DVD** Domain-agnostic Video Discriminator trains a functional similarity between robot behaviors and human video clips [34], which can be used for skill retrieval to execute downstream tasks conditioned on a human video [35].
- **H2RIL (ours)** extracts task-relevant skill embedding and learns a cross-domain skill mapping policy, where the video encoder can be implemented by different models, such as pre-trained **R3M** [19], **3DCNN** [45], and **ViViT** [46].

3) Implementation Details: The fixed length H is set as 10 for both on-robot trajectories and human video clips. During the semantic alignment phase, we divide each subtask into an approaching process and an interaction process such as "move to microwave" and "open microwave", where a complete task consists of 4 subtasks. Besides, we use the following hyper-parameters: the regularization weight $\beta = 0.05$, the margin $\alpha = 3.0$, the positive window $w_+ = 10$, and the negative window $w_- = 30$.

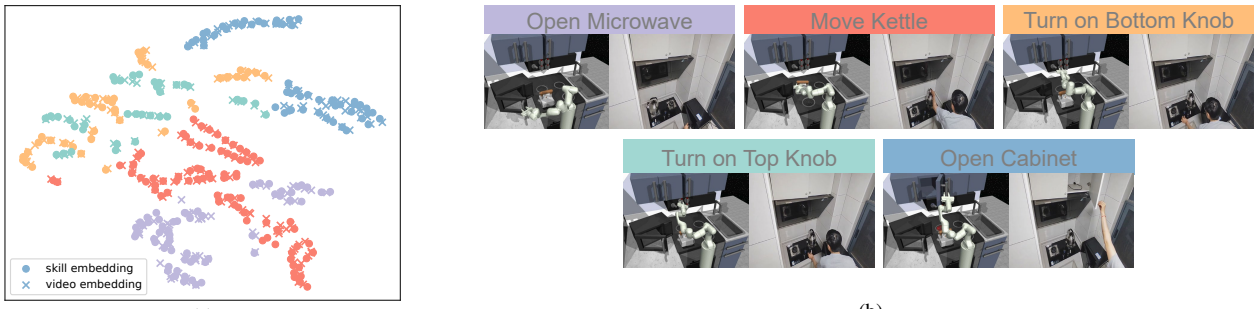
B. Quantitative Evaluation

The primary experimental question in this paper is whether the proposed H2RIL method is capable of solving downstream long-horizon tasks conditioned on a prompt human video. We compare our method to prior skill-based imitation learning with human videos in three levels of protocols: a **seen** setting, in which the whole task is seen during skill extraction, a **composable** setting, in which the permutation of the 4 subtasks are unseen and the method is required to generalize zero-shot to new tasks with a human video, and an **unseen** setting, in which the part of subtasks remains unavailable in the skill embedding space and 5 paired on-robot demonstrations and human videos of the target task are provided for fine-tuning. During inference, the reward is the number of subtasks completed in order.

The final success rate for baselines and H2RIL are shown in Table I. Note that the metrics we compared here are the average success rate after 10 rollouts. It can be seen that H2RIL significantly outperforms other baselines with an average success rate of 3.4 in all protocols. FIST (aligned) fails to leverage cross-domain human videos and makes no progress on skill alignment without access to on-robot demonstrations of target tasks. In comparison, DVD shows slight improvements in the seen kitchen manipulation tasks but still struggles with generalizing to new tasks, which can be attributed to the solely coarse semantic alignment between robot trajectories and human videos. We also observe a decline in performance with R3M, primarily ascribed to the pre-trained model's inability to align effectively with robot skills. Our method, H2RIL can generalize zero-shot to new tasks in the composable setting and achieve successful task completion after fine-tuning in the unseen setting. This demonstrates that our method is capable of extracting task-relevant skill embedding and encapsulating fine-grained skill memories into cross-domain human videos.

C. Qualitative Analysis

The H2RIL discovers the task-relevant skill embedding space from on-robot trajectories while maintaining the information about object-agent interactions and learns a robust cross-domain skill mapping. If the skill encoder is efficient, sub-trajectories indicating the same subtasks will be close together in the latent space. To analyze the skill representation and video embedding acquired through H2RIL, we sample 6 paired on-robot trajectories and human videos and



(a) (b)

Fig. 4. **Cross-domain skill embedding.** (a) The t-SNE visualization of the alignment of the skill embeddings among cross-domain on-robot demonstrations and human videos when executing the identical subtasks. Different shapes represent different domains: *point*=on-robot skill embeddings, *cross*=human video embeddings. Different colors indicate features from different subtasks. (b) The paired cross-domain data after the skill alignment is sampled from $\mathcal{D}^r_{demo}$ and $\mathcal{D}^h_{demo}$ for each subtask.

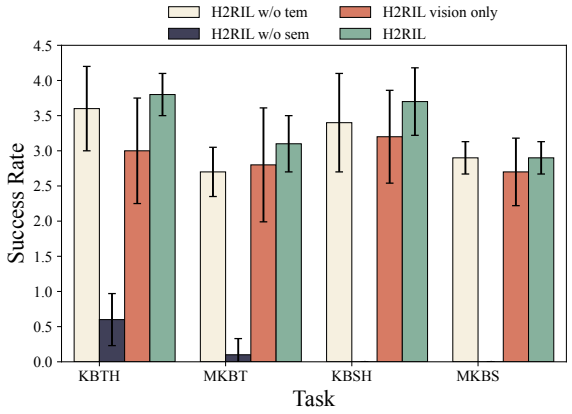


Fig. 5. Ablations on the impact of semantic pre-alignment, temporal sequence contrastive learning objective, and the robot data used for skill alignment.

subsequently visualize the cross-domain skill embeddings using t-SNE as shown in Fig. 4(a). It can be observed that the skill embeddings from the same subtasks are grouped together and clearly separated from the others. Simultaneously, the video embeddings share the same compact subspace as the matched skill embeddings. Fig. 4(b) shows the paired cross-domain data after the skill alignment. The phenomenon demonstrates that the distilled skill embeddings are sensitive to task-relevant features such as object-agent interactions and the cross-domain skill mapping policy can perform both semantic and temporal alignment in paired data.

D. Ablations

In this section, we compare three variants of our method to provide insight into the effectiveness of our architecture design: (1) **H2RIL (w/o sem):** variant of our method without performing semantic pre-alignment. We instead uniformly sample over the full video; (2) **H2RIL (w/o tem):** variant without using the temporal sequence contrastive learning objective for learning cross-domain skill mapping from paired data; (3) **H2RIL (vision only):** extracts skill embeddings solely from visual observations without proprioceptive states and actions.

From the results in Fig. 5, we observe that both the semantic pre-alignment and the temporal sequence contrastive learning objective have a significant impact on generalizing to composable tasks. These results showcase that there exists a substantial gap between on-robot demonstrations and human videos. However, both semantic pre-alignment and the temporal sequence contrastive learning objective help to bridge the gap via transferring the domain-invariant task semantics and temporal structures when training the cross-domain skill mapping policy. Comparing H2RIL (vision only) to H2RIL, we see a drop in performance, which can be attributed to the fact that complete trajectories contain more task-relevant information than visual observations alone.

V. CONCLUSIONS AND LIMITATIONS

We have presented H2RIL, a hierarchical human-to-robot imitation learning method that extracts interaction-aware skill embedding from task-agnostic on-robot play data and learns a robust cross-domain skill mapping policy via the temporal sequence contrastive learning objective to both semantically and temporally align the on-robot demonstrations and human videos in a different environment. Our method is shown to significantly outperform the prior skill-based imitation learning methods in cross-domain scenarios. It brings forth a cost-efficient way of specifying new long-horizon tasks using solely a non-expert human video.

There are multiple limitations and directions for future work. First, our H2RIL executes the sequence of skills specified by human video clips in an open-loop fashion. Future research could focus on developing a reinforcement learning method to interactively choose the video clip as the current skill in terms of the current state. In addition, the quality of frozen skill representations significantly impacts subsequent alignment operations. Future works may resort to large models for better skill representations. Lastly, Another important avenue for future work is to investigate the practicality of H2RIL for solving long-horizon tasks in real-world scenarios.

REFERENCES

[1] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "Bc-z: Zero-shot task generalization with robotic imitation learning," in *Conference on Robot Learning*. PMLR, 2022, pp. 991–1002.

[2] A. Yu and R. Mooney, "Using both demonstrations and language instructions to efficiently learn robotic tasks," in *The Eleventh International Conference on Learning Representations*, 2022.

[3] O. Mees, L. Hermann, and W. Burgard, "What matters in language conditioned robotic imitation learning over unstructured data," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 205–11 212, 2022.

[4] K. Hakhamaneshi, R. Zhao, A. Zhan, P. Abbeel, and M. Laskin, "Hierarchical few-shot imitation with skill transition models," *arXiv preprint arXiv:2107.08981*, 2021.

[5] S. Nasiriany, T. Gao, A. Mandlekar, and Y. Zhu, "Learning and retrieval from prior data for skill-based imitation learning," *arXiv preprint arXiv:2210.11435*, 2022.

[6] S. Bahl, A. Gupta, and D. Pathak, "Human-to-robot imitation in the wild," *arXiv preprint arXiv:2207.09450*, 2022.

[7] C. Wang, L. Fan, J. Sun, R. Zhang, L. Fei-Fei, D. Xu, Y. Zhu, and A. Anandkumar, "Mimicplay: Long-horizon imitation learning by watching human play," *arXiv preprint arXiv:2302.12422*, 2023.

[8] K. Pertsch, R. Desai, V. Kumar, F. Meier, J. J. Lim, D. Batra, and A. Rai, "Cross-domain transfer via semantic skill imitation," in *Conference on Robot Learning*. PMLR, 2023, pp. 690–700.

[9] O. Kroemer, S. Niekum, and G. Konidaris, "A review of robot learning for manipulation: Challenges, representations, and algorithms," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 1395–1476, 2021.

[10] P. Englert and M. Toussaint, "Learning manipulation skills from a single demonstration," *The International Journal of Robotics Research*, vol. 37, no. 1, pp. 137–154, 2018.

[11] C. Finn, T. Yu, T. Zhang, P. Abbeel, and S. Levine, "One-shot visual imitation learning via meta-learning," in *Conference on robot learning*. PMLR, 2017, pp. 357–368.

[12] S. James, M. Bloesch, and A. J. Davison, "Task-embedded control networks for few-shot imitation learning," in *Conference on robot learning*. PMLR, 2018, pp. 783–795.

[13] T. Yu, P. Abbeel, S. Levine, and C. Finn, "One-shot hierarchical imitation learning of compound visuomotor tasks," *arXiv preprint arXiv:1810.11043*, 2018.

[14] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.

[15] D. A. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network," *Advances in neural information processing systems*, vol. 1, 1988.

[16] P. Florence, C. Lynch, A. Zeng, O. A. Ramirez, A. Wahid, L. Downs, A. Wong, J. Lee, I. Mordatch, and J. Tompson, "Implicit behavioral cloning," in *Conference on Robot Learning*. PMLR, 2022, pp. 158–168.

[17] J. Ho and S. Ermon, "Generative adversarial imitation learning," *Advances in neural information processing systems*, vol. 29, 2016.

[18] V. Petrík, M. Tapaswi, I. Laptev, and J. Sivic, "Learning object manipulation skills via approximate state estimation from real videos," in *Conference on Robot Learning*. PMLR, 2021, pp. 296–312.

[19] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3m: A universal visual representation for robot manipulation," *arXiv preprint arXiv:2203.12601*, 2022.

[20] R. Strudel, A. Pashevich, I. Kalevatykh, I. Laptev, J. Sivic, and C. Schmid, "Learning to combine primitive skills: A step towards versatile robotic manipulation §," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 4637–4643.

[21] T. Shankar and A. Gupta, "Learning robot skills with temporal variational inference," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8624–8633.

[22] T. Kipf, Y. Li, H. Dai, V. Zambaldi, A. Sanchez-Gonzalez, E. Grefenstette, P. Kohli, and P. Battaglia, "Compile: Compositional imitation learning and execution," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3418–3428.

[23] D. Tanneberg, K. Ploeger, E. Rueckert, and J. Peters, "Skid raw: Skill discovery from raw trajectories," *IEEE robotics and automation letters*, vol. 6, no. 3, pp. 4696–4703, 2021.

[24] Y. Zhu, P. Stone, and Y. Zhu, "Bottom-up skill discovery from unsegmented demonstrations for long-horizon robot manipulation," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4126–4133, 2022.

[25] K. Shiarlis, M. Wulfmeier, S. Salter, S. Whiteson, and I. Posner, "Taco: Learning task decomposition via temporal alignment for control," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4654–4663.

[26] K. Pertsch, Y. Lee, and J. Lim, "Accelerating reinforcement learning with learned skill priors," in *Conference on robot learning*. PMLR, 2021, pp. 188–204.

[27] C. Lynch, M. Khansari, T. Xiao, V. Kumar, J. Tompson, S. Levine, and P. Sermanet, "Learning latent plans from play," in *Conference on robot learning*. PMLR, 2020, pp. 1113–1132.

[28] A. Ajay, A. Kumar, P. Agrawal, S. Levine, and O. Nachum, "Opal: Offline primitive discovery for accelerating offline reinforcement learning," *arXiv preprint arXiv:2010.13611*, 2020.

[29] A. Nguyen, D. Kanoulas, L. Muratore, D. G. Caldwell, and N. G. Tsagarakis, "Translating videos to commands for robotic manipulation with deep recurrent neural networks," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 3782–3788.

[30] J. Rothfuss, F. Ferreira, E. E. Aksoy, Y. Zhou, and T. Asfour, "Deep episodic memory: Encoding, recalling, and predicting episodic experiences for robot action execution," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4007–4014, 2018.

[31] Y. Liu, A. Gupta, P. Abbeel, and S. Levine, "Imitation from observation: Learning to imitate behaviors from raw video via context translation," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1118–1125.

[32] M. Xu, Z. Xu, C. Chi, M. Veloso, and S. Song, "Xskill: Cross embodiment skill discovery," *arXiv preprint arXiv:2307.09955*, 2023.

[33] T. Franzmeyer, P. Torr, and J. F. Henriques, "Learn what matters: cross-domain imitation learning with task-relevant embeddings," *Advances in Neural Information Processing Systems*, vol. 35, pp. 26 283–26 294, 2022.

[34] A. S. Chen, S. Nair, and C. Finn, "Learning generalizable robotic reward functions from" in-the-wild" human videos," *arXiv preprint arXiv:2103.16817*, 2021.

[35] E. Chane-Sane, C. Schmid, and I. Laptev, "Learning video-conditioned policies for unseen manipulation tasks," *arXiv preprint arXiv:2305.06289*, 2023.

[36] M. Monfort, B. Pan, K. Ramakrishnan, A. Andonian, B. A. McNamara, A. Lascelles, Q. Fan, D. Gutfreund, R. S. Feris, and A. Oliva, "Multi-moments in time: Learning and interpreting models for multi-action video understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9434–9445, 2021.

[37] Z. J. Cui, Y. Wang, N. M. M. Shafiullah, and L. Pinto, "From play to policy: Conditional behavior generation from uncurated robot data," in *The Eleventh International Conference on Learning Representations*, 2022.

[38] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh, "Recycle-gan: Unsupervised video retargeting," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 119–135.

[39] M. Douze, J. Revaud, J. Verbeek, H. Jégou, and C. Schmid, "Circulant temporal encoding for video retrieval and temporal alignment," *International Journal of Computer Vision*, vol. 119, pp. 291–306, 2016.

[40] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari, "Actor and observer: Joint modeling of first and third-person videos," in *proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7396–7404.

[41] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfreund, C. Vondrick, *et al.*, "Moments in time dataset: one million videos for event understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 2, pp. 502–508, 2019.

[42] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, and G. Brain, "Time-contrastive networks: Self-supervised learning from video," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 1134–1141.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[44] A. Gupta, V. Kumar, C. Lynch, S. Levine, and K. Hausman, "Relay policy learning: Solving long-horizon tasks via imitation and reinforce-

ment learning," in *Conference on Robot Learning*. PMLR, 2020, pp. 1025–1037.
[45] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.
[46] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6836–6846.