Unifying What and How: Distilling a Pre-trained Unified Skill Representation for Efficient Adaptation

Jusuk Lee[1] Daesol Cho[2] Jonghun Shin[1] Jonghae Park[1] Taekbeom Lee[1] H. Jin Kim[1]

[1]Seoul National University [2]Georgia Institute of Technology

Abstract: Skill abstraction, the process of learning reusable and temporally extended behaviors, has emerged as a key focus in robot learning for its potential to improve sample efficiency and generalization. However, existing methods exhibit complementary strengths and weaknesses, typically modeling either high-level semantic intent (*'what to do'*) while sacrificing motion fidelity, or fine-grained motion dynamics (*'how to do it'*) while lacking semantic context. To address these limitations, we introduce Unified Skill Representation (USR) that unifies both the semantic intent and motion dynamics into a single skill representation. USR employs a cross-modal VQ-VAE to learn a semantically grounded and dynamically aware skill codebook. Furthermore, we propose a decoupled training framework that reconciles large-scale skill pre-training with practical deployment. Our approach builds transferable skills from vast, diverse datasets and then trains a lightweight policy on small, in-domain data. Extensive experiments on the LIBERO benchmark demonstrate that USR not only achieves near expert-level performance on the training distribution but also substantially outperforms prior methods in few-shot transfer to unseen tasks. Our results highlight the importance of both unifying skill representations and decoupling the training pipeline, offering a step toward more generalizable and practical robotic agents.

Keywords: Robot learning, Skill abstraction, Imitation learning

1 Introduction

A longstanding goal in robot learning is to develop a general-purpose agent capable of rapidly adapting to novel tasks in a zero-shot or few-shot manner. Recent advances in vision-language foundation models have demonstrated remarkable success by leveraging powerful, pre-trained representations for diverse downstream tasks [1, 2, 3, 4]. In robot learning, imitation learning similarly employs supervised training on demonstration data, analogous to the training of vision-language foundation models. However, this approach typically yields specialized agents that fail to generalize beyond their training distribution and lack the ability to transfer their knowledge to new tasks. This limitation underscores the necessity of developing *reusable, pretrained representations* that can transfer to new tasks, environments, and embodiments.

Skill abstraction has emerged as a prominent approach toward achieving compact and reusable representations of temporally extended behaviors. Existing skill abstraction methods can broadly be categorized into two types, each exhibiting complementary strengths and limitations. Low-level skill abstraction methods [5, 6, 7, 8] capture fine-grained motion dynamics directly from action trajectories. Yet, they lack an explicit understanding of the semantic context, necessitating substantial task-specific data to bridge semantic information and low-level skills during downstream policy training. Conversely, high-level skill abstraction methods [9, 10, 11, 12] effectively encode semantic intent but often sacrifice motion fidelity. This leads to an oversimplified skill representation that fails to differentiate dynamically distinct behaviors, severely limiting the diversity of the learned

9th Conference on Robot Learning (CoRL 2025), Seoul, Korea.

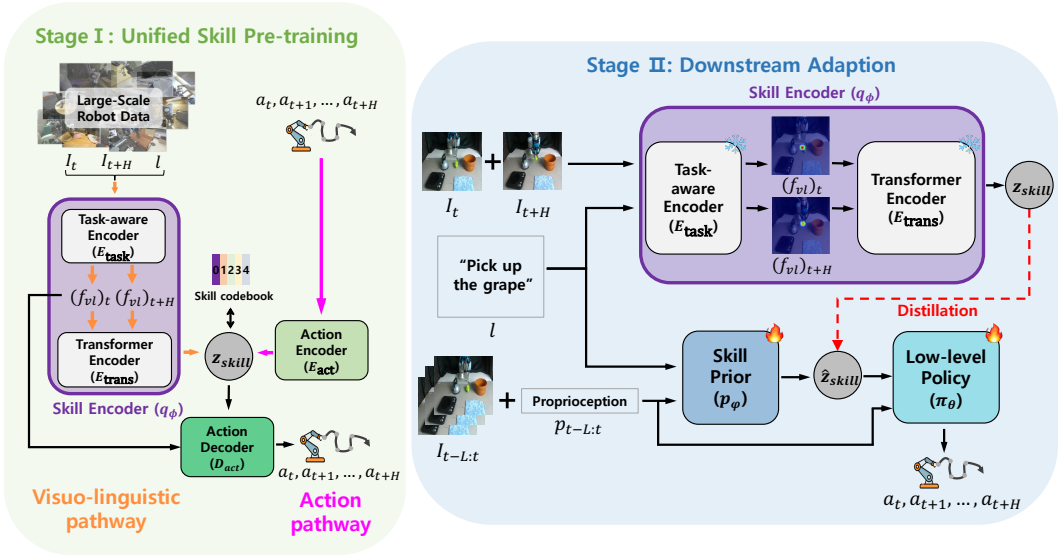


Figure 1: **Overview of Unified Skill Representation (USR) framework.** (Left) **Stage I: Unified Skill Pre-training.** Learning a generalizable, unified skill representation from large-scale data. (Right) **Stage II: Downstream Adaptation.** Efficiently distilling the unified skill representation into a lightweight skill prior and policy for the target task. The heatmaps inside the skill encoder visualize attention maps, showing that the model focuses on task-relevant objects.

skills. To overcome these limitations, it is essential to develop a unified skill representation that simultaneously integrates high-level semantic intent and low-level motion dynamics.

Furthermore, capturing rich semantic intent typically requires large-scale models trained on extensive, heterogeneous datasets [13, 14]. However, deploying such models in real-world robotic applications is computationally prohibitive due to strict latency and resource constraints. Consequently, it is necessary to distill these rich semantics into a lightweight model to enable fast, resource-efficient inference. This challenge has remained largely unexplored in existing skill-conditioned robot learning literature [9, 10, 11, 12, 15, 16]. Therefore, our goal is to establish a framework that learns a unified skill representation, capturing both semantic and motion dynamics, while remaining lightweight enough for practical robotic deployment.

In this paper, we propose **Unified Skill Representation (USR)**, a novel framework that operates via a decoupled, two-stage training pipeline. In the **first stage**, USR pre-trains a unified skill representation from large-scale, diverse datasets, capturing both semantic and motion dynamics. In the **second stage**, this rich representation is distilled efficiently into a lightweight policy and skill prior, optimized specifically for the target tasks. The core of our framework is cross-modal VQ-VAE [17, 18], which integrates two complementary pathways to learn a unified skill codebook. The visuo-linguistic pathway captures high-level semantic intent ('what') by extracting generalizable visual features from a pre-trained CLIP model [19], and further fusing them with task instructions via a cross-attention mechanism. Simultaneously, the action pathway grounds these high-level semantics into low-level physical motion ('how'), using action sequences as supervision. This forces the model to learn a compact representation that is robust to visual distractors irrelevant to the action. These combined pathways make our skill representations highly transferable. This unification of 'what' and 'how' yields robust skills that are both task-aware and temporally coherent. The resulting representation is compactly distilled into a lightweight skill prior, facilitating efficient downstream adaptation.

We validate USR on the LIBERO benchmark, evaluating both in-distribution performance and few-shot adaptation. Our results demonstrate that USR substantially outperforms previous skill-based methods in few-shot adaptation, particularly excelling in extreme low-data regimes (5-10 demon-

strations). Moreover, USR achieves near expert-level performance on tasks within the training distribution. To summarize, our main contributions are:

1. We propose USR, a framework that learns a unified skill representation by integrating high-level semantic intent and low-level motion dynamics into a single skill codebook.
2. We introduce a decoupled training pipeline that reconciles large-scale pre-training with deployment efficiency, enabling efficient adaptation of a lightweight policy to new tasks in a few-shot manner.
3. We demonstrate superior performance on the LIBERO benchmark, with competitive results on the training distribution and significant outperformance in few-shot adaptation against existing skill-based approaches.

2 Related Works

2.1 Representation Pre-training on Large-Scale Data

Recent works have explored pre-training on large-scale video datasets to improve downstream policy learning. One prominent approach is learning visual representations through self-supervised objectives, such as time-contrastive learning [20, 21] or video-language alignment [20, 22]. However, these methods typically overlook explicit dynamic labels (i.e., robot actions, proprioceptions), resulting in representations that are insufficiently grounded for manipulation. Although recent work [23] contrasts visual features with proprioceptive and action data, it still relies on time-contrastive objectives, inherently lacking meaningful temporal abstraction. Our work overcomes both limitations by learning a skill representation that is inherently not only temporally coherent but also explicitly grounded in motion dynamics through action-based supervision.

Another line of work involves video foundation models [24, 25, 26, 27, 28, 29], which train generative networks to predict future frames. These generated visuals guide policies either directly as image goals [24] or indirectly through intermediate representations like object flow [28]. Despite capturing implicit dynamics, their high computational demands limit real-time deployment and closed-loop control. Our method circumvents this limitation by learning compact skill representations instead of generating high-dimensional pixels.

2.2 Skill Abstraction from Offline Data

Low-level skill abstractions Low-level methods [5, 6, 7, 8] focus exclusively on action trajectories, quantizing action sequences into discrete motion primitives. As these skills derive solely from the agent's behavior, they inherently lack semantic context, such as identifying relevant objects to interact with. The absence of rich scene understanding significantly limits downstream task performance [30, 20, 22]. In contrast, our framework explicitly enriches motion primitives with semantic context.

High-level skill abstractions High-level methods embed rich semantic context into discrete skills by leveraging visual and language data. One subset of approaches [9, 11, 10, 31, 32] learns interpretable skills from limited, in-domain datasets. However, these methods produce representations overly reliant on specific visual sequences, as they predict skills conditioned on past observation histories. Consequently, the learned skills become sensitive to tiny visual details, limiting transferability to novel scenes and tasks. Recent methods [15, 16] employ vision foundation models [2, 20] to enhance generalization. However, by defining skills based on visual change without explicit grounding in actions or language, these representations become susceptible to visual distractors and fail to capture meaningful temporal abstractions. Our proposed framework differs by jointly leveraging task-aware visual features and action supervision, yielding skill representations robust to visual distractors and explicitly grounded in meaningful temporal abstractions. Moreover, our approach conditions skills on current and future visual contexts, rather than past observations alone. This de-

sign choice reduces dependence on visual sequences specific to demonstrations, thereby enhancing transferability.

3 Problem Setting

Our goal is to learn a policy that enables a robot to perform a wide range of tasks specified by natural language instructions. We formulate this problem as a conditional Markov Decision Process (MDP), denoted by the tuple $\big(\mathcal{O}, \mathcal{A}, P, R, \mathcal{L}, \gamma\big)$ [33]. It consists of an observation space $\mathcal{O}$, an action space $\mathcal{A}$, transition dynamics P, a reward function R, a space of natural language instructions $\mathcal{L}$, and a discount factor γ. At each timestep t, the agent receives an observation O_t comprising a front camera image I_t, a gripper camera image I_t^{gripper} (if available), and its proprioceptive state p_t. We assume access to a dataset of N language-conditioned trajectories $\mathcal{D} = \big\{\big(l^{(i)},\, s_1^{(i)}, a_1^{(i)}, \ldots, s_{T_i}^{(i)}, a_{T_i}^{(i)}\big)\big\}_{i=1}^{N}$. Our objective is to learn a language-conditioned policy $\Pi(a_{t:t+H} \mid O_{t-L:t}, l)$. Here, L denotes the observation history length and H is the action prediction horizon.

4 Method

Our method employs a modular, two-stage pipeline that decouples large-scale skill pre-training from small-scale downstream adaptation, as illustrated in Figure 1. Section 4.1 presents the probabilistic framework based on the Evidence Lower Bound (ELBO), enabling our two-stage learning strategy. Section 4.2 describes the skill pre-training phase, where a unified skill representation is learned from diverse large-scale offline datasets. Section 4.3 outlines the downstream adaptation phase, where a lightweight skill prior and policy are trained efficiently for target tasks. This modular approach enables USR to distill knowledge from large datasets into a compact and sample-efficient policy suitable for real-world deployment.

4.1 Decoupling Skill Pre-Training and Downstream Adaptation

Training a monolithic, large-scale policy Π on vast datasets produces powerful but impractical models due to their computational costs, making them unsuitable for resource-constrained robotic deployment. Therefore, our approach focuses on creating a lightweight policy π_θ that preserves the benefits of large-scale pre-training.

We achieve this by decomposing the monolithic policy Π. Our goal is to optimize its log-likelihood, $\log \Pi(a_{t:t+H} \mid O_{t-L:t}, l)$. We introduce a latent skill variable z to reformulate this objective in terms of a skill-conditioned policy $\pi_\theta(a_{t:t+H} \mid z_t, O_{t-L:t})$ and a skill prior $p_\psi(z_t \mid O_{t-L:t}, l)$. Following the variational inference framework, we then introduce a skill encoder $q_\phi(z_t \mid I_t, I_{t+H}, l)$. This allows us to optimize the evidence lower bound of the log-likelihood, formulated as (see Appendix A for the detailed derivation):

$$\begin{aligned}\mathcal{L}(\theta, \psi, \phi) = \; & -\mathbb{E}_{z_t \sim q_\phi}[\log \pi_\theta(a_{t:t+H} \mid z_t, O_{t-L:t})] \\ & + D_{\mathrm{KL}}(q_\phi(z_t \mid I_t, I_{t+H}, l) \parallel p_\psi(z_t \mid O_{t-L:t}, l)).\end{aligned} \tag{1}$$

To enable the lightweight prior and policy to utilize rich knowledge from large-scale data, we introduce a two-stage training framework. **Stage I** pre-trains the skill encoder q_ϕ on large-scale data. This process builds a general-purpose skill representation. In **Stage II**, we freeze the pre-trained encoder. Its general knowledge is then distilled into the lightweight skill prior p_ψ and policy π_θ as they are trained efficiently on the target dataset. This decoupled, distillation-based approach ensures our policy is compact yet informed by large-scale data.

4.2 Stage I: Unified Skill Pre-training

The goal of Stage I is to pre-train a unified skill representation from large-scale, diverse offline data, jointly capturing high-level semantic intent ("what to do") and low-level motion dynamics ("how to do it"). To achieve this, we employ a cross-modal VQ-VAE with a discrete skill codebook containing reusable skill primitives. This codebook is trained using information from two complementary

pathways: a visuo-linguistic pathway for semantic understanding and an action pathway for motion dynamics. We alternately optimize these two pathways to ensure a balanced and robust skill representation.

The visuo-linguistic pathway provides semantic context through our skill encoder q_ϕ, defined as $q_\phi(\cdot) = E_{\text{trans}}(E_{\text{task}}(\cdot))$. First, the task-aware encoder E_{task} generates task-aware visual features $((f_{vl})_t, (f_{vl})_{t+H})$ from the current image I_t, the future image I_{t+H}, and language instruction l. Inspired by recent works [34, 35], we fuse patch-level visual tokens from a pre-trained CLIP ViT with the language instruction via a temperature-scaled cross-attention mechanism:

$$f_{\text{vl}} = \text{softmax}\Big(\frac{\hat{f}_l\,\hat{f}_v^\top}{\tau}\Big)\,(\hat{f}_v + \text{PE}), \tag{2}$$

where $\hat{f}_v$ and $\hat{f}_l$ are the normalized visual and language features, τ is a learnable temperature, and PE denotes positional embeddings. This allows the model to selectively attend to instruction-relevant visual regions. Then, the transformer encoder E_{trans} captures temporal relationships, yielding $f_{\text{enc},vl} = E_{\text{trans}}((f_{vl})_t, (f_{vl})_{t+H})$. In parallel, the action pathway encoder E_{act} directly processes action trajectories, producing dynamic features $f_{\text{enc, act}} = E_{\text{act}}(a_{t:t+H})$. Each pathway extracts only task-relevant information, which is then quantized into a discrete codebook. The continuous feature vector f_{enc} from either pathway (i.e., $f_{\text{enc},vl}$ or $f_{\text{enc, act}}$) is then quantized to the nearest vector c_k in the codebook, $\mathcal{C} = \{c_k\}_{k=1}^K$, to obtain the discrete skill z_t:

$$z_t = \underset{k\in\{1,\dots,K\}}{\arg\min} \big\|f_{\text{enc}} - c_k\big\|_2^2. \tag{3}$$

The framework is optimized via a single supervisory signal, which is action reconstruction. An action decoder D_{act} reconstructs action sequences from the skill z_t and current task-aware visual feature $(f_{vl})_t$: $\hat{a}_{t:t+H} = D_{\text{act}}(z_t, (f_{vl})_t)$. Unlike prior work that reconstructs high-dimensional images [36, 37], our framework uses the low-dimensional action sequence as a supervisory signal. This compels the model to discard visual information irrelevant to the action, thus enhancing its robustness to distractors. To ensure training stability, we mitigate gradient collapse using NSVQ [38] and prevent codebook collapse using a codebook replacement technique [38, 39, 40]. The pre-training objective is therefore a total reconstruction loss over both pathways:

$$\mathcal{L}_{\text{pretrain}}(\phi) \;=\; \underbrace{\|a_{t:t+H} - \hat{a}_{vl}\|_2^2}_{\text{visuo-linguistic pathway}} \;+\; \underbrace{\|a_{t:t+H} - \hat{a}_{act}\|_2^2}_{\text{action pathway}}, \tag{4}$$

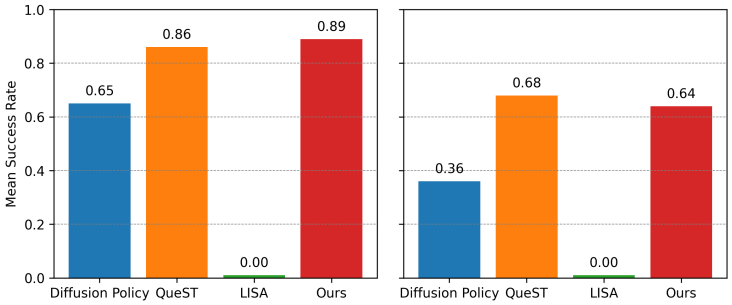
where $a_{t:t+H}$ is the ground-truth action sequence, and $\hat{a}_{vl}$ and $\hat{a}_{act}$ are the reconstructions from the two pathways.

4.3 Stage II: Downstream Adaptation

In Stage II, we distill the rich knowledge encoded in the pre-trained skill encoder q_ϕ into two lightweight components: a skill prior p_ψ and a skill-conditioned policy π_θ, both designed for efficient deployment. Unlike Stage I, which leverages future image to learn predictive representations, Stage II uses only past observations, as future inputs are unavailable during execution. Both p_ψ and π_θ are trained on a smaller, in-domain dataset and implemented as small Transformer models that process multimodal input stream consisting of camera features and proprioception. The policy head follows BAKU [41] and predicts a continuous action distribution modeled as an isotropic Gaussian.

$$\begin{aligned}\mathcal{L}_{\text{downstream}}(\theta,\psi;\phi) = &\underbrace{\mathbb{E}_{z_t^q\sim q_\phi}\big[-\log p_\psi(z_t^q \mid O_{t-L:t}, l)\big]}_{\text{Prior Distillation Loss}} \\ &+\; \alpha\,\underbrace{\mathbb{E}_{z_t\sim p_\psi}\Big[-\log \pi_\theta(a_{t:t+H} \mid \text{sg}(z_t), O_{t-L:t})\Big]}_{\text{Policy Behavior Cloning Loss}}.\end{aligned} \tag{5}$$

The first term trains the prior p_ψ to approximate the output of the frozen skill encoder, where $z_t^q \sim q_\phi(I_t, I_{t+H}, l)$ serves as a pseudo-label. The second term trains the policy π_θ to imitate expert



(a) Multitask on LIBERO-90 (b) Multitask on LIBERO-LONG

Figure 2: **In-distribution performance.** Average success rate on the (a) LIBERO-90 and (b) LIBERO-LONG benchmarks. Results are averaged over 5 runs.

actions, conditioned on sampled skills. For stable training, the stop-gradient operator sg isolates the prior from the policy's learning signal, ensuring it is guided solely by the stable output of the pre-trained skill encoder.

5 Experiments

We evaluate USR along two key axes: **(1) In-Distribution Performance** on the training data, and **(2) Few-Shot Adaptation Performance** to new tasks with limited demonstrations.

5.1 Experimental Setup

Benchmark. We evaluate our method on the LIBERO benchmark. LIBERO-90 and LIBERO-LONG are used for skill pre-training. For few-shot adaptation, we evaluate on LIBERO-OBJECT, LIBERO-SPATIAL, and LIBERO-GOAL, each introducing novel objects, spatial configurations, or goals.

Baselines. We compare USR with three representative baselines and a key variant of our own model. We evaluate against Diffusion Policy [42], a non-hierarchical imitation learning method that directly maps observations to action sequences; LISA [9], a high-level skill abstraction method that jointly learns semantic skills and a policy; and QueST [6], a low-level skill abstraction method that learns motion primitives from action trajectories. Additionally, we introduce USR (No Skill), a variant where the policy is conditioned directly on the language instruction instead of our learned skills. This allows us to isolate the contribution of the unified skill representation itself.

Evaluation Protocol. For all experiments, we evaluate each method by running it 5 times with different random seeds. We report the average success rate across these runs.

5.2 In-Distribution Performance

We first evaluate the performance on the pre-training distributions (LIBERO-90 and LIBERO-LONG) to assess the fundamental quality of our learned skills. As shown in Figure 2, USR achieves the highest success rate on LIBERO-90, outperforming all baselines. On the more complex LIBERO-LONG suite, USR achieves strong performance, significantly surpassing both Diffusion Policy and LISA.

Due to its end-to-end framework with joint training, LISA exhibits significant instability when applied to large and diverse datasets, resulting in complete failure (0% success rate) and issues such as codebook collapse. This result provides strong empirical evidence for our core motivation: decou-

Table 1: **Few-shot adaptation performance**. We report the average success rate over 5 evaluation runs on the three LIBERO benchmarks, training the downstream policy and prior on a varying number of demonstrations (50, 10, and 5). **Bold** indicates the best performance in each column.

Method	LIBERO-OBJECT (# demos)			LIBERO-SPATIAL (# demos)			LIBERO-GOAL (# demos)		
	50	10	5	50	10	5	50	10	5
Diffusion Policy [42]	0.92	0.76	**0.54**	0.78	0.36	0.34	0.72	0.44	0.24
QueST [6]	0.90	0.60	0.38	0.78	0.50	0.10	0.80	0.54	0.12
LISA [9]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
USR (No Skill)	**0.94**	0.50	0.16	0.74	0.34	0.14	0.84	0.28	0.18
USR (Ours)	0.92	**0.78**	0.52	**0.84**	**0.64**	**0.52**	**0.90**	**0.62**	**0.54**

pling skill pre-training from downstream adaptation is essential for robust training with large-scale data.

5.3 Few-Shot Adaptation Performance

We now test the central claim of our work that USR's unified skills enable highly efficient adaptation. In a few-shot adaptation setting, we freeze the pre-trained skill encoder q_ϕ and train only the lightweight prior p_ψ and policy π_θ on a limited number of demonstrations (5, 10, or 50). For a fair comparison, the other skill-based baselines (QueST and LISA) follow the same protocol. Their pre-trained skill modules are frozen, and only the downstream components are trained on the target data. The non-hierarchical Diffusion Policy, in contrast, is trained from scratch on the few-shot demonstrations.

As shown in Table 1, USR demonstrates markedly superior performance over all baselines across the three benchmarks. This advantage is most pronounced in the extreme low-data regimes (5 and 10 demonstrations), which underscores the high transferability of our unified skills. Diffusion Policy performs competently on the simpler tasks (LIBERO-OBJECT). However, its performance degrades on complex tasks that require both high-level semantic understanding and coherent low-level motion dynamics, revealing the limitations of an unstructured, end-to-end approach. QueST exhibits a sharp performance decline as the number of demonstrations decreases. While its motion primitives are effective with sufficient adaptation data, they lack the semantic grounding necessary to infer the correct high-level behavior in a novel visual context from few examples. LISA fails on all transfer tasks. Its retrospective approach of defining skills from past observations causes it to overfit to the training domain's specific visual details. Furthermore, its joint training paradigm is ill-suited for the large-scale pre-training necessary for generalization. Finally, our variant, USR (No Skill) provides direct evidence that the unified skill representation itself is the primary driver of this high sample efficiency.

5.4 Latent Skill Analysis

We conduct a qualitative analysis to understand the properties of the learned latent skills. As shown in Figure 3, the learned skills are fundamentally task-centric and robust to visual distractors. For instance, a skill codebook (e.g., 'Place') consistently represents trajectories with the same underlying semantic intent and motion dynamics, even amidst visual variations such as different backgrounds or object layouts. Furthermore, the t-SNE visualization in Figure 4 demonstrates that the learned skill space is well-structured, where skills with similar semantics and motion characteristics form distinct clusters.

6 Conclusion

In this work, we introduce USR that learns a unified skill representation to reconcile the need for large-scale pre-training with the demands of efficient, real-world policy deployment. Our approach

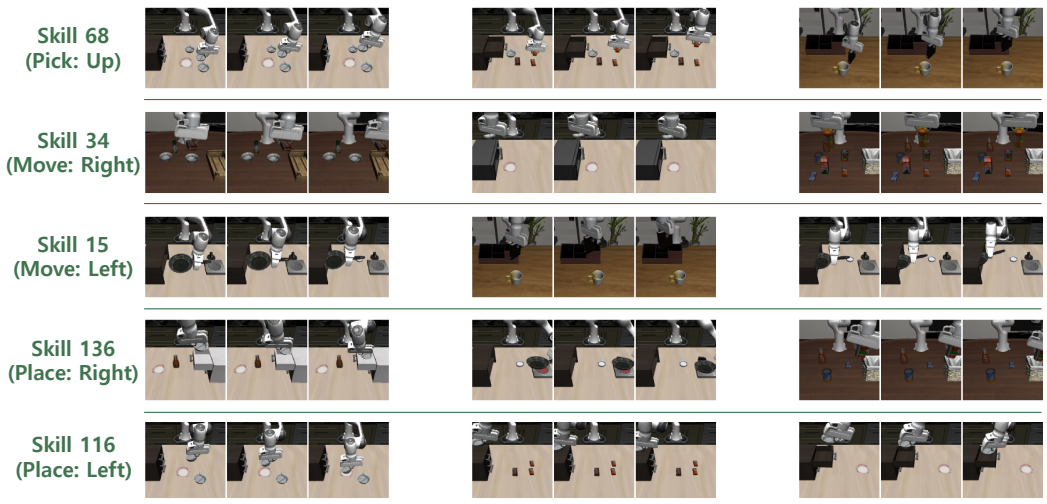


Figure 3: **Visualizations of skills.** Each row shows different trajectories that map to the same skill codebook, demonstrating that our skills are task-centric and robust to visual distractors.

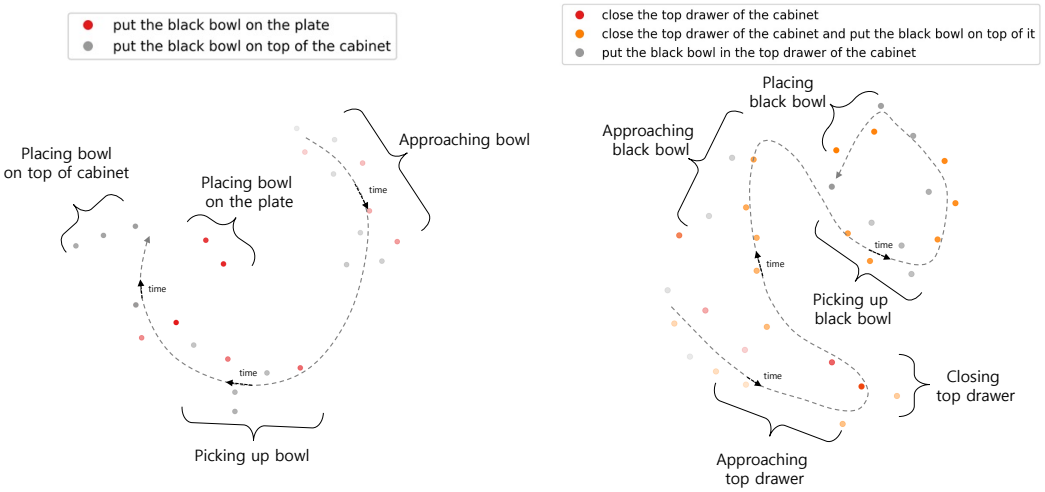


Figure 4: **t-SNE visualization of unified skill embedding.** Our learned skills form distinct clusters based on semantic-motion similarity, demonstrating their reusability across tasks. Each point is a skill embedding, with opacity denoting time (more transparent is earlier).

unifies high-level semantic intent ('what to do') and low-level motion dynamics ('how to do it') via a cross-modal VQ-VAE with a shared skill codebook. Our decoupled, two-stage pipeline enables skill pre-training on large datasets by separating it from the few-shot adaptation phase of a compact, downstream policy. Experiments on the LIBERO benchmark demonstrate that our method achieves strong in-distribution performance and significantly outperforms baselines in few-shot adaptation scenarios. Despite these promising results, a key limitation lies in our framework's reliance on action-labeled trajectories. A crucial direction for future work is to extend our method to learn from the vast and diverse corpora of action-free videos, such as human demonstrations available online. Addressing this limitation would unlock the potential of web-scale data for building generalist robotic agents.

A Derivation of the Evidence Lower Bound

This appendix provides a comprehensive derivation of the Evidence Lower Bound (ELBO) presented in Eq. 1. Our framework decomposes this objective into three distinct modules: a skill encoder q_ϕ, a skill prior p_ψ, and a skill-conditioned policy π_θ. For notational simplicity, we denote the $O_{t-L:t}$ as O and the action sequence $a_{t:t+H}$ as a.

The derivation begins with the log-likelihood policy $\Pi(a \mid O, l)$:

$$\log \Pi(a \mid O, l)$$

We introduce the latent skill variable z by marginalizing over its distribution. This allows us to conceptually decompose the monolithic policy Π into a conditional policy π_θ and a skill prior p.

$$= \log \sum_z \Pi(a, z \mid O, l) = \log \sum_z \pi_\theta(a \mid z, O, l) p(z \mid O, l)$$

Next, we introduce a variational distribution $q_\phi(z \mid I_t, I_{t+H}, l)$ as an approximation to the true posterior $p(z \mid a, O, l)$. Multiplying and dividing by q_ϕ within the summation yields:

$$= \log \sum_z q_\phi(z \mid I_t, I_{t+H}, l) \frac{\pi_\theta(a \mid z, O, l) p(z \mid O, l)}{q_\phi(z \mid I_t, I_{t+H}, l)}$$

This is equivalent to the logarithm of an expectation with respect to q_ϕ:

$$= \log \mathbb{E}_{z \sim q_\phi} \left[\frac{\pi_\theta(a \mid z, O, l) p(z \mid O, l)}{q_\phi(z \mid I_t, I_{t+H}, l)} \right]$$

By applying Jensen's inequality, we establish a lower bound on the log-likelihood:

$$\geq \mathbb{E}_{z \sim q_\phi} \left[\log \frac{\pi_\theta(a \mid z, O, l) p(z \mid O, l)}{q_\phi(z \mid I_t, I_{t+H}, l)} \right]$$

Expanding the logarithm gives:

$$= \mathbb{E}_{z \sim q_\phi} \big[\log \pi_\theta(a \mid z, O, l) + \log p(z \mid O, l) - \log q_\phi(z \mid I_t, I_{t+H}, l) \big]$$

Herein, we posit a key modeling assumption: the latent skill z serves as a sufficient statistic for the language instruction l with respect to the policy's action generation. This implies the conditional independence $\pi_\theta(a \mid z, O, l) = \pi_\theta(a \mid z, O)$. Applying this assumption, we obtain:

$$= \mathbb{E}_{z \sim q_\phi} [\log \pi_\theta(a \mid z, O)] + \mathbb{E}_{z \sim q_\phi} [\log p(z \mid O, l) - \log q_\phi(z \mid I_t, I_{t+H}, l)]$$

Finally, we substitute the true prior $p(z \mid O, l)$ with a learnable skill prior $p_\psi(z \mid O, l)$. The second expectation term can then be expressed as the negative Kullback-Leibler (KL) divergence, which yields the final ELBO objective:

$$\mathcal{L}(\theta, \phi, \psi) = \mathbb{E}_{z \sim q_\phi(z \mid I_t, I_{t+H}, l)} \big[\log \pi_\theta(a \mid z, O) \big] - D_{\mathrm{KL}} \big(q_\phi(z \mid I_t, I_{t+H}, l) \big| p_\psi(z \mid O, l) \big)$$

Acknowledgments

This work was supported by Samsung Research Funding & Incubation Center of Samsung Electronics under Project Number SRFC-IT2402-17

References

[1] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[2] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[3] 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*, 30, 2017.

[4] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[5] S. Lee, Y. Wang, H. Etukuru, H. J. Kim, N. M. M. Shafiullah, and L. Pinto. Behavior generation with latent actions. *arXiv preprint arXiv:2403.03181*, 2024.

[6] A. Mete, H. Xue, A. Wilcox, Y. Chen, and A. Garg. Quest: Self-supervised skill abstractions for learning continuous control. *Advances in Neural Information Processing Systems*, 37: 4062–4089, 2024.

[7] K. Pertsch, Y. Lee, and J. Lim. Accelerating reinforcement learning with learned skill priors. In *Conference on robot learning*, pages 188–204. PMLR, 2021.

[8] R. Zheng, C.-A. Cheng, H. Daumé III, F. Huang, and A. Kolobov. Prise: Llm-style sequence compression for learning temporal action abstractions in control. *arXiv preprint arXiv:2402.10450*, 2024.

[9] D. Garg, S. Vaidyanath, K. Kim, J. Song, and S. Ermon. Lisa: Learning interpretable skill abstractions from language. *Advances in Neural Information Processing Systems*, 35:21711–21724, 2022.

[10] Z. Liang, Y. Mu, H. Ma, M. Tomizuka, M. Ding, and P. Luo. Skilldiffuser: Interpretable hierarchical planning via skill abstractions in diffusion-based task execution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16467–16476, 2024.

[11] H. Jiang, J. Wang, and Z. Lu. Discrete latent plans via semantic skill ab.

[12] Z. Ju, C. Yang, F. Sun, H. Wang, and Y. Qiao. Rethinking mutual information for language conditioned skill discovery on imitation learning. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 34, pages 301–309, 2024.

[13] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

[14] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[15] J. Zhang, M. Heo, Z. Liu, E. Biyik, J. J. Lim, Y. Liu, and R. Fakoor. Extract: Efficient policy learning by extracting transferable robot skills from offline data. *arXiv preprint arXiv:2406.17768*, 2024.

[16] W. Wan, Y. Zhu, R. Shah, and Y. Zhu. Lotus: Continual imitation learning for robot manipulation through unsupervised skill discovery. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 537–544. IEEE, 2024.

[17] S. Yoo, S. Jung, Y. Lee, D. Shim, and H. J. Kim. Mono-camera-only target chasing for a drone in a dense environment by cross-modal learning. *IEEE Robotics and Automation Letters*, 9(8): 7254–7261, 2024.

[18] A. Spurr, J. Song, S. Park, and O. Hilliges. Cross-modal deep variational hand pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 89–98, 2018.

[19] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[20] 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.

[21] T. Xiao, I. Radosavovic, T. Darrell, and J. Malik. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022.

[22] Y. J. Ma, V. Kumar, A. Zhang, O. Bastani, and D. Jayaraman. Liv: Language-image representations and rewards for robotic control. In *International Conference on Machine Learning*, pages 23301–23320. PMLR, 2023.

[23] G. Jiang, Y. Sun, T. Huang, H. Li, Y. Liang, and H. Xu. Robots pre-train robots: Manipulation-centric robotic representation from large-scale robot datasets. *arXiv preprint arXiv:2410.22325*, 2024.

[24] Y. Du, S. Yang, B. Dai, H. Dai, O. Nachum, J. Tenenbaum, D. Schuurmans, and P. Abbeel. Learning universal policies via text-guided video generation. *Advances in neural information processing systems*, 36:9156–9172, 2023.

[25] P.-C. Ko, J. Mao, Y. Du, S.-H. Sun, and J. B. Tenenbaum. Learning to act from actionless videos through dense correspondences. *arXiv preprint arXiv:2310.08576*, 2023.

[26] Y. Wen, J. Lin, Y. Zhu, J. Han, H. Xu, S. Zhao, and X. Liang. Vidman: Exploiting implicit dynamics from video diffusion model for effective robot manipulation. *Advances in Neural Information Processing Systems*, 37:41051–41075, 2024.

[27] Y. Hu, Y. Guo, P. Wang, X. Chen, Y.-J. Wang, J. Zhang, K. Sreenath, C. Lu, and J. Chen. Video prediction policy: A generalist robot policy with predictive visual representations. *arXiv preprint arXiv:2412.14803*, 2024.

[28] M. Xu, Z. Xu, Y. Xu, C. Chi, G. Wetzstein, M. Veloso, and S. Song. Flow as the cross-domain manipulation interface. *arXiv preprint arXiv:2407.15208*, 2024.

[29] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790, 2023.

[30] S. Li, Y. Gao, D. Sadigh, and S. Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025.

[31] L. X. Shi, J. J. Lim, and Y. Lee. Skill-based model-based reinforcement learning. *arXiv preprint arXiv:2207.07560*, 2022.

[32] K. Pertsch, Y. Lee, Y. Wu, and J. J. Lim. Guided reinforcement learning with learned skills. *arXiv preprint arXiv:2107.10253*, 2021.

[33] V. Myers, A. W. He, K. Fang, H. R. Walke, P. Hansen-Estruch, C.-A. Cheng, M. Jalobeanu, A. Kolobov, A. Dragan, and S. Levine. Goal representations for instruction following: A semi-supervised language interface to control. In *Conference on Robot Learning*, pages 3894–3908. PMLR, 2023.

[34] M. Lan, C. Chen, Y. Ke, X. Wang, L. Feng, and W. Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *European Conference on Computer Vision*, pages 143–160. Springer, 2024.

[35] H. Huang, F. Liu, L. Fu, T. Wu, M. Mukadam, J. Malik, K. Goldberg, and P. Abbeel. Otter: A vision-language-action model with text-aware visual feature extraction. *arXiv preprint arXiv:2503.03734*, 2025.

[36] S. Ye, J. Jang, B. Jeon, S. Joo, J. Yang, B. Peng, A. Mandlekar, R. Tan, Y.-W. Chao, B. Y. Lin, et al. Latent action pretraining from videos. *arXiv preprint arXiv:2410.11758*, 2024.

[37] H. Kim, J. Kang, H. Kang, M. Cho, S. J. Kim, and Y. Lee. Uniskill: Imitating human videos via cross-embodiment skill representations. *arXiv preprint arXiv:2505.08787*, 2025.

[38] M. H. Vali and T. Bäckström. Nsvq: Noise substitution in vector quantization for machine learning. *IEEE Access*, 10:13598–13610, 2022.

[39] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021.

[40] R. Gray. Vector quantization. *IEEE Assp Magazine*, 1(2):4–29, 1984.

[41] S. Haldar, Z. Peng, and L. Pinto. Baku: An efficient transformer for multi-task policy learning. *Advances in Neural Information Processing Systems*, 37:141208–141239, 2024.

[42] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.