Behavior Generation with Latent Actions

VQ-BeT: Action multi-modality through tokenization

Seungjae Lee, Yibin Wang, Haritheja Etukuru, H. Jin Kim,
Nur Muhammad Mahi Shafiullah*, Lerrel Pinto*

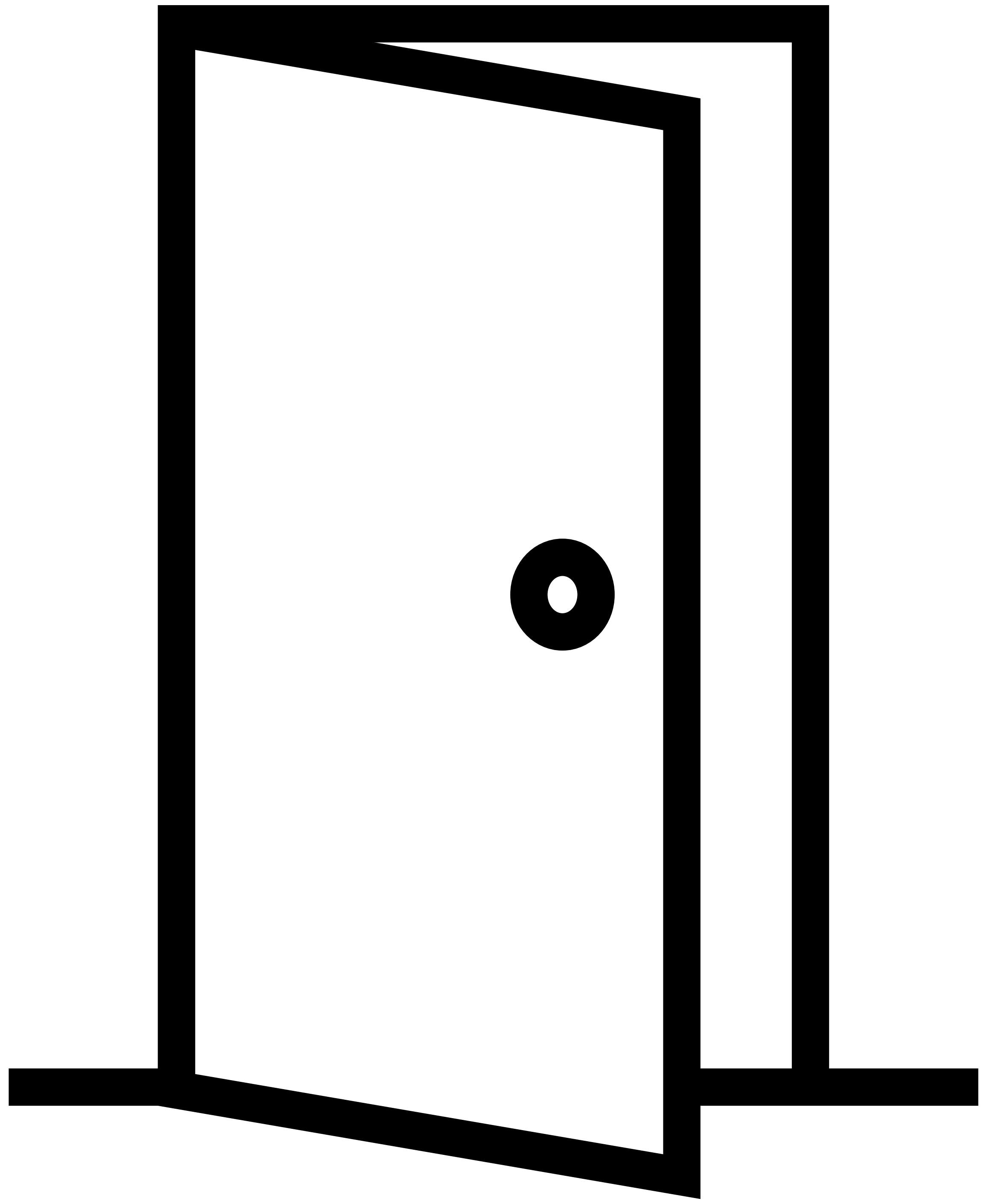
https://sjlee.cc/vq-bet

Difficulties of adapting BC to real world

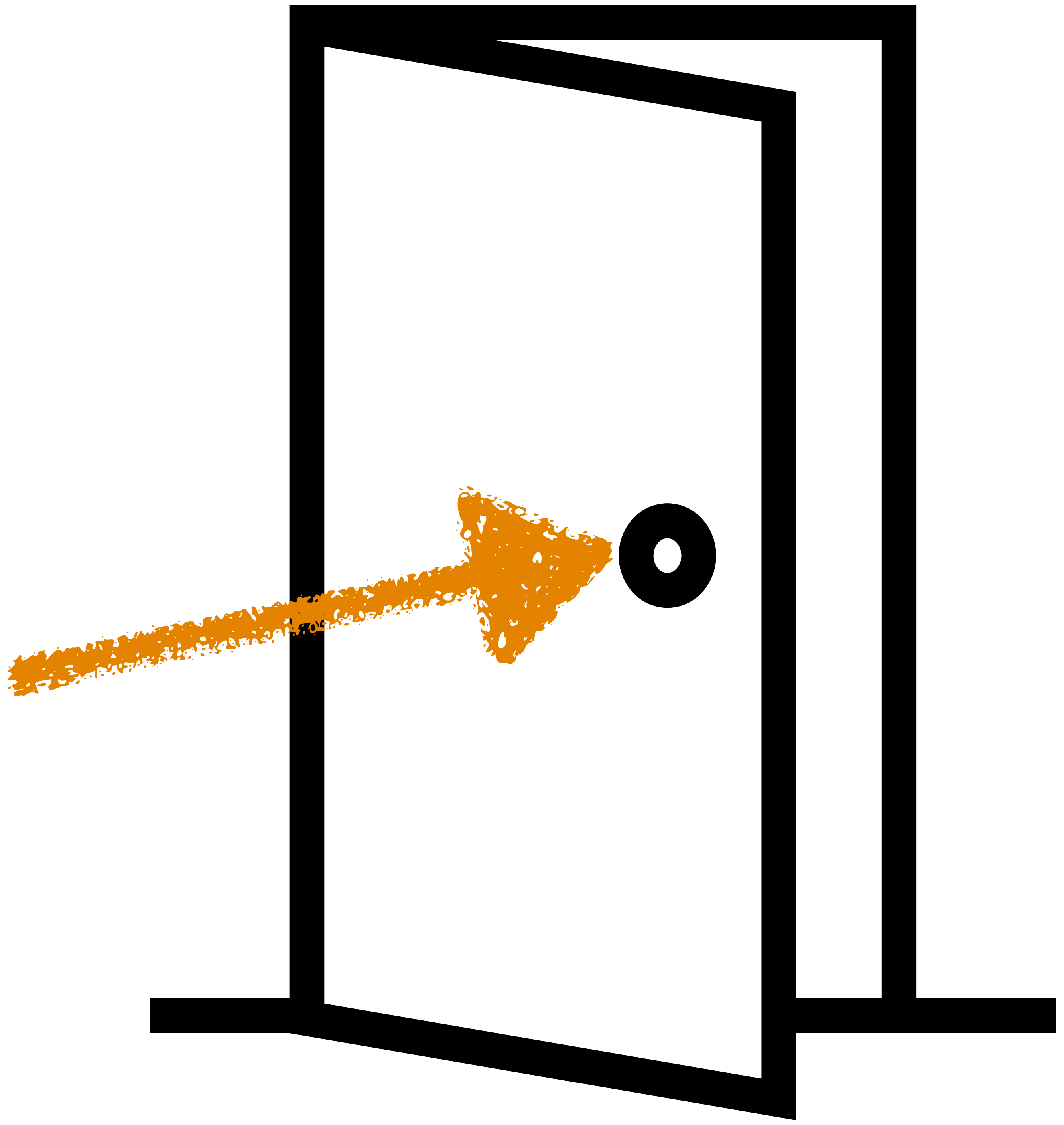
What are the big challenges?

- Expert demos are expensive and sometimes come without a reward label.
- Modeling behavior from demonstrations can have multiple modes.
- Environments are not Markovian.

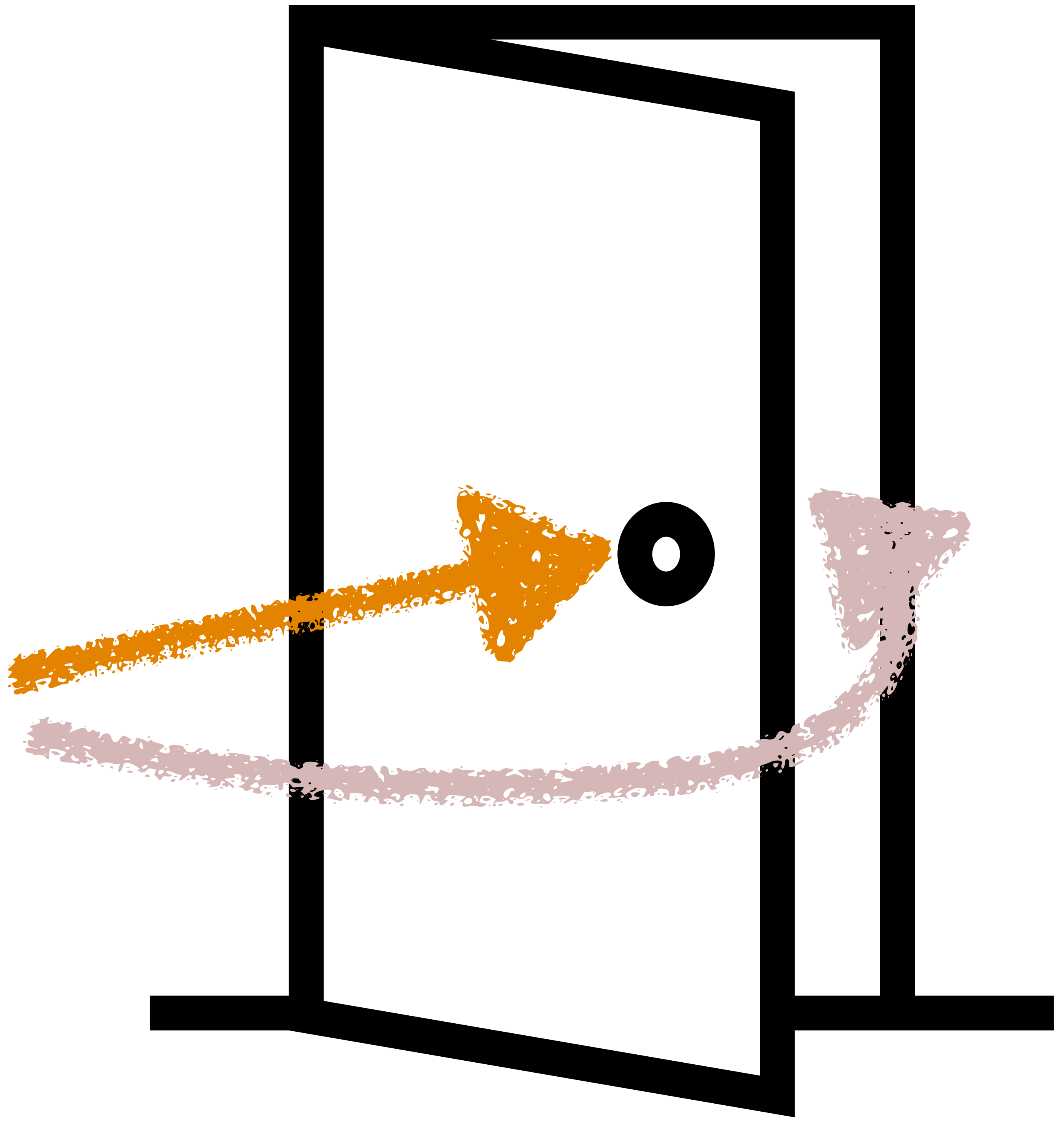
Consider opening this door



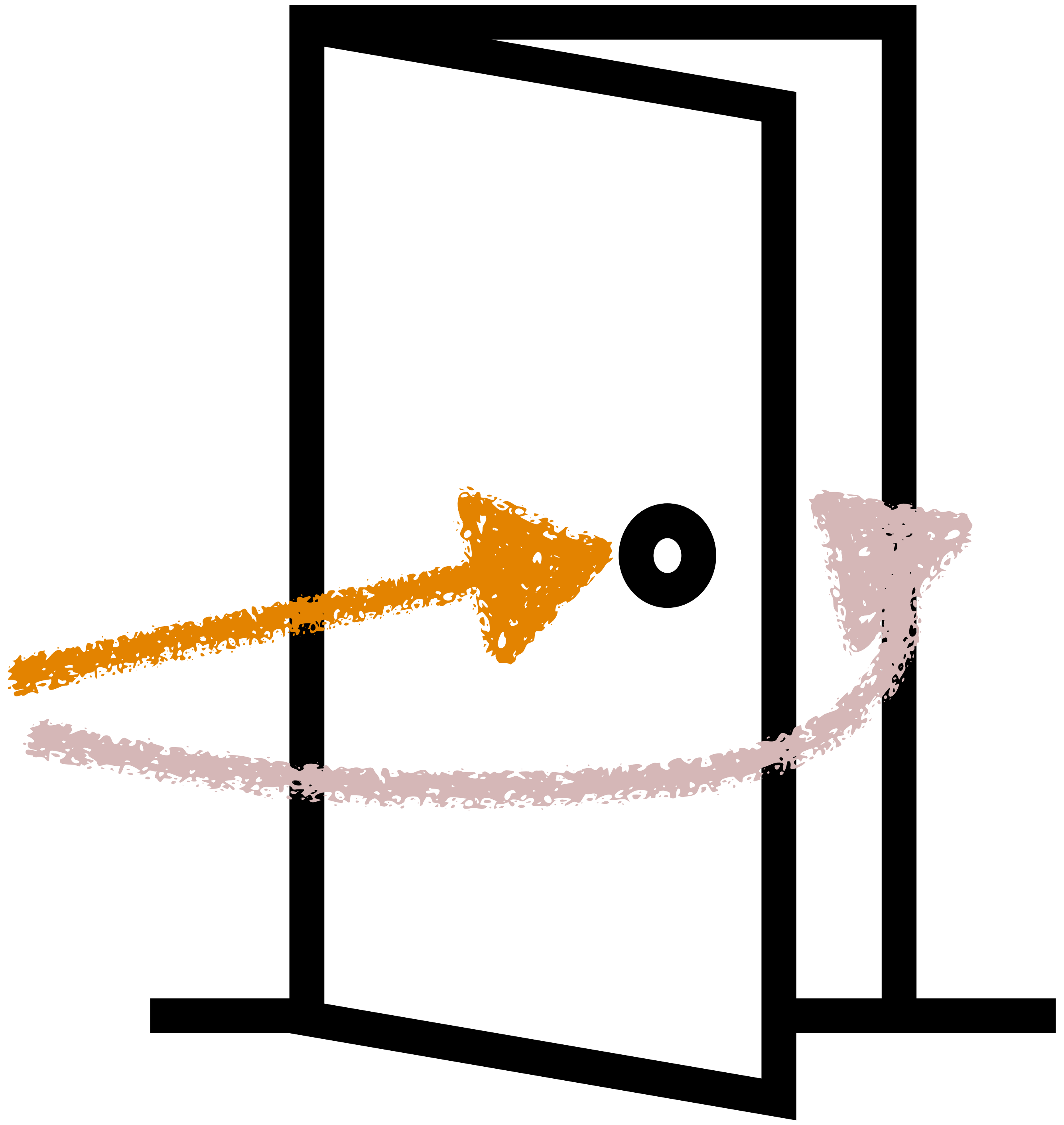
Consider opening this door



Consider opening this door

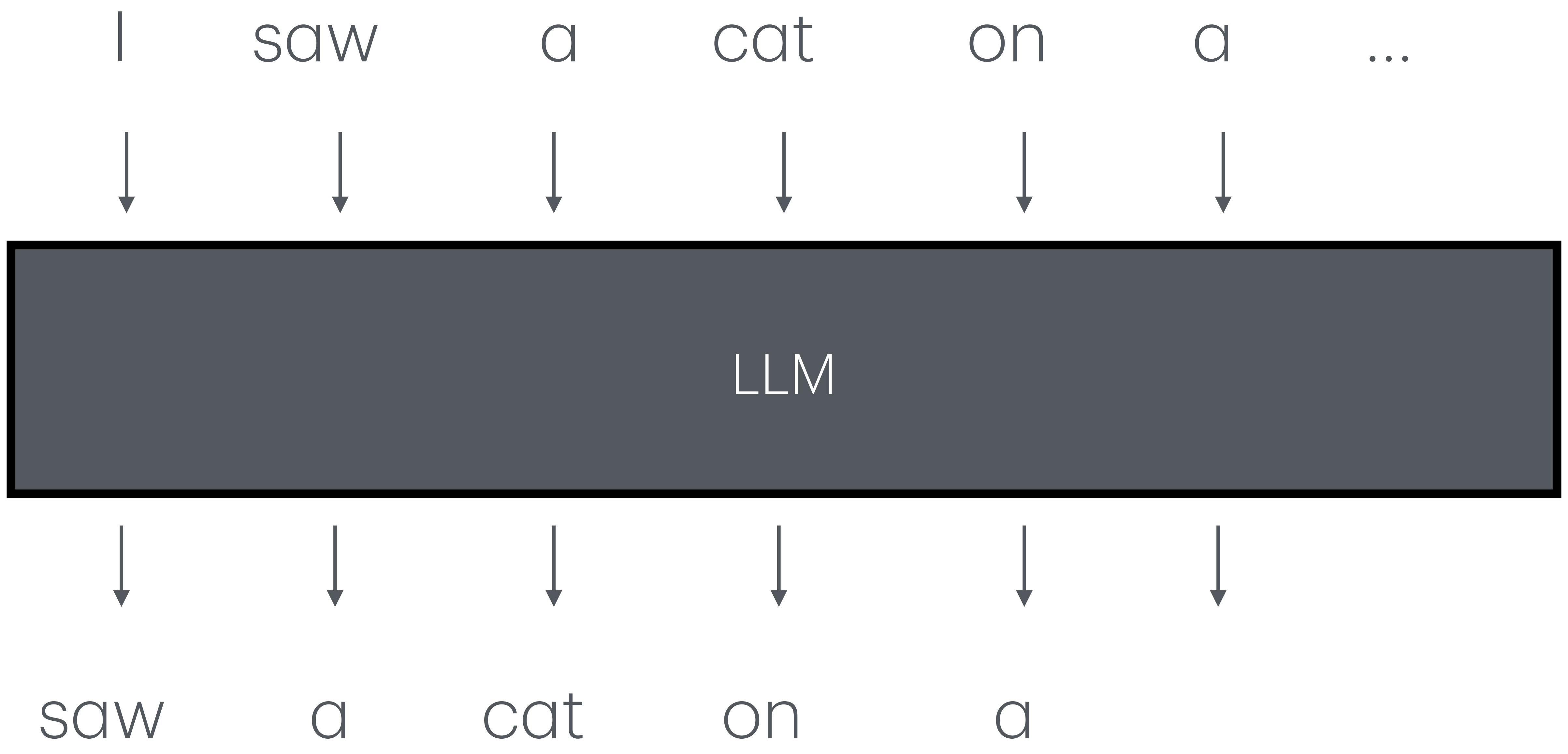


Multi-modality: 🔑 for large action dataset



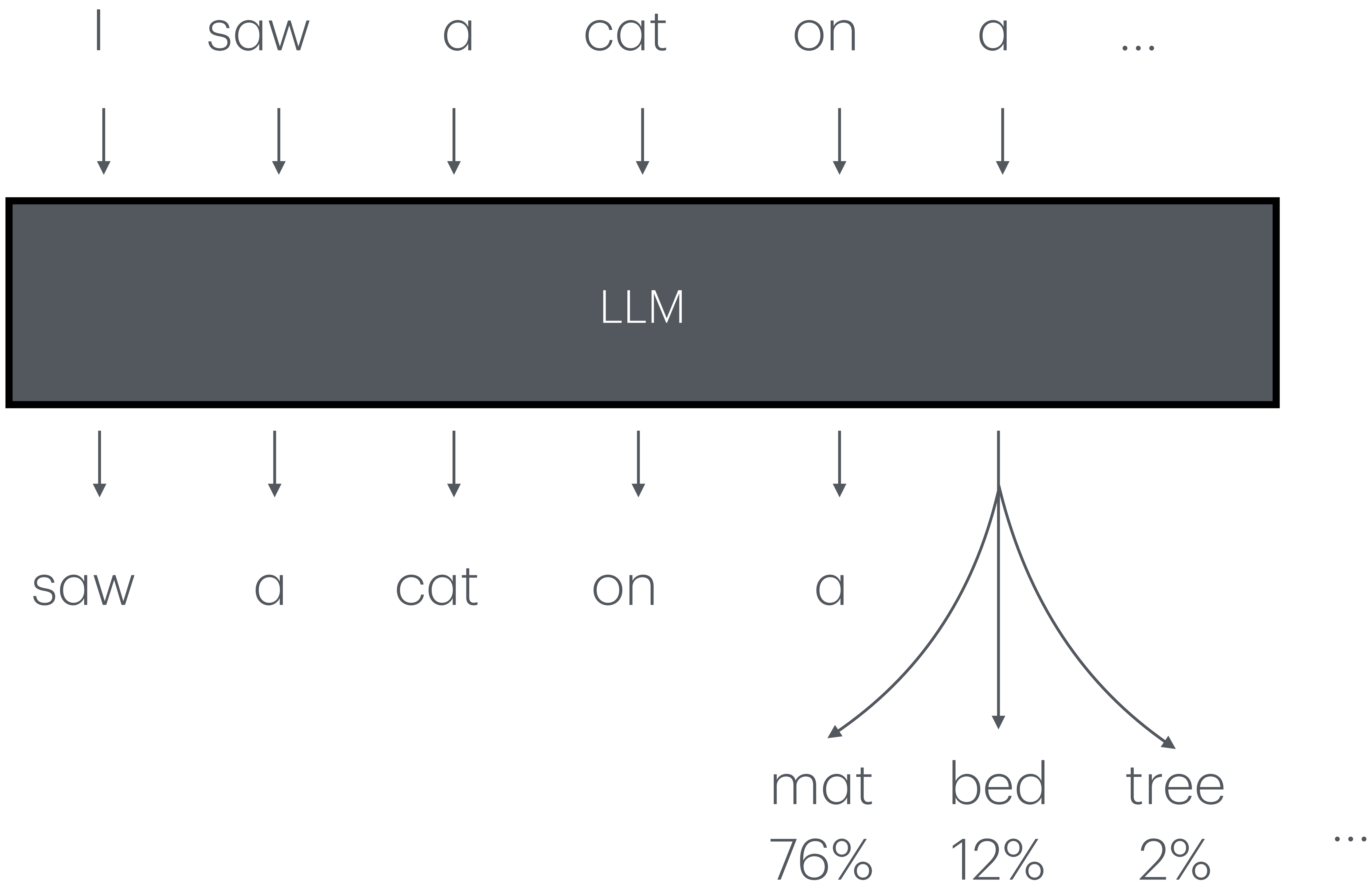
How do language models do it?

Predicting tokenized language, one token at a time

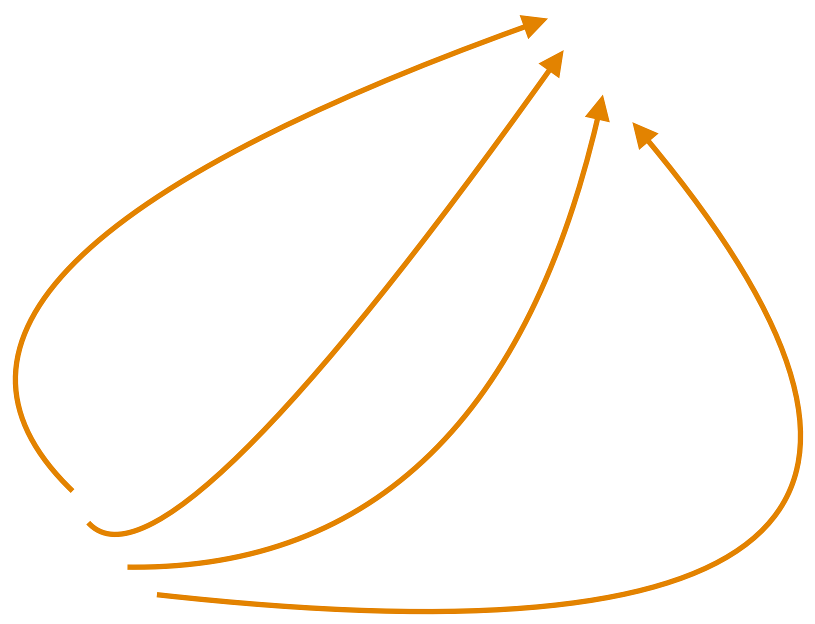


How do language models do it?

Predicting tokenized language, one token at a time



Learning the alphabet of actions

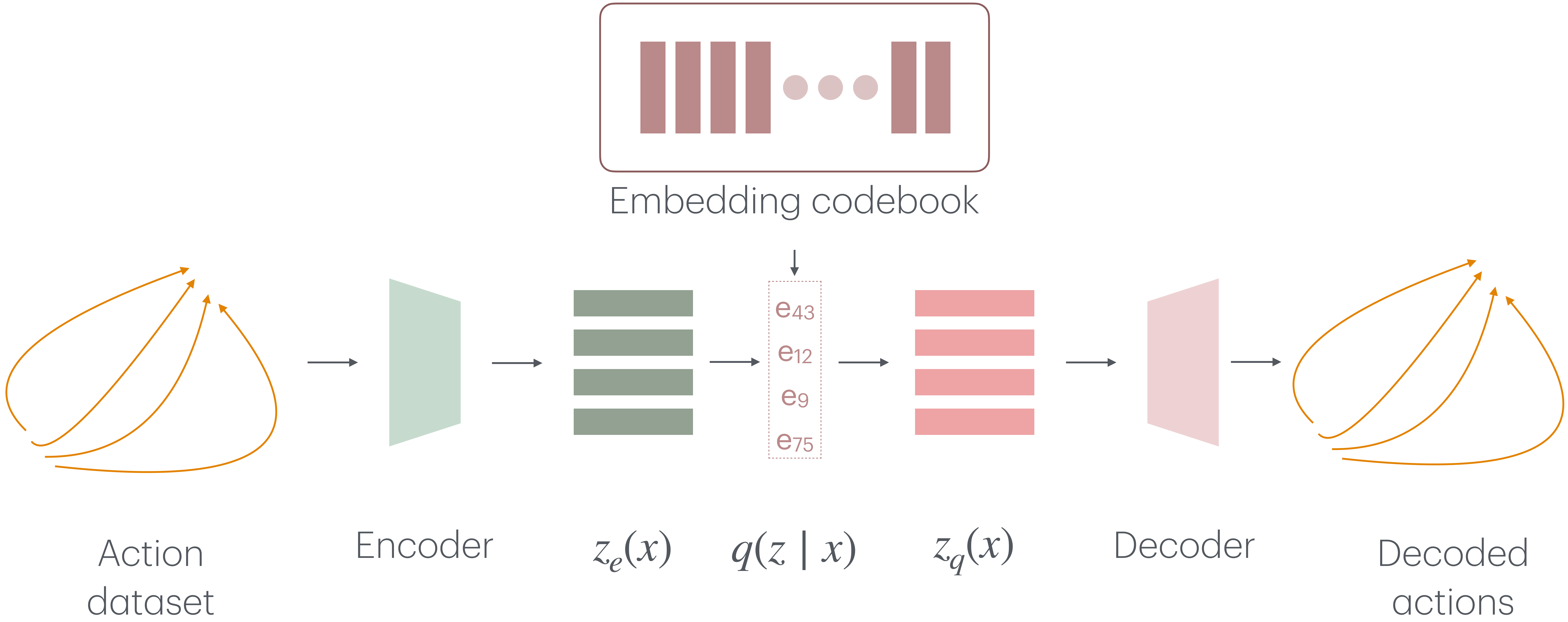


Action dataset

This is continuous

⇒ hard to learn multi-modal distributions over

Learning the alphabet of actions



Embedding codebook

e_{43} e_{12} e_{9} e_{75}

Action dataset

Encoder

$z_e(x)$

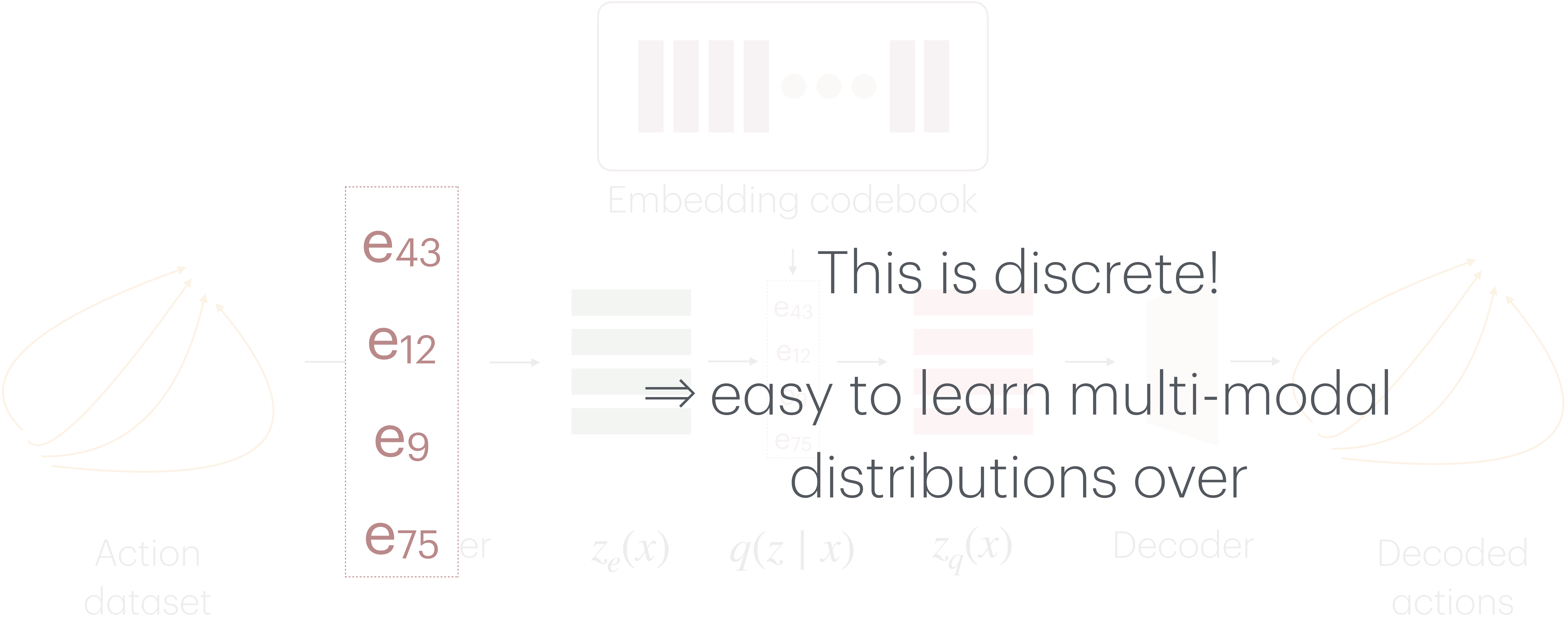
$q(z \mid x)$

$z_q(x)$

Decoder

Decoded actions

Learning the alphabet of actions

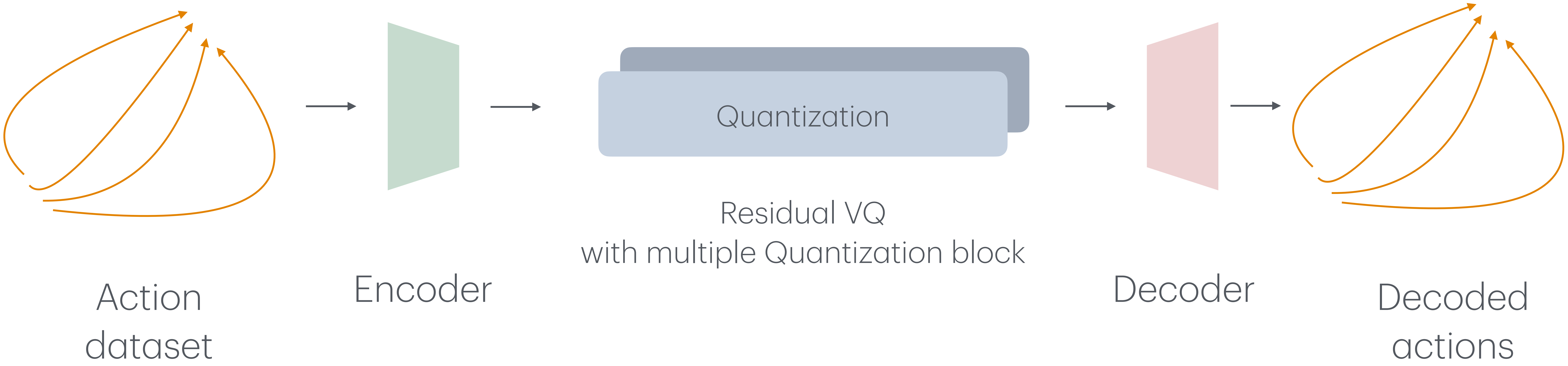


This is discrete!

⇒ easy to learn multi-modal distributions over

Modeling behavior like GPT

Step 1: Tokenizing actions

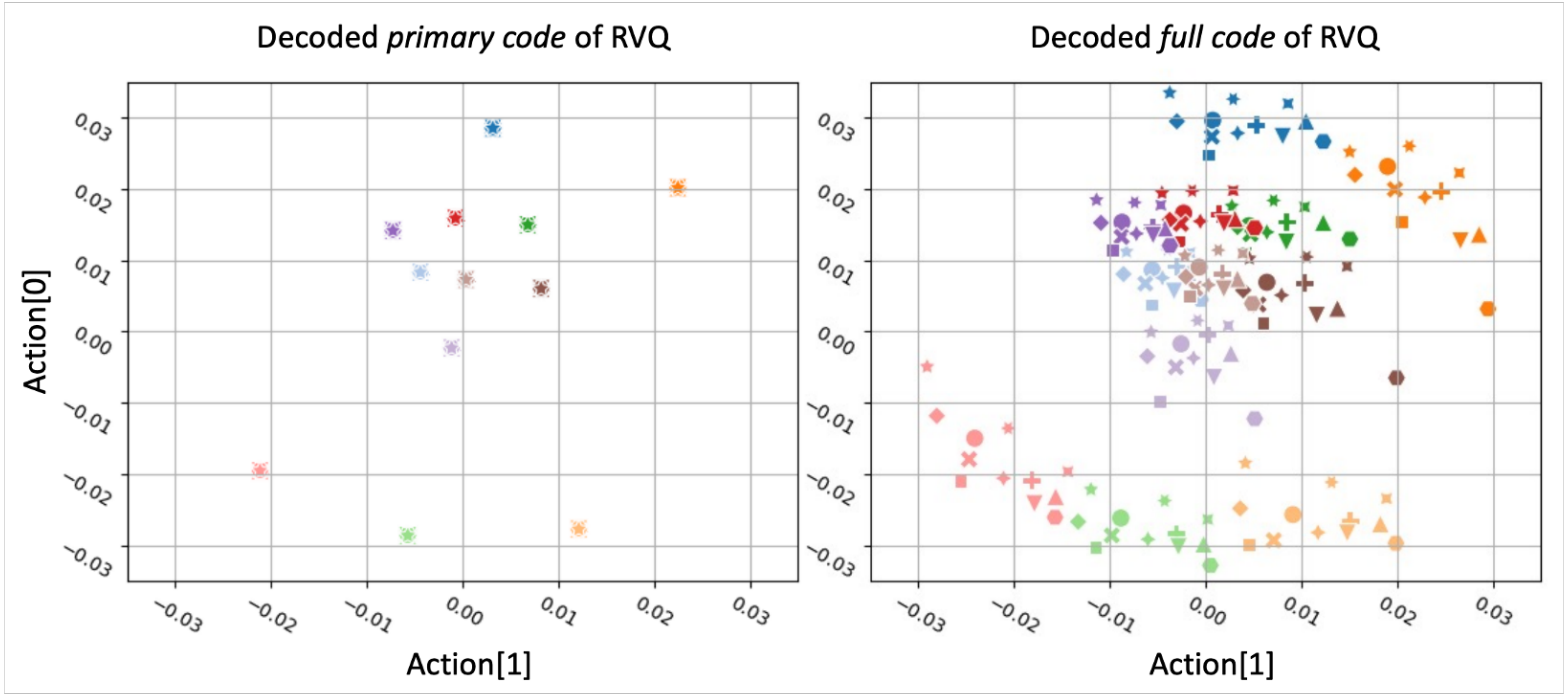


Quantization

Residual VQ
with multiple Quantization block

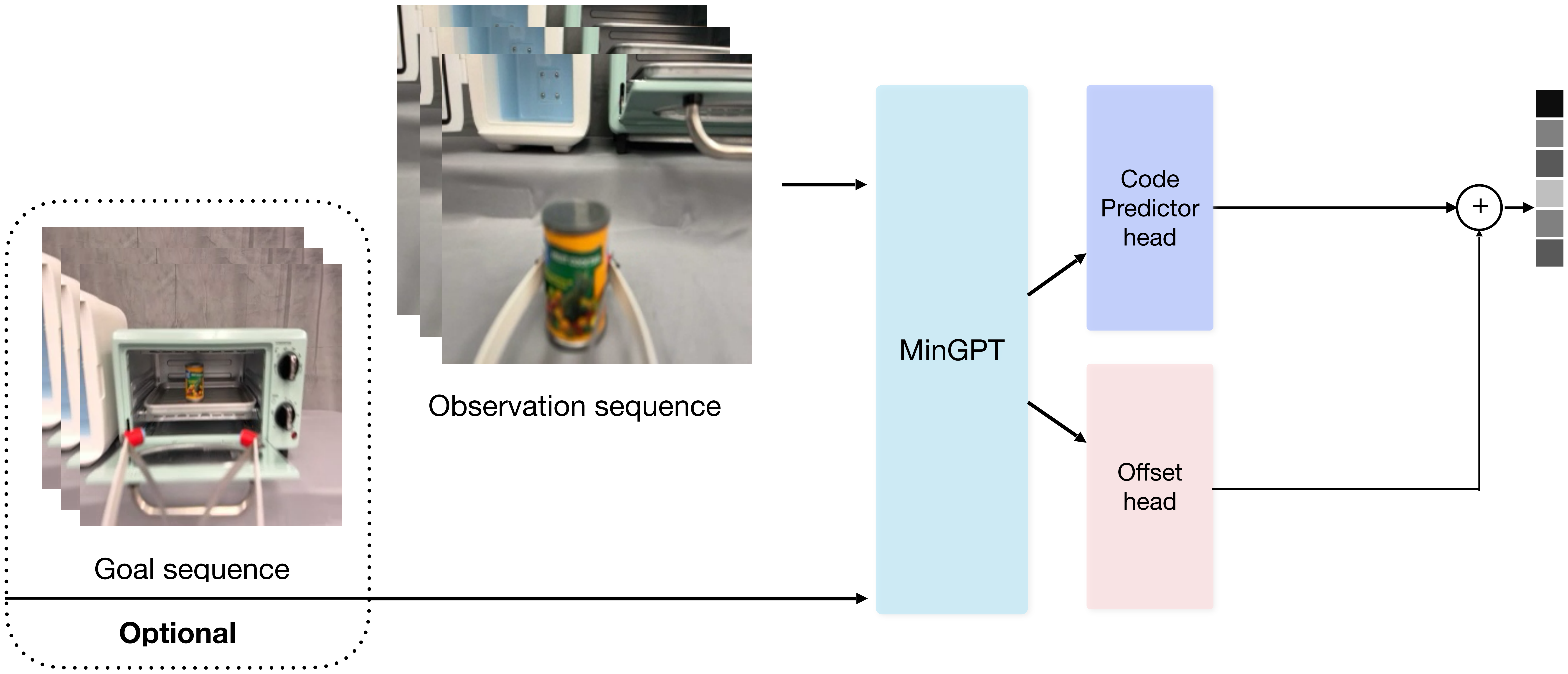
Modeling behavior like GPT

Step 1: Tokenizing actions

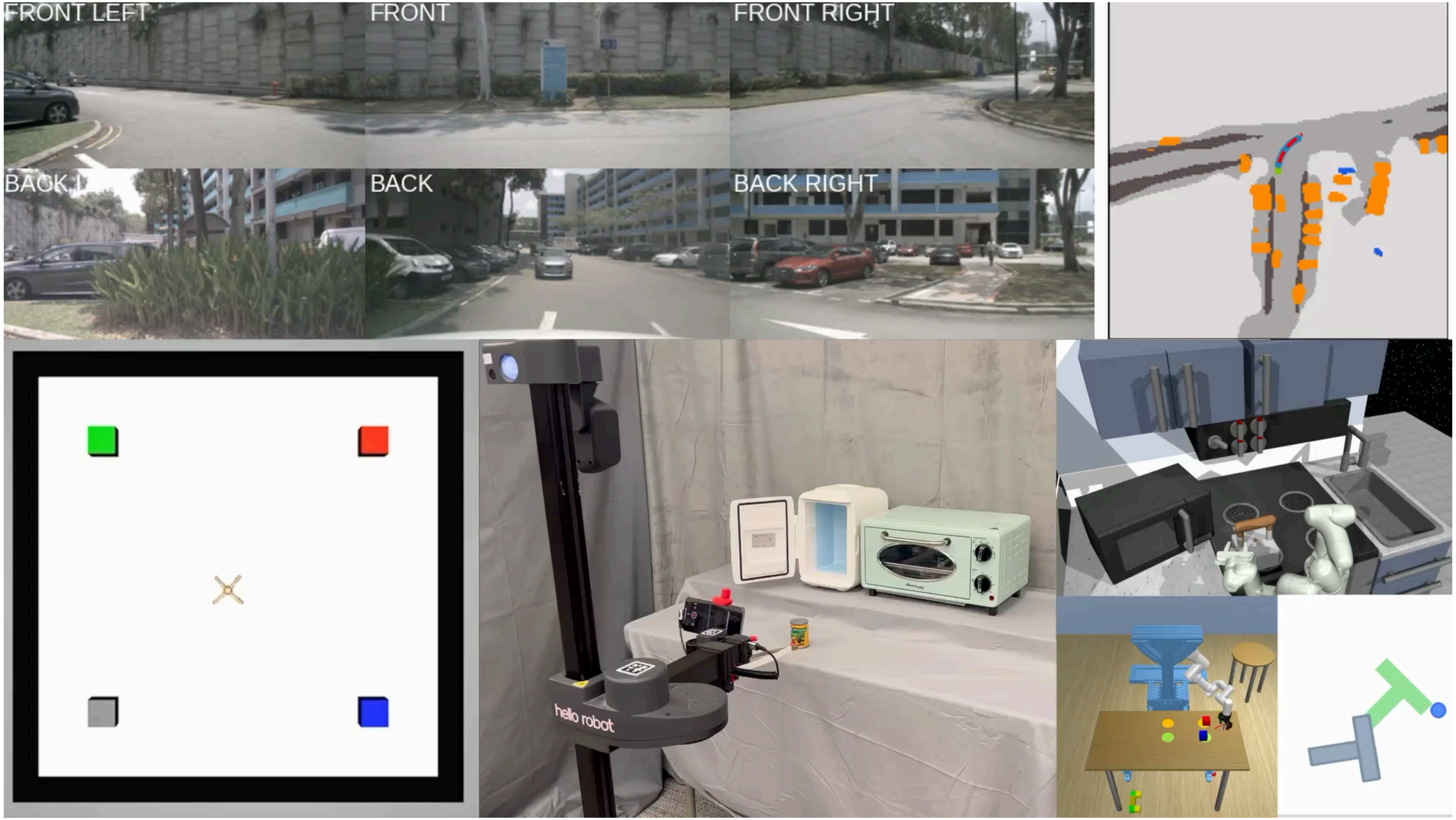


Modeling behavior like GPT

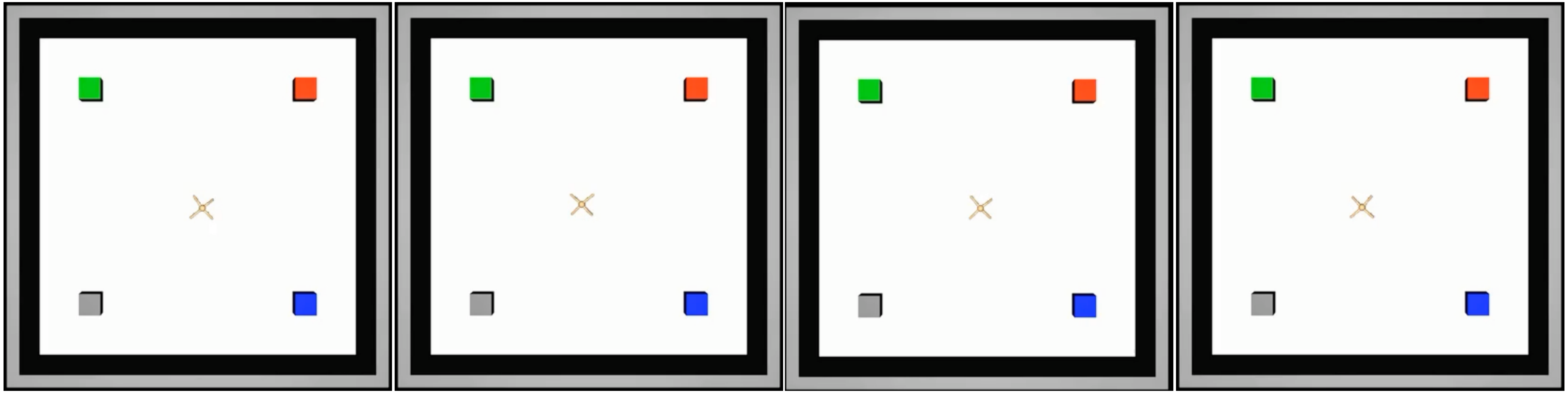
Step 2: Predicting actions using a transformer decoder



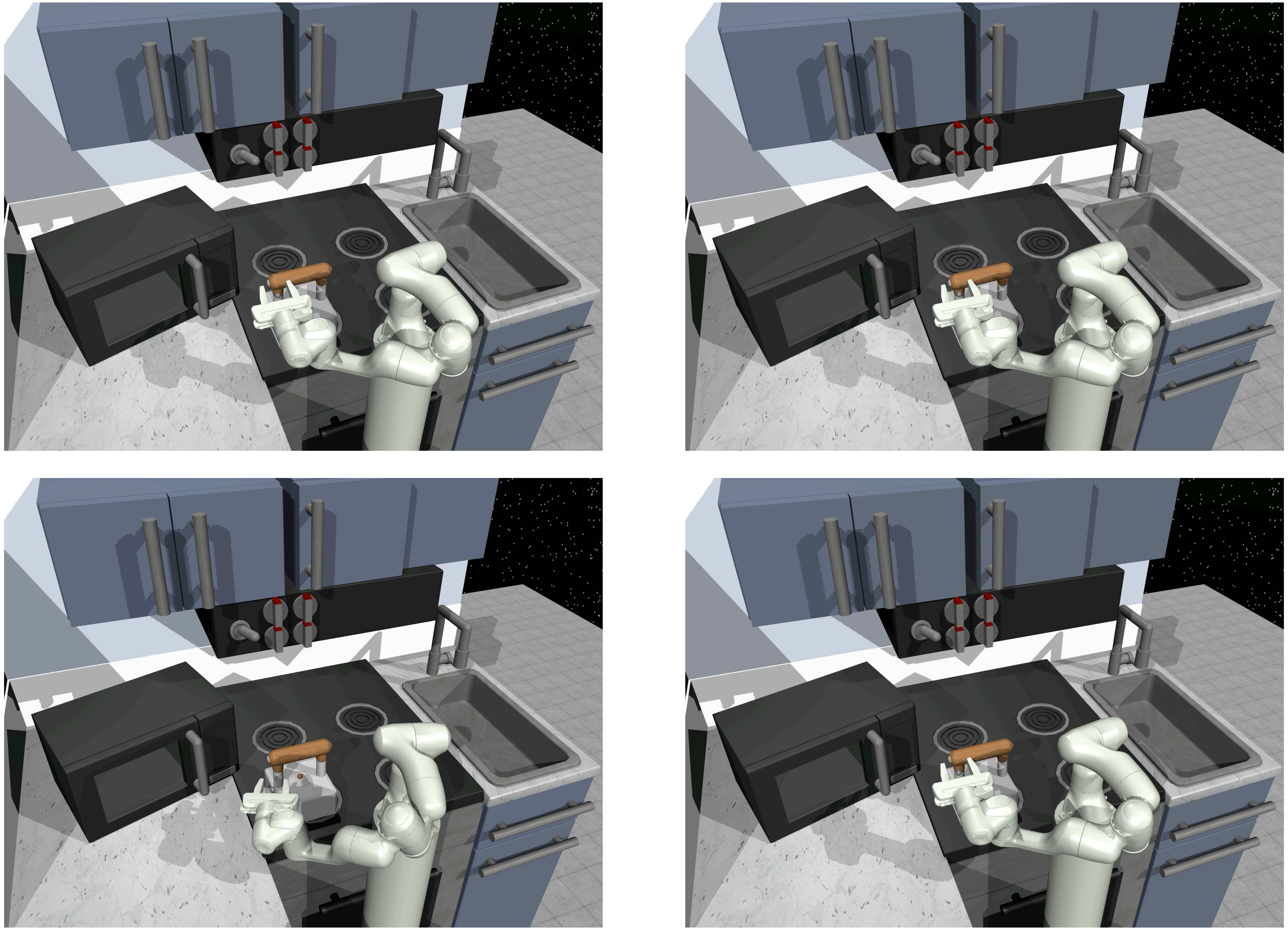
VQ-BeT in Various Decision-making Problems



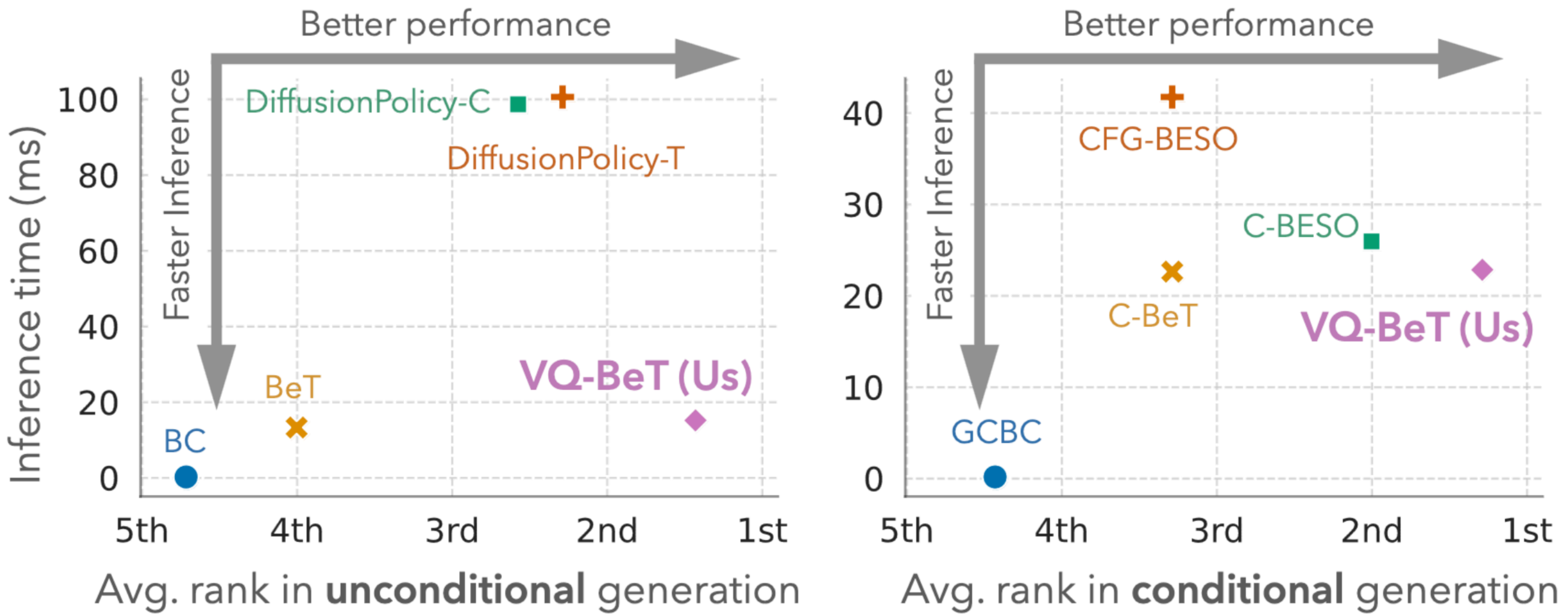
Multimodal Behaviors of VQ-BeT



Multimodal Behaviors of VQ-BeT



Outperforming Diffusion policy



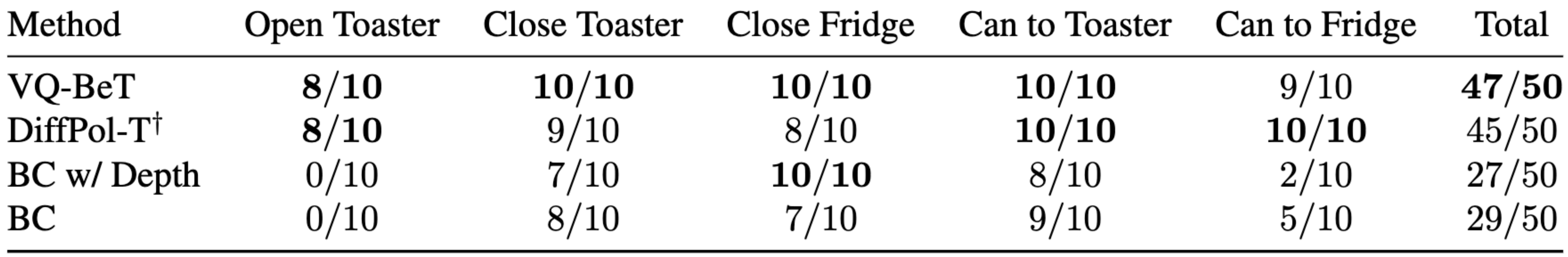
Fast and Light-weighted Model

14-15% of Diffusion Policy!

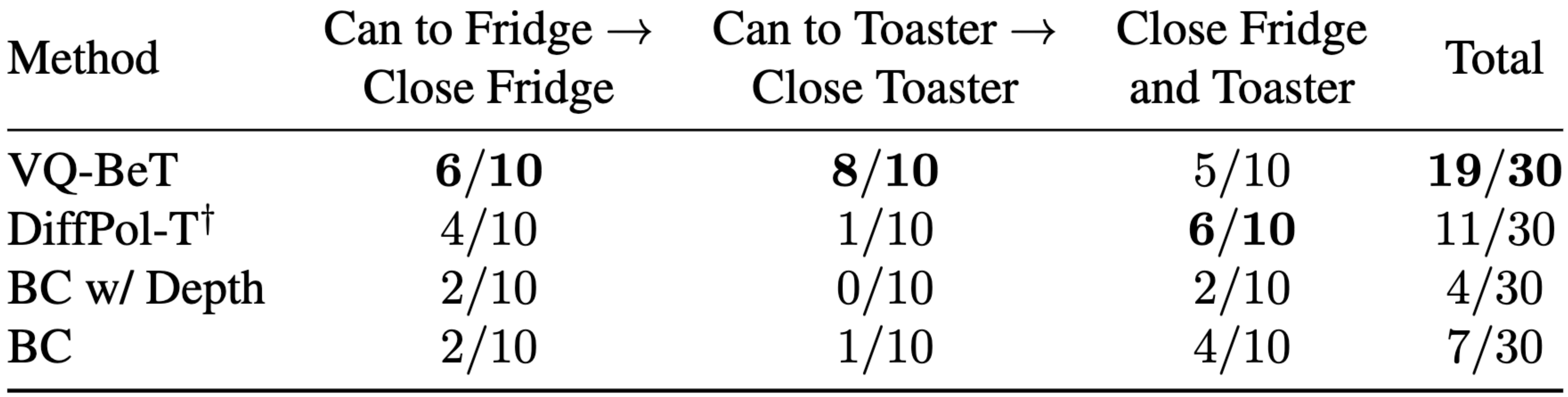
38MB including Image Encoder (HuggingFace LeRobot Implementation)

	On GPU (A6000)	On CPU
Inference time (50 envs batch)	**12ms** for 5-step action chunk	**43ms** for 5-step action chunk
Inference time (per single action)	**2.4ms** for 5-step action chunk	**8ms** for single-step

Real-world Experiments

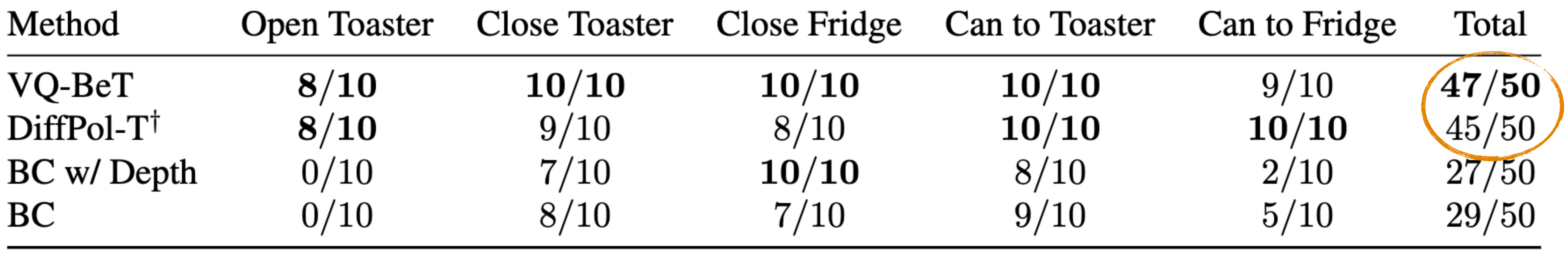


Method	Open Toaster	Close Toaster	Close Fridge	Can to Toaster	Can to Fridge	Total
VQ-BeT	**8/10**	**10/10**	**10/10**	**10/10**	9/10	**47/50**
DiffPol-T†	**8/10**	9/10	8/10	**10/10**	**10/10**	45/50
BC w/ Depth	0/10	7/10	**10/10**	8/10	2/10	27/50
BC	0/10	8/10	7/10	9/10	5/10	29/50

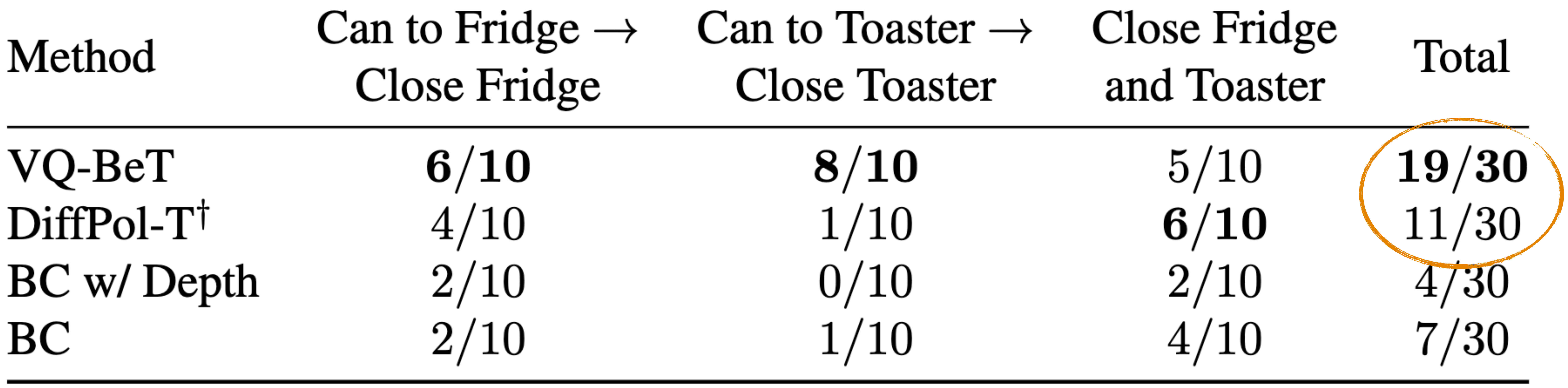


Method	Can to Fridge → Close Fridge	Can to Toaster → Close Toaster	Close Fridge and Toaster	Total
VQ-BeT	**6/10**	**8/10**	5/10	**19/30**
DiffPol-T†	4/10	1/10	**6/10**	11/30
BC w/ Depth	2/10	0/10	2/10	4/30
BC	2/10	1/10	4/10	7/30

Real-world Experiments



Method	Open Toaster	Close Toaster	Close Fridge	Can to Toaster	Can to Fridge	Total
VQ-BeT	**8/10**	**10/10**	**10/10**	**10/10**	9/10	**47/50**
DiffPol-T†	**8/10**	9/10	8/10	**10/10**	**10/10**	45/50
BC w/ Depth	0/10	7/10	**10/10**	8/10	2/10	27/50
BC	0/10	8/10	7/10	9/10	5/10	29/50



Method	Can to Fridge → Close Fridge	Can to Toaster → Close Toaster	Close Fridge and Toaster	Total
VQ-BeT	**6/10**	**8/10**	5/10	**19/30**
DiffPol-T†	4/10	1/10	**6/10**	11/30
BC w/ Depth	2/10	0/10	2/10	4/30
BC	2/10	1/10	4/10	7/30

Especially true for long horizon and low-data regime

Behavior Generation with Latent Actions

VQ-BeT: Action multi-modality through tokenization

Seungjae Lee, Yibin Wang, Haritheja Etukuru, H. Jin Kim,
Nur Muhammad Mahi Shafiullah*, Lerrel Pinto*

https://sjlee.cc/vq-bet