**PoCo: Policy Composition from and for Heterogeneous Robot Learning Paper Review**

Paper Review by Tyler Kim

**Summary**

In Lirui Wang et al’s *PoCo*: *Policy Composition from and for Heterogenous Robot Learning*, the researchers describe an approach to train a robot to perform various actions using data that is derived from different sources and modalities. Current robot learning pipelines trained specialized models for a single robot for a single task without any behavior. Additionally, training on data that are inherently from different distributions and sensors proved difficult. Therefore, the authors propose a framework called Policy Composition (PoCo) which aims to compose data derived from different modalities, sources, and tasks. **The main contribution of the paper are PoCo, the framework that uses diffusion models for combine data from different domains and modalities, developing task-, behavior-, and domain-level composition for creating policies without retraining, and illustrating scene- and task-level generalization of PoCo across simulation and real-world settings.**

Policy Composition (PoCo) uses a diffusion model to represent a policy where a trajectory is generated given a history of observation denoted as using as the loss function. The general process of PoCo to input the heterogenous data has multiple steps. Modalities are defined as where each modality is defined as information from a particular sensor. Next, data domains , which could be simulations, real-world robots, and human demonstrations, are considered as long as they share the same action space. Then, constraints are considered on a desired behavior via a cost function which in the paper are and . Finally, a robot task is specified via a natural language command. A separate probabilistic model is learned on tuple .

A diagram of a computer program

Description automatically generated with medium confidence

The paper proposes to sample from the product of different distributions in hopes to combine the score predictions from diffusion policies at inference time. The product of the composition will yield a trajectory that will most likely accomplish the task under multiple distributions. However, assumptions must be made: mutual dependence of tasks and costs and conditional independence of tasks and costs given a trajectory . The implementation is represented as energy-based models.

The experiments were divided into two types: simulated and real-world. During the simulation experiments, the researchers reported gains in smoothness and safety constraints adding prior information when composing behaviors. They also report that task composition performs the best in multitask policy evaluation. Finally, they found that their approach can generalize across multiple distractors.

A black and white text with numbers

Description automatically generated

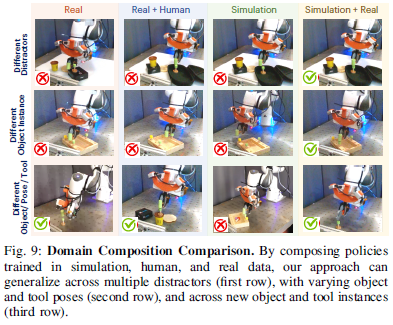
A collage of several images of a machine

Description automatically generated

A graph of different sizes and colors

Description automatically generated with medium confidence

For the real-world experiments, the researchers found that policy composition improves success rate across different scenes on four generalization axes: varying object poses, varying robot initial pose, varying tool poses, adding distractor objects, and replacing objects with novel instances from the same class. The researchers also found that multitask policies can perform on par with task-specific policies, stable in dexterous tasks, classifier-free training performs better than naively concatenating features, and composition hyperparameters must stay within a range to remain effective.



A table with numbers and text

Description automatically generated

A table with text and numbers

Description automatically generated

In the ablation study, researchers reported that a naïve approach results in a 75% performance drop.

A close-up of a text

Description automatically generated

A graph with a line

Description automatically generated

**Strengths**

The paper provides a new approach to handle heterogeneous data via policy composition. More importantly, it opens the door to further research in a more generalized robot particularly with more sensors and various data distributions. More interestingly, composing policies such that a trajectory could be sampled in a way that has the highest likelihood among multiple distributions. One thing I learned from this paper was a possible approach to handle heterogeneous data for a unified goal. More specifically, I learned that simply multiplying distributions together could serve as a feasible way to create a common distribution among multiple modalities. In fact, I thought it was a very clever way to approach the problem. The PoCo framework in general, I thought was very novel in the realms of robotics.

**Potential Improvements**

One thing I think the paper could improve upon to get closer to its stated goal/contributions to generalize PoCo more. Unfortunately, policy composition works only under certain assumptions as stated in the paper, that is, the action space must be the same. Another improvement that I think the paper could improve is handling the time consumption off the models. Since a separate diffusion model is trained for various tuples, this would mean parallelism is extremely important or else the framework would take too long to train/infer. The paper seems to be using the latest techniques or models but one thing they could do is have some sort of memory compression system as seen in LRLL. This would help reduce memory use significantly. In addition, using primitive abilities to learn new skills as an addition to PoCo would take advantage of the versatility of multiple data distributions and modalities.

**Extensions**

There are a couple of extensions or follow-ups I could think of for this paper. One possible extension would be creating a more generalized diffusion model for the different modality, domain, task, and behavioral cost tuples instead of having a separate one for each instance. Another extension would be to try a new approach to compose data from varying distributions and modalities such that it does not rely on the assumption that the action space must be the same. The final extension would be to expand PoCo to a larger variety of tasks instead of just using a couple of tools.