

Bayesian Imitation Learning for End-to-End Mobile Manipulation

Yuqing Du, Daniel Ho, Alexander A. Alemi, Eric Jang, Mohi Khansari

Motivation

Challenges we would like generalist robots to handle:

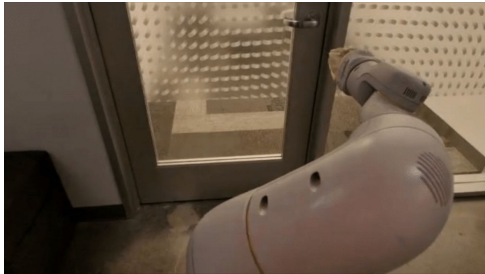
- A range of different skills – navigation, manipulation, or combinations of the two.
- Unstructured, dynamic, real world environments.
- Using egocentric, on-board sensor inputs (i.e. no object tracking).

Our goal: use learning from demonstrations (LfD) to train a multi-sensor visuomotor policy for mobile manipulation in natural human environments.

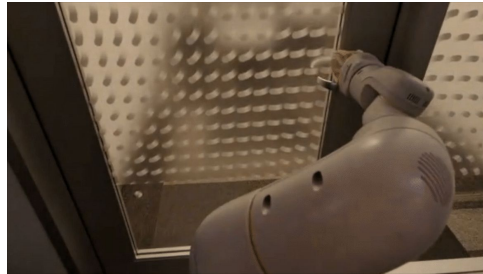


Why latched door opening?

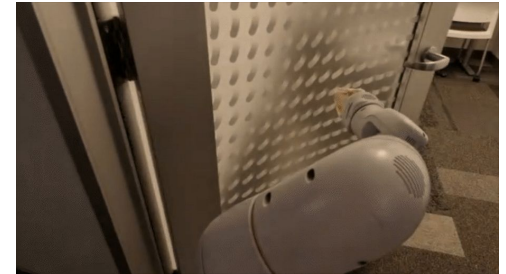
- There are multiple subtasks within a single episode,
- Need to coordinate both navigation and manipulation,
- Operate in a natural human environment.



Going to the door



Unlatching & Opening



Going inside

Meat of the Problem: Sim-to-Real

Evaluating robot policies for the real world is expensive, dangerous, and time-consuming, so we typically rely on sim (to some degree).



Sim-to-real gap: Differences in dynamics and visuals between simulation and reality \Rightarrow policies can perform differently between the two domains.

This is only exacerbated when we introduce multiple sensor modalities.

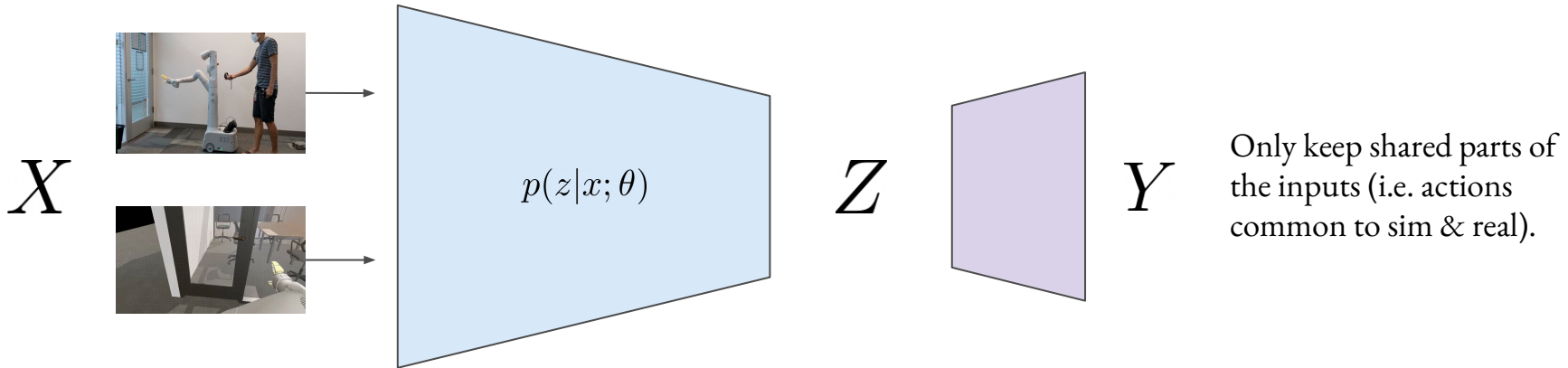
Bayesian Imitation Learning for End-to-End Mobile Manipulation

How can we learn a multi-sensor visuomotor policy with the following properties?

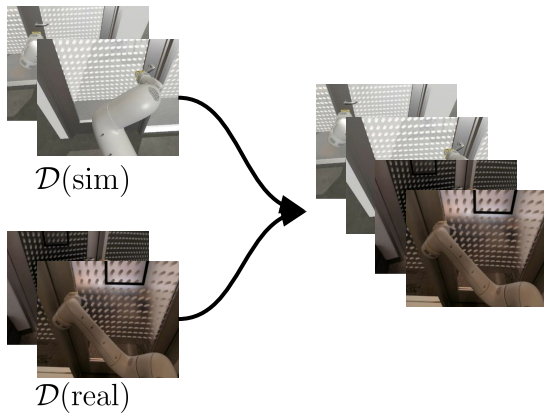
- 1) Sim and real domain agnostic for each sensor,
- 2) Able to estimate each sensor's 'uncertainty' level at each timestep,
- 3) Enable easy fusion or switching between different sensor modalities.

Variational Information Bottleneck (Alemi et al., 2016)

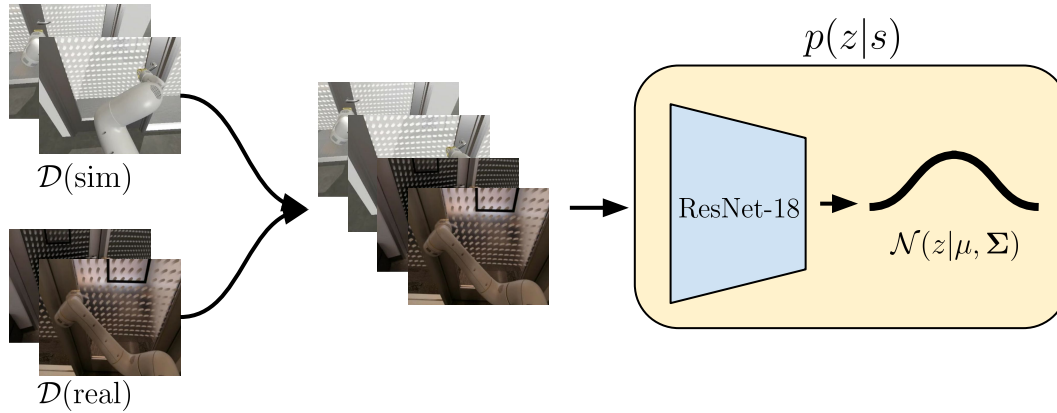
$$\max_{\theta} I(Z; Y | \theta) - \beta I(Z; X | \theta)$$



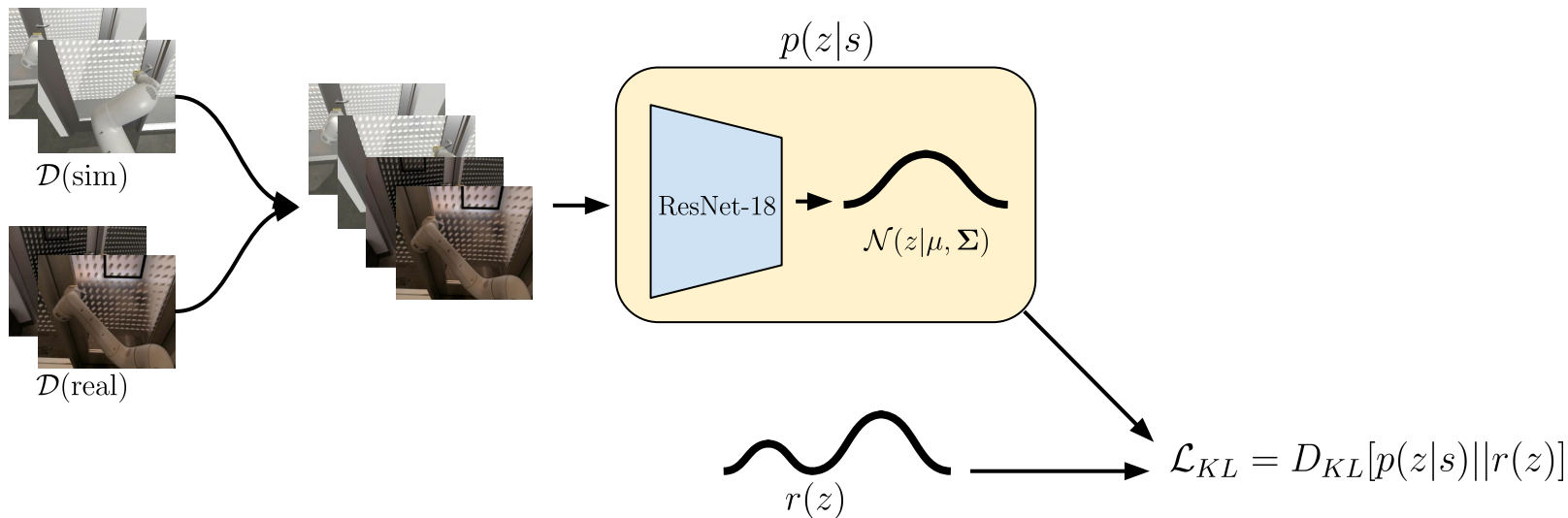
VIB Architecture



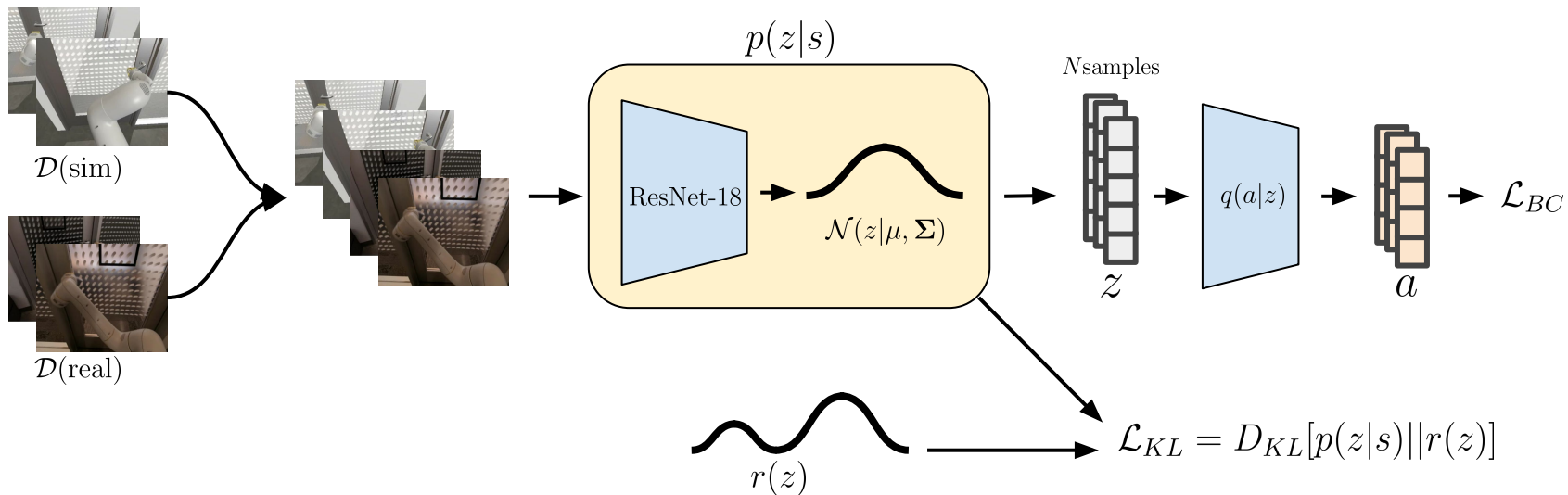
VIB Architecture



VIB Architecture



VIB Architecture



Model Expressiveness

approach



unlatch



enter



Model Expressiveness

approach



141



161



169

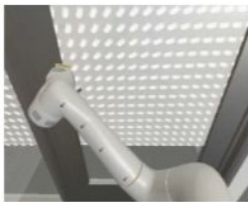


172



173

unlatch



200



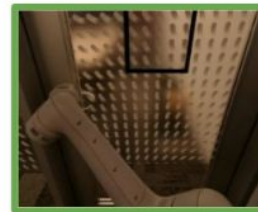
225



232



232



233

enter



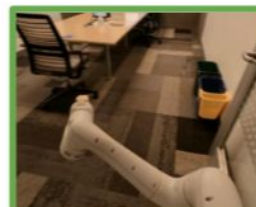
57



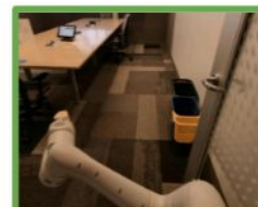
71



92



96



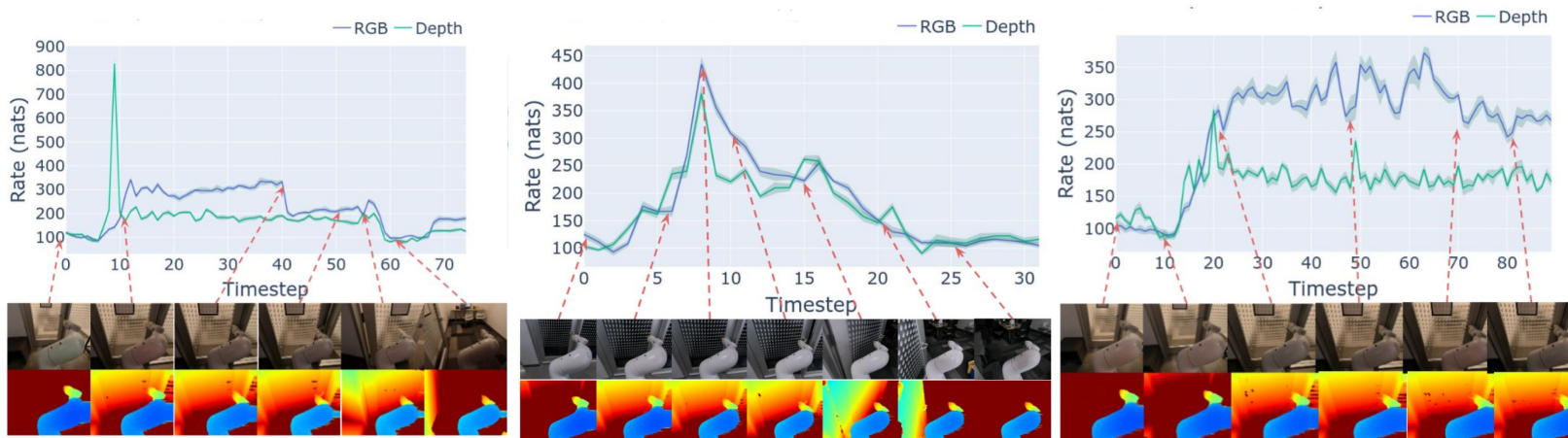
98

Model Explainability

$$\mathcal{L}_{KL} = D_{KL}[p(z|s)||r(z)] \rightarrow \text{'rate'}$$

The rate tells us how much information is needed to encode a particular input.

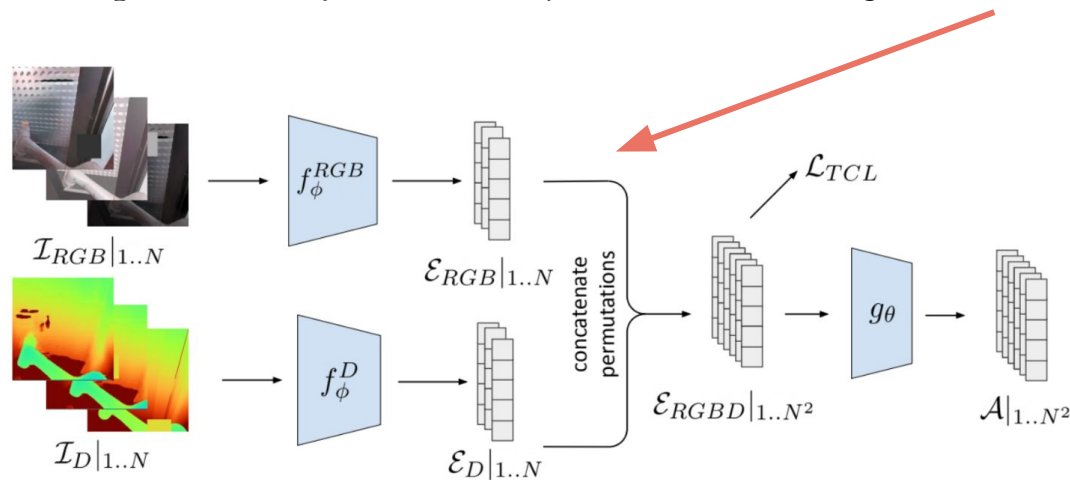
We find that the rate across a trajectory has a pattern of being highest when opening the door.



Model Explainability

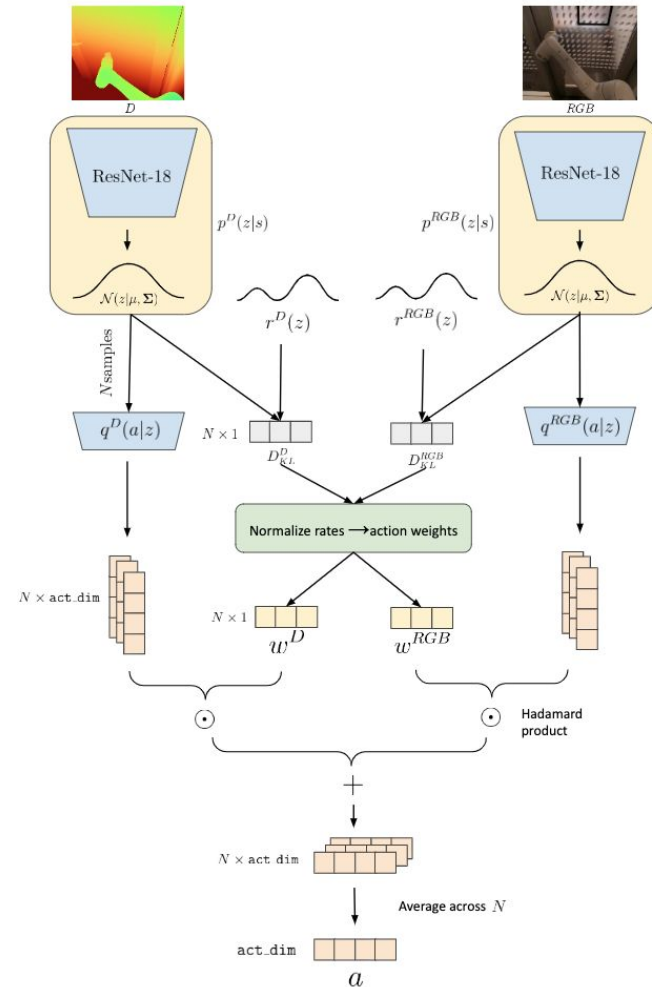
$$\mathcal{L}_{KL} = D_{KL}[p(z|s)||r(z)] \rightarrow \text{'rate'}$$

Rarer inputs (i.e. different from training data) typically correlate with higher rates, so we propose using it as a measure of sensor modality ‘uncertainty’. This allows us to fuse modalities in a more interpretable way, instead of just concatenating embeddings.



Sensor Fusion

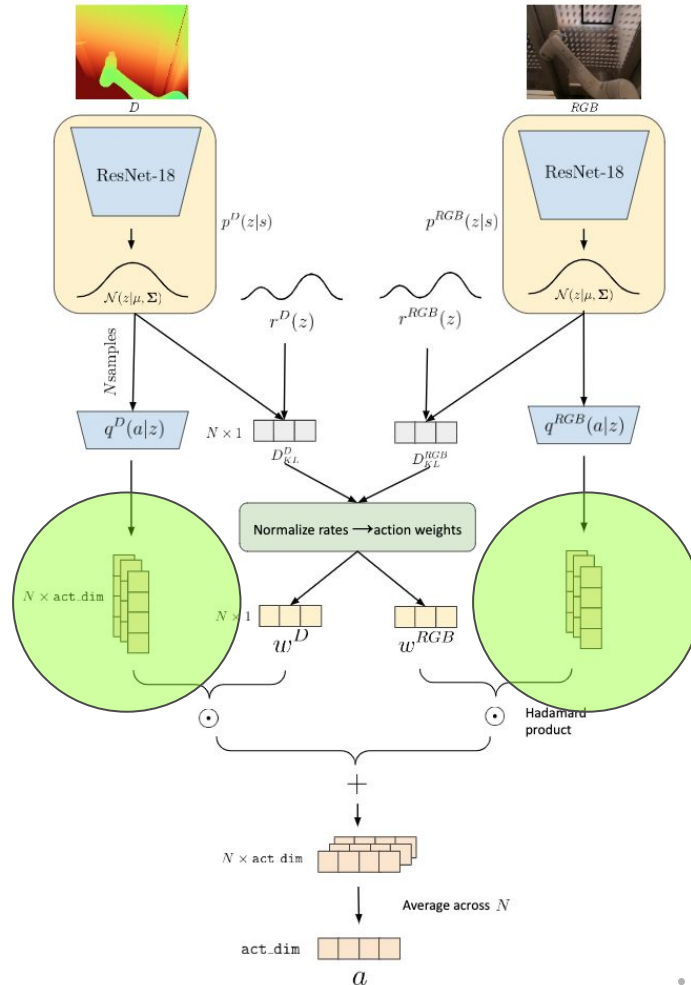
Model architecture for sensor fusion



Sensor Fusion

Model architecture for sensor fusion

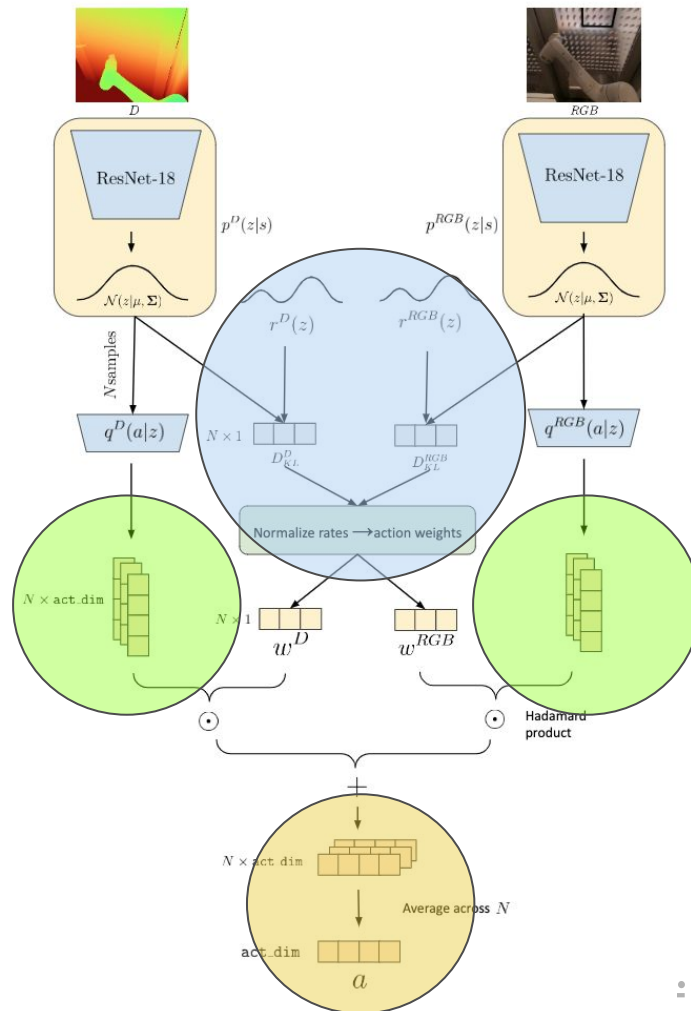
- Each sensor tower predicts actions individually



Sensor Fusion

Model architecture for sensor fusion

- Each sensor tower predicts actions individually
- We use inverse of rates as to balance the contribution of each tower's action predictions.



Latched Door Opening: Evaluations

Method	Total (10 rooms)	Seen (6 rooms)	Unseen (4 rooms)
<i>RGB - Naive (baseline)</i>	46% ± 2.5	48% ± 3.7	42% ± 4.5
<i>RGB - GAN (baseline)</i>	62% ± 2.5	56% ± 3.7	71% ± 4.2
<i>RGB - TCL (baseline)</i>	80% ± 2.3	75% ± 3.2	87% ± 3.1
<i>Depth - TCL (baseline)</i>	77% ± 2.2	79% ± 3.1	75% ± 4.2
<i>RGBD - TCL (baseline)</i>	75% ± 2.2	79% ± 3.0	69% ± 4.3
<i>RGBD - TCL - VIB (Softmax Fusion)</i>	96% ± 1.6	98% ± 1.0	93% ± 2.3
<i>RGBD - TCL - VIB (Linear Fusion)</i>	93% ± 1.5	93% ± 1.9	93% ± 2.3
<i>RGBD - TCL - VIB (Concat.)</i>	96 % ± 1.1	94% ± 1.8	98% ± 1.3

Thank you!