Adaptive Robot Skill Transfer Through Multimodal Demonstration Learning and Uncertainty-Aware Policy Optimization

Leni Weller
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
lweller@ku.edu

Vilhelm Sjodin
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
sjodin930@missouri.edu

Katja Klingelhofer
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
klingelhoferkatja@colostate.edu

Jeremi Dydak
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
jeremi.dydak@uc.edu

Abstract

Adaptive transfer of manipulation skills from human demonstrations remains difficult because demonstrations are multimodal, task contexts shift, and policy errors are unevenly costly. This manuscript develops a methodological framework that couples multimodal demonstration learning with uncertainty-aware policy optimization for context-adaptive robot skill transfer. The framework learns a shared skill latent from vision, proprioception, force or tactile signals, and language-conditioned task descriptors; aligns modalities through learned attention; and infers a posterior over skills from limited adaptation data. The policy optimization objective decomposes predictive uncertainty into aleatoric and epistemic components, penalizes conditional value-at-risk of return, and enforces trust-region and safety constraints during adaptation. The central contribution is a coherent integration of representation, probabilistic context inference, and risk-sensitive optimization, rather than a single new network component. Analytically, the framework clarifies when multimodal fusion improves transfer, why epistemic uncertainty should regulate policy updates under distribution shift, and how risk penalties alter the exploration-exploitation trade-off. The paper also identifies failure modes: spurious cross-modal correlations, overconfident extrapolation, contact-rich aleatoric noise, and constraint violation under abrupt embodiment changes. The resulting design provides a technical basis for deploying demonstration-learned robot skills that remain reliable when task, object, or embodiment conditions differ from the demonstration distribution.

Keywords

multimodal imitation learning; robot skill transfer; uncertainty quantification; meta-learning; risk-sensitive policy optimization; safe robot learning

1. Introduction

Robot skill transfer concerns the reuse of demonstrated competencies across objects, task variants, and embodiments without requiring complete relearning in every new setting. Imitation learning provides a direct route from human demonstrations to robot policies [1]. Dynamical movement primitives encode attractor dynamics for stable motion generation [2], while deep visuomotor policies map raw observations to actions through learned representations [3]. Adaptation to new task contexts has been approached through meta-learning and probabilistic context inference, allowing policies to infer task variables from limited interaction data [4], [5]. Yet deterministic policy learning can produce overconfident behavior in unfamiliar contexts; predictive uncertainty is therefore not ancillary but constitutive of safe transfer [6]. A recent design-space analysis of imitation-learning policies further shows that encoder choice, action parameterization, and data regime interact strongly [7]. Offline demonstration studies also indicate that data coverage and action representation affect manipulation performance [8], and vision-based multi-task imitation has demonstrated the feasibility of end-to-end skill reuse [9]. Visual pre-training [10] and large-scale transformer policies [11] improve representation scale, but scale alone does not resolve context ambiguity, multimodal conflict, or risk-sensitive action selection.

The central problem addressed here is how a robot can acquire a compact, transferable skill representation from multimodal demonstrations and then adapt that representation under uncertainty without violating safety constraints. The contribution is an integrated methodological framework with three linked components. First, multimodal demonstrations are encoded into a shared latent skill space that separates task intent, motion style, and contact strategy. Second, a context posterior infers which skill latent is appropriate for a new task or embodiment condition using limited adaptation data. Third, policy optimization treats uncertainty as a first-class quantity: epistemic uncertainty governs adaptation step size and exploration, aleatoric uncertainty shapes risk-sensitive action selection, and safety constraints remain outside the soft reward trade-off. The framework is developed analytically and connected to existing work on imitation learning, probabilistic dynamics models, calibrated uncertainty, and constrained reinforcement learning.

2. Problem Formulation and Design Requirements

Consider a robot with state s in S, action a in A, and task context c in C. Demonstrations for a context are trajectories whose observations combine vision, proprioception, force or tactile measurements, and optional language instructions. A skill latent z summarizes reusable behavioral factors such as motion primitives, grasp strategies, approach directions, and compliance patterns. The transfer target is a context-conditioned policy obtained by marginalizing over the posterior distribution of skill latents rather than selecting a single task label.

$$\pi*(a|s,c) = \int \pi(a|s,z,c)\, p(z|D,c)\, dz$$

Here $\pi(a|s,z,c)$ is a skill-conditioned policy, $p(z|D,c)$ is the posterior over skill latents given demonstration or adaptation data D and context c, and the integral expresses transfer as probabilistic averaging over plausible skills. This formulation avoids brittle hard assignment

when two demonstrations are consistent with the same context or when a new object admits more than one grasp strategy. It also makes the quality of transfer depend explicitly on posterior calibration: a narrow posterior around an inappropriate latent produces confident but wrong behavior, whereas an overly diffuse posterior produces hesitant and inefficient control.

Four requirements follow. First, multimodal fusion must be robust to missing or conflicting modalities, because force sensors may saturate, vision may be occluded, and language may be underspecified. Second, context inference must update from few trials and must represent ambiguity rather than collapsing to a point estimate. Third, uncertainty must distinguish sensor and contact noise from model ignorance caused by distribution shift. Fourth, policy optimization must respect safety constraints and avoid abrupt policy changes when the context posterior is poorly identified. These requirements motivate the architecture and optimization logic developed in the following sections.

3. Multimodal Demonstration Learning

3.1 Observation representation and cross-modal fusion

Each modality provides a partial view of the manipulation state. Vision captures object geometry and spatial relations; proprioception captures joint configuration and velocity; force or tactile signals capture contact onset and slip; language descriptors provide task intent when available. Multimodal learning benefits from representations that preserve modality-specific information while allowing task-dependent weighting [12]. The proposed fusion step uses attention over modality encoders so that the policy can emphasize the most informative signal for the current decision.

$$h_t = \sum_m \alpha_{t,m} f_m(o_{t,m}), \alpha_{t,m} \propto \exp\left(q^T f_m(o_{t,m})\right)$$

In this expression, o_{t,m} is the observation from modality m at time t, f_m is its encoder, h_t is the fused representation, q is a learned query vector, and α_{t,m} is the normalized attention weight for modality m. The proportionality symbol indicates normalization over modalities. Attention is not merely a feature-combination device; it provides an interpretable reliability signal. When force feedback is absent, a masked attention variant can renormalize over available modalities. When visual and tactile evidence conflict, the learned weights indicate which modality the policy is relying on, enabling later checks against calibration and safety margins.

3.2 Skill latent and probabilistic dynamics

The fused representation feeds a skill encoder that maps demonstration segments to a latent z and a context-conditioned dynamics model that predicts the next state distribution. Latent world models motivate this separation between compact state abstraction and predictive rollout [13], and probabilistic ensemble dynamics provide a practical basis for estimating model uncertainty [14]. The dynamics model is not treated as an oracle. It supplies a predictive distribution used for uncertainty estimation and, where appropriate, short-horizon planning. A heteroscedastic output captures aleatoric noise from contact and actuation, while disagreement across an ensemble or approximate posterior samples captures epistemic uncertainty about unexplored state-action regions.

This separation matters for skill transfer. A new object may introduce high aleatoric uncertainty because its surface friction is noisy, even when the policy has extensive experience with similar objects. Conversely, a new embodiment may produce high epistemic

uncertainty because the mapping from actions to observed motion has shifted outside the training distribution. The former calls for cautious contact control and impedance adjustment; the latter calls for slower adaptation, information-seeking behavior, or additional demonstration data. Conflating the two would either make the robot unnecessarily timid in familiar but noisy tasks or dangerously confident in unfamiliar dynamics.

3.3 Context-conditioned adaptation

Given limited adaptation data from a new context, the skill posterior is updated from a prior over reusable skills and a likelihood induced by demonstration or interaction outcomes.

$$p(z|c, D) \propto p(D|c, z)p(z|c)$$

Here p(z|c) is a context-conditioned prior learned from related tasks, and p(D|c,z) is the likelihood of observed trajectories or outcomes under skill z. The posterior supports online or few-shot adaptation without discarding the structural knowledge encoded in the prior. In practice, a variational or particle approximation can represent multimodality in the posterior, and the policy can condition on posterior samples during optimization. The analytical consequence is that adaptation speed is controlled by the informativeness of D and the alignment between c and the prior context distribution. If a new task differs only in object color or position, the likelihood should sharpen quickly. If it changes contact mechanics or kinematic structure, the prior may be misleading, and the posterior must remain broad enough to signal epistemic uncertainty.

3.4 Uncertainty estimation and calibration

Predictive uncertainty is decomposed into aleatoric and epistemic terms according to the action and context under consideration.

$$u(a, s, c) = u_a le(a, s, c) + u_e pi(a, s, c)$$

The term u_ale(a,s,c) represents irreducible noise from sensing, contact, and actuation, while u_epi(a,s,c) represents reducible uncertainty due to limited data or model mismatch. Deep ensembles provide a robust epistemic estimator through disagreement among independently trained predictors [15], and approximate Bayesian inference through dropout or variational layers offers a complementary route to posterior uncertainty [16]. Calibration is essential: the policy should not treat a high-variance but well-centered prediction as necessarily dangerous, nor should it interpret a low-variance prediction as reliable when the ensemble agrees for the wrong reason. Calibration checks should therefore consider held-out contexts and safety-relevant tail events, not only average predictive error.

The uncertainty decomposition directly informs control. Aleatoric uncertainty should enter risk-sensitive action selection and compliance control, because no amount of additional data removes contact noise. Epistemic uncertainty should enter adaptation gating, exploration, and constraint tightening, because it indicates that the current policy or dynamics model may be outside its support. This distinction also prevents a common failure in demonstration learning: a policy that becomes progressively more conservative as sensor noise increases, even though the correct response is to filter or adapt impedance rather than to stop exploring task-relevant strategies.

4. Uncertainty-Aware Policy Optimization

4.1 Risk-sensitive objective

Expected-return maximization weights rare high-cost outcomes only through their probability, which is inadequate for contact-rich transfer where a small number of collisions or excessive forces can dominate deployment cost. Let R be the discounted return over a trajectory. The optimization objective can penalize the lower tail of the return distribution while retaining an expected-performance term.

$$J(\pi) = E[R] \; - \; \lambda \, CVaR_\beta(-R)$$

Here E[R] is the expected return, CVaR_β(−R) is the conditional value-at-risk of the negative return at tail level β in (0,1), and λ is a nonnegative risk weight. The conditional value-at-risk term penalizes the average of the worst β fraction of outcomes, making the objective sensitive to low-probability but high-cost events [18]. Constrained policy optimization provides a related language for separating reward maximization from constraint satisfaction [17]. In the present framework, λ need not be fixed across contexts: it can increase when epistemic uncertainty is high or when the task enters a safety-critical phase, and decrease when the policy is calibrated and constraints have sufficient margin.

4.2 Trust-region and safety constraints

Adaptation can be unstable when the context posterior changes abruptly or when the dynamics model is biased. A trust-region update limits the change in policy distribution per update, while explicit constraints encode hard safety requirements such as joint limits, maximum contact force, collision margins, or workspace boundaries.

$$max_\theta \; E[A_\theta(s,a)] \; s.t. \; E[D_KL(\pi_\theta \; || \; \pi_\theta old)] \; \leq \; \delta, E[C_i] \; \leq \; d_i$$

In this expression, θ denotes policy parameters, A_θ(s,a) is the advantage estimate, D_KL is the Kullback-Leibler divergence between the updated policy π_θ and the previous policy π_θold, δ is the trust-region radius, C_i is the i-th safety cost, and d_i is its permitted expectation. Safe reinforcement learning surveys emphasize that constraints should be treated as distinct from the reward signal rather than folded into a single scalar penalty [19]. The trust-region radius can be made adaptive: when u_epi is high, a smaller δ restricts policy change and forces the system to gather evidence before committing to a new skill interpretation. Hard constraints remain active even if the posterior favors an aggressive action, which is essential when demonstration data underrepresent near-contact failure modes.

4.3 Adaptation procedure

The policy update alternates three operations. It first updates the multimodal representation and skill posterior using available demonstration or interaction data. It then estimates aleatoric and epistemic uncertainty from the probabilistic dynamics model and policy ensemble. Finally, it solves the constrained risk-sensitive policy update, using epistemic uncertainty to modulate exploration and trust-region size and aleatoric uncertainty to modulate risk sensitivity. The underlying decision process follows standard Markov decision process assumptions [20], but the optimization layer adds context-conditioned skill inference and explicit uncertainty handling. Model-based uncertainty estimates can be used even when policy optimization is model-free, provided that model bias is monitored and does not dominate the risk estimate. When model bias is large, the framework should rely more on ensemble policy disagreement and conservative constraints than on simulated rollouts.

5. Analytical Assessment and Technical Discussion

The proposed framework is best understood as a coordination mechanism among representation, inference, and optimization. Behavioral cloning alone can reproduce

demonstrated actions but often lacks a calibrated mechanism for deciding when a new context is out of distribution. Dynamical movement primitives provide stable motion structure but do not by themselves solve multimodal context inference or risk-sensitive policy selection. Meta-learning and probabilistic context variables support rapid adaptation, but their benefit depends on task distributions that cover the relevant variation. The present design uses a skill latent and context posterior to connect these approaches while preserving uncertainty information that point-estimate adaptation would discard.

Multimodal fusion improves transfer when modalities provide complementary evidence. Vision may identify object geometry, force may indicate slip, and proprioception may reveal unexpected resistance. Attention can exploit this complementarity, but it can also learn spurious correlations if one modality is always present during training and absent during deployment. Reliability gating, modality dropout, and calibration on missing-modality conditions are therefore not optional refinements; they are conditions for trustworthy fusion. A related failure occurs when language instructions are too coarse to disambiguate grasp strategy. In that case, the posterior over z should remain multimodal, and the policy should either seek information or choose a low-risk action that is acceptable across plausible skills.

The uncertainty decomposition has clear analytical consequences. High aleatoric uncertainty in contact-rich manipulation implies that increased data collection will not eliminate the risk; the appropriate response is compliant control, force limiting, or conservative impedance selection. High epistemic uncertainty implies that the policy is extrapolating; the appropriate response is to reduce trust-region size, increase exploration cautiously, or request additional demonstration. If the two are conflated, the system may either overfit to noisy contact signals or fail to recognize genuine model ignorance. The risk-sensitive objective further changes the exploration-exploitation trade-off: increasing λ penalizes low-tail outcomes and can produce safer but less aggressive behavior, while decreasing λ recovers expected-return behavior when the context is well understood and constraints provide sufficient protection.

Constraint handling is the main boundary condition for the framework. Trust-region updates and CVaR penalties improve risk awareness, but they cannot compensate for an incorrectly specified cost function or an unmodeled contact mode. Safety constraints must therefore be derived from physical limits and verified where possible, and the policy should treat near-boundary states as requiring additional caution. Abrupt embodiment changes are especially challenging because the dynamics model, action space, and observation statistics may shift simultaneously. In such cases, the prior over skills may be misleading, and the posterior should widen rather than collapse to the nearest demonstrated behavior. The framework's reliance on epistemic uncertainty as an adaptation gate is designed for this condition, but its effectiveness depends on ensemble diversity, calibration, and the availability of at least a small amount of informative adaptation data.

The framework also clarifies a design trade-off between interpretability and capacity. A shared latent skill space with attention over modalities is more structured than a monolithic end-to-end policy, which aids diagnosis when transfer fails. However, excessive structure can limit the ability to represent novel skills that are not compositional combinations of demonstrated ones. The balance should be set by the expected degree of context shift: modest object and position changes favor strong priors and rapid posterior updates, whereas substantial embodiment changes favor broader priors, stronger epistemic signals, and more conservative policy updates.

6. Conclusion

This manuscript developed an integrated framework for adaptive robot skill transfer that couples multimodal demonstration learning with uncertainty-aware policy optimization. Demonstrations are encoded through modality-specific representations and attention-based fusion, a skill latent captures reusable behavioral structure, and a context posterior supports adaptation without hard task assignment. The policy optimization layer separates aleatoric from epistemic uncertainty, penalizes conditional value-at-risk, and enforces trust-region and safety constraints during adaptation. The central analytical insight is that transfer reliability depends less on a single representation choice than on the alignment among multimodal evidence, calibrated context inference, and risk-sensitive constrained updates. The framework also identifies concrete failure modes, including spurious cross-modal correlations, posterior collapse under misleading priors, overconfident extrapolation, and constraint violation under abrupt embodiment changes. These conditions define the technical requirements for deploying demonstration-learned robot skills beyond the narrow distribution of the original demonstrations.

References

1. Schaal, S. (1999). Is imitation learning the route to humanoid robots? Trends in Cognitive Sciences, 3(6), 233–242.

2. Ijspeert, A. J., Nakanishi, J., Hoffmann, H., Pastor, P., & Schaal, S. (2013). Dynamical movement primitives: Learning attractor models for motor behaviors. Neural Computation, 25(2), 328–373.

3. Levine, S., Finn, C., Darrell, T., & Abbeel, P. (2016). End-to-end training of deep visuomotor policies. Journal of Machine Learning Research, 17(39), 1–40.

4. Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the 34th International Conference on Machine Learning (pp. 1126–1135). PMLR.

5. Rakelly, K., Zhou, A., Quillen, D., Finn, C., & Levine, S. (2019). Efficient off-policy meta-reinforcement learning via probabilistic context variables. In Proceedings of the 36th International Conference on Machine Learning (pp. 5331–5340). PMLR.

6. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In Advances in Neural Information Processing Systems 30 (pp. 5574–5584).

7. Jia, X., Donat, A., Huang, X., Zhao, X., Blessing, D., Zhou, H., ... & Neumann, G. (2025). X-il: Exploring the design space of imitation learning policies. arXiv preprint arXiv:2502.12330.

8. Mandlekar, A., Xu, D., Wong, J., Nasiriany, S., Wang, C., Kwan, R., ... & Martín-Martín, R. (2021). What matters in learning from offline human demonstrations for robot manipulation. In Proceedings of the 5th Conference on Robot Learning (pp. 1678–1690). PMLR.

9. Rahmatizadeh, R., Abolghasemi, P., Bölöni, L., & Levine, S. (2018). Vision-based multi-task manipulation for inexpensive robots using end-to-end learning from demonstration. In 2018 IEEE International Conference on Robotics and Automation (pp. 3758–3765). IEEE.

10. Radosavovic, I., Xiao, T., James, S., Abbeel, P., Malik, J., & Darrell, T. (2022). Real-world robot learning with masked visual pre-training. In Proceedings of the 6th Conference on Robot Learning (pp. 416–426). PMLR.

11. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., ... & Zitkovich, B. (2022). RT-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817.

12. Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443.

13. Ha, D., & Schmidhuber, J. (2018). Recurrent world models facilitate policy evolution. In Advances in Neural Information Processing Systems 31 (pp. 2450–2462).

14. Chua, K., Calandra, R., McAllister, R., & Levine, S. (2018). Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In Advances in Neural Information Processing Systems 31 (pp. 4754–4765).

15. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in Neural Information Processing Systems 30 (pp. 6402–6413).

16. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning (pp. 1050–1059). PMLR.

17. Achiam, J., Held, D., Tamar, A., & Abbeel, P. (2017). Constrained policy optimization. In Proceedings of the 34th International Conference on Machine Learning (pp. 22–31). PMLR.

18. Rockafellar, R. T., & Uryasev, S. (2000). Optimization of conditional value-at-risk. Journal of Risk, 2(3), 21–42.

19. García, J., & Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. Journal of Machine Learning Research, 16(1), 1437–1480.

20. Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.