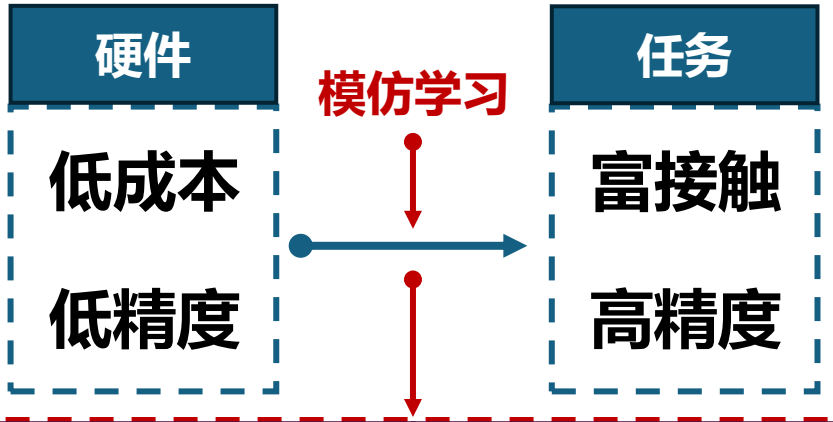
Abstract

Title: Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware

ALOHA：A Low-cost Open-source Hardware System for Bimanual Teleoperation

ACT：Action chunking with Transformers



Motivation

Learning
>
Modelling

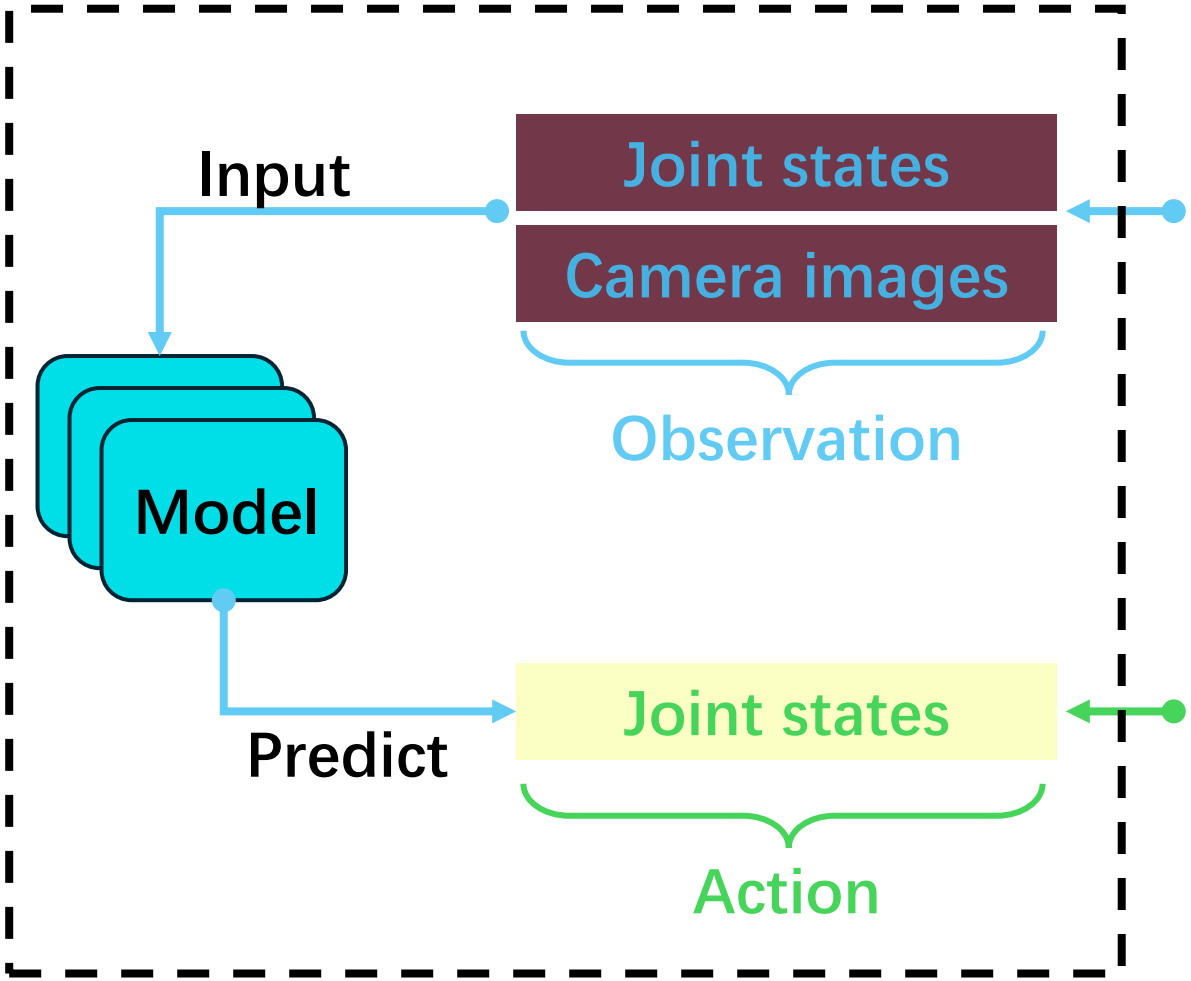
创新点	问题
Action chunking + Transformers II ACT	人类演示非平稳 CVAE 策略误差会放大 Action chunking

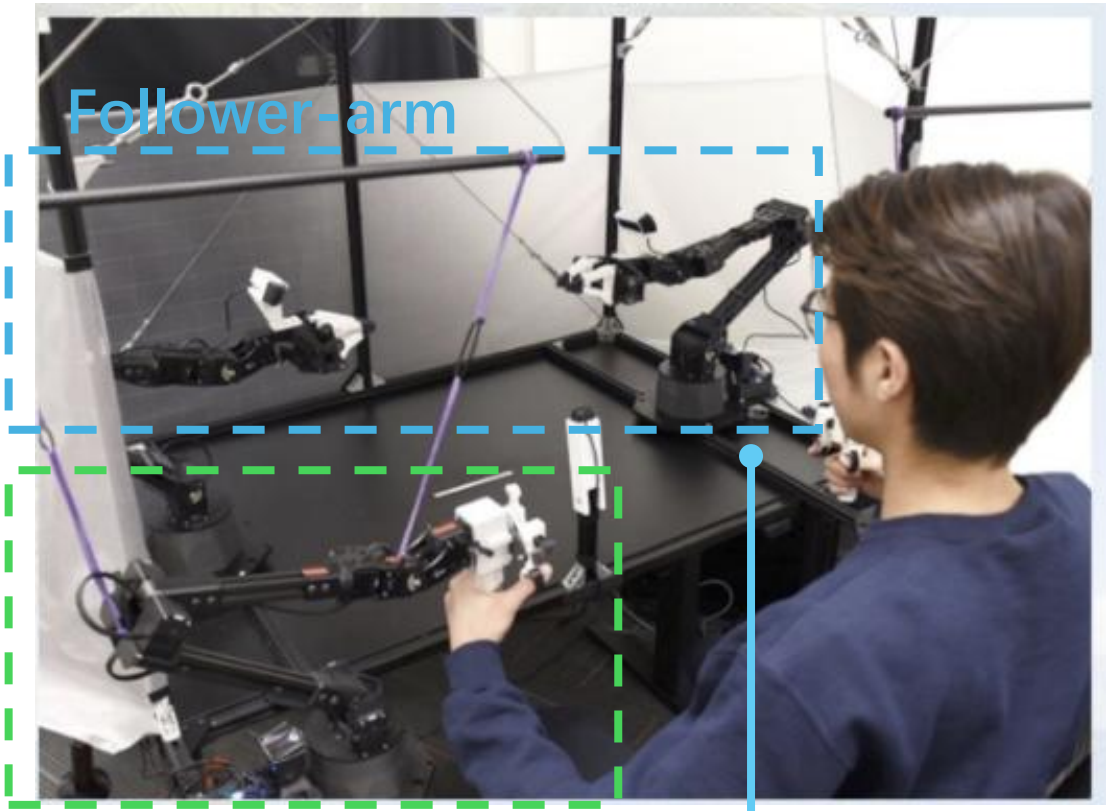
Trick

temporal ensembling

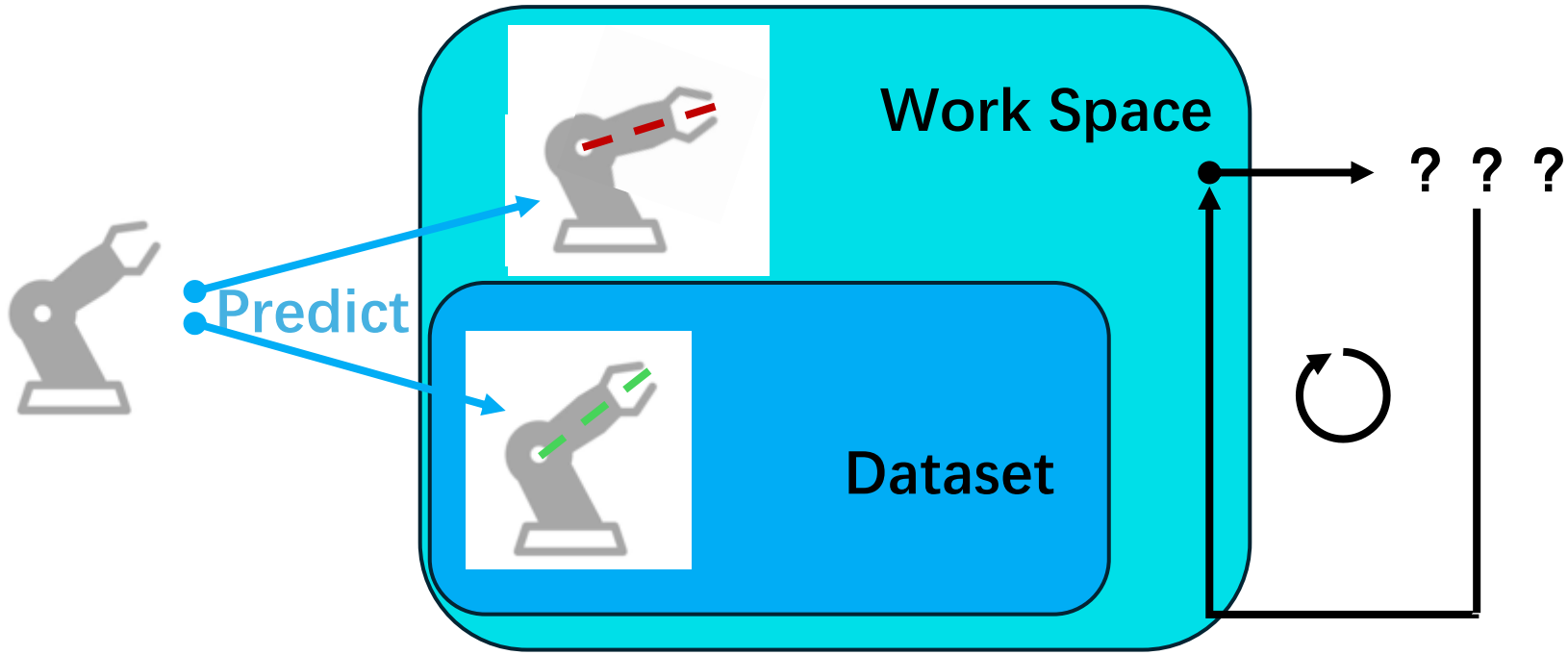
pixel-to-action

Training


Input
Joint states
Camera images
Observation
Model
Predict
Joint states
Action
Frequency: 50 Hz


Follower-arm
Leader-arm
人控制
PID控制电机
跟随leader-arm

Compounding error



Single-step policy $\pi_\theta(a_t \mid s_t)$	当前的action只由当前的states决定	Motivation 单个误差是不致命的，可以由下一次predict修正，高频误差是致命的
Action-chunking policy $\pi_\theta(a_{t:t+k} \mid s_t)$	一个states决定接下来k步的action	

Why action chunking?

1 Reduce effective horizon of Long trajectory

2 Model temporal reliance

假设任务：“把油放到锅里，等三秒……” ⟶ **单步策略无法完成**

Single-step policy

$$\pi_\theta(a_t|s_t)$$

Action-chunking policy

$$\pi_\theta(a_{t:t+k}|s_t)$$

History-conditioned policy

$$\pi_\theta(a_t|s_t, s_{t-1}, s_{t-2}, \dots)$$

Casual Misidentification

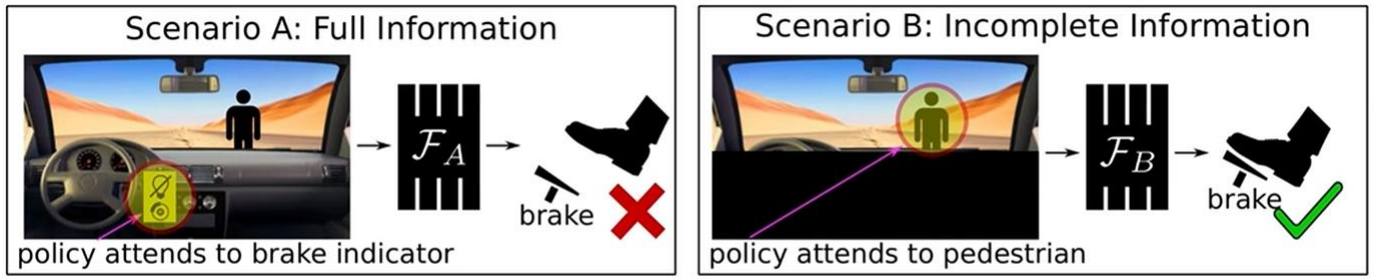
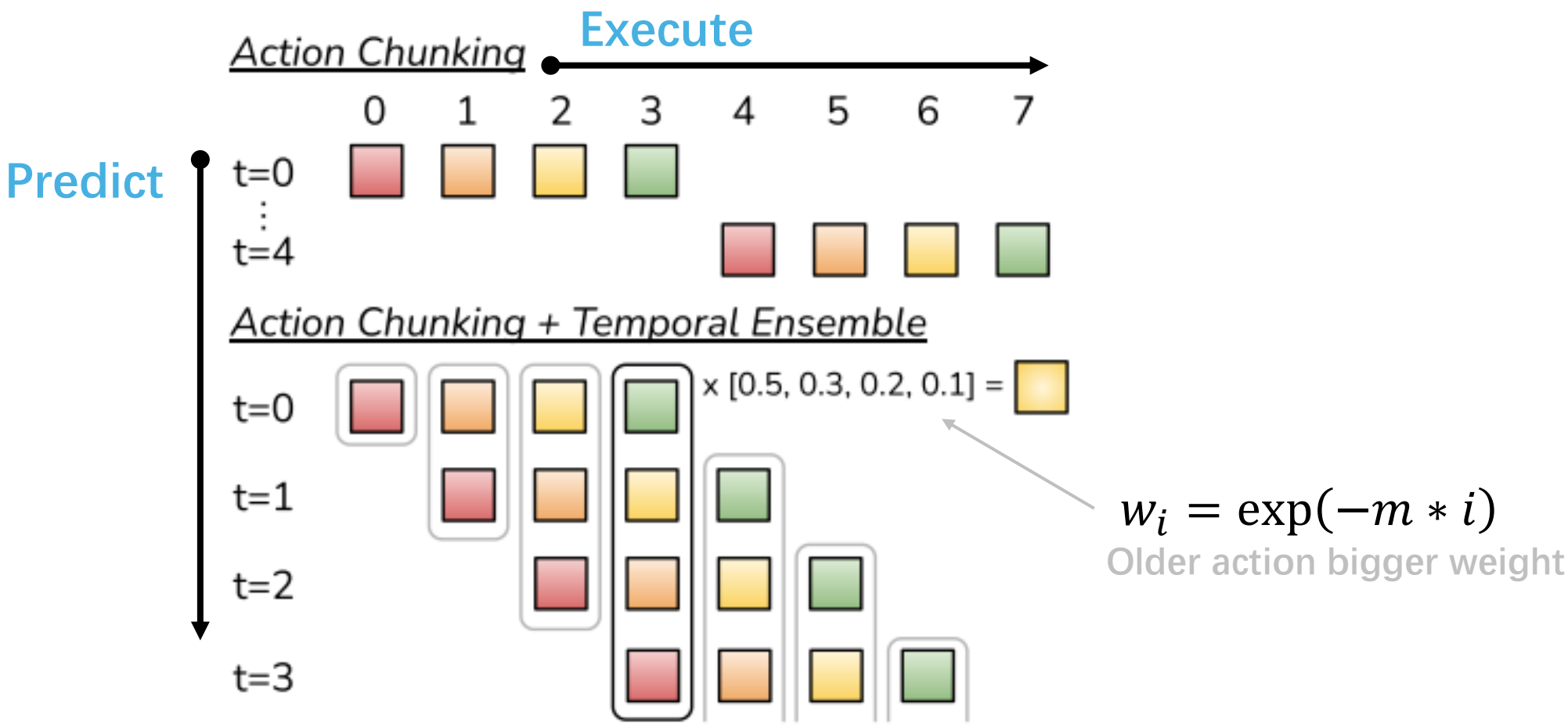


Figure 1: Causal misidentification: *more* information yields worse imitation learning performance. Model A relies on the braking indicator to decide whether to brake. Model B instead correctly attends to the pedestrian.

Trick

Temporal ensembling



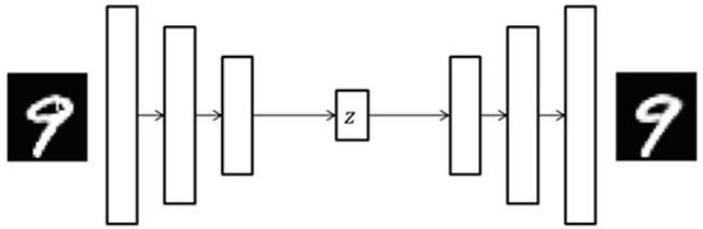
$$w_i = \exp(-m * i)$$

Older action bigger weight

Model

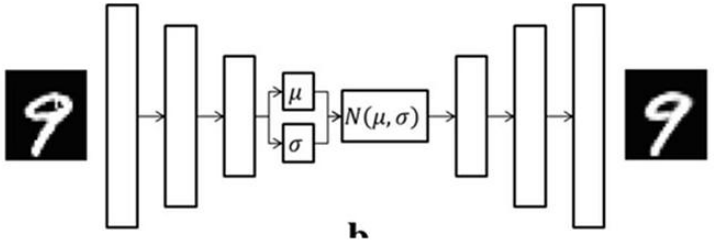
CVAE 简介

Auto-Encoder (AE)



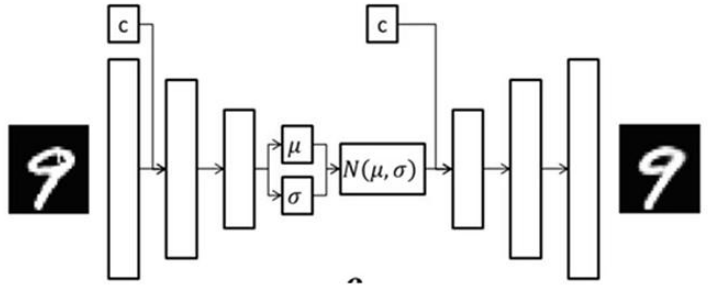
Compress data using middle representation z

Variational Auto-Encoder (VAE)



Add probabilistic to improve generative capabilities

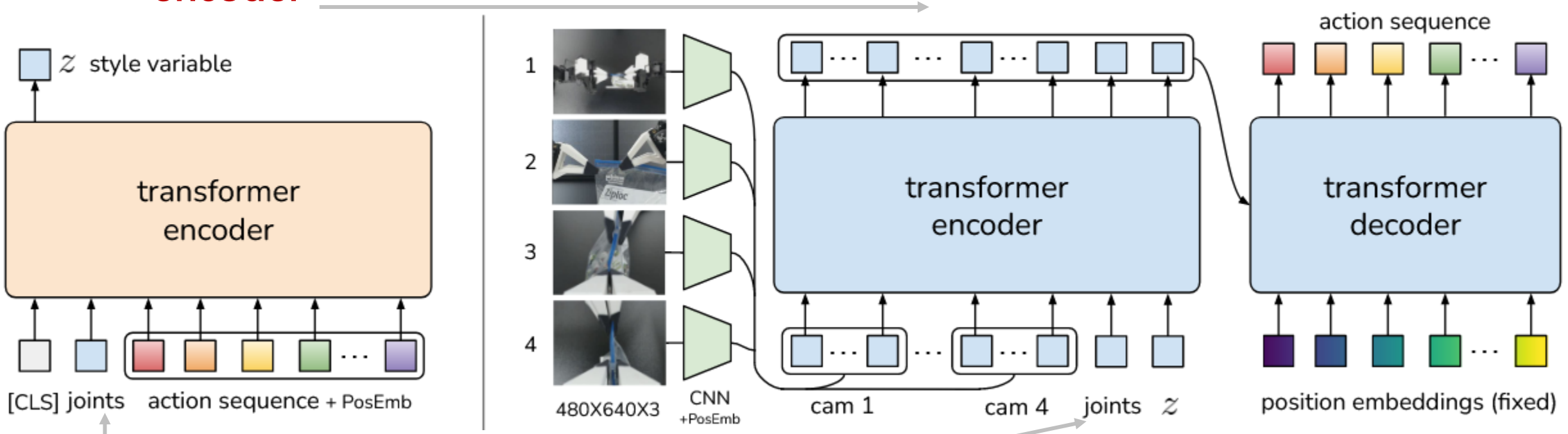
Conditional Variational Auto-Encoder (CVAE)



Add conditioning information to guide the generative process

Model (train)

encoder 仅为了训练decoder，test时丢弃 decoder

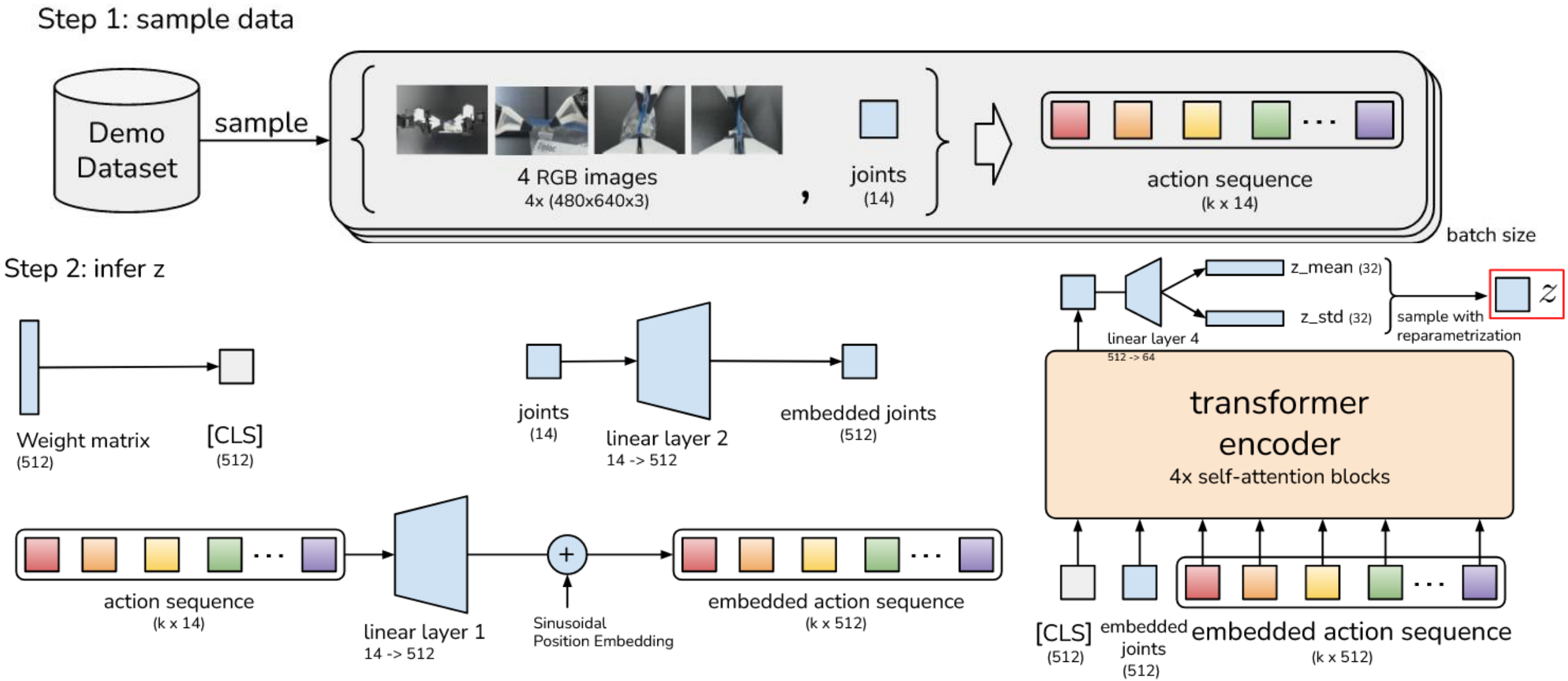


without image

Just for faster training

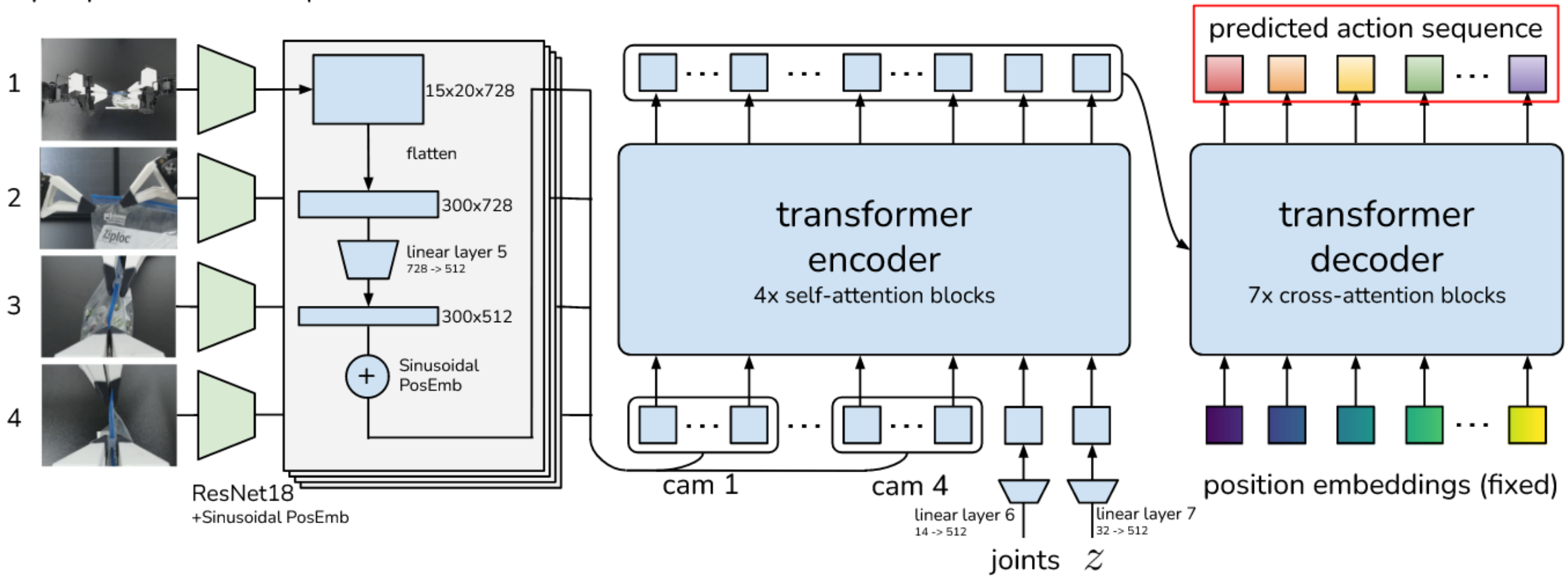
Loss Function

Model (train)


Step 1: sample data
Demo Dataset
sample
4 RGB images
4x (480x640x3)
joints
(14)
action sequence
(k x 14)
batch size
Step 2: infer z
Weight matrix
(512)
[CLS]
(512)
joints
(14)
linear layer 2
14 -> 512
embedded joints
(512)
z_mean (32)
z_std (32)
sample with reparametrization
z
linear layer 4
512 -> 64
transformer encoder
4x self-attention blocks
action sequence
(k x 14)
linear layer 1
14 -> 512
Sinusoidal Position Embedding
embedded action sequence
(k x 512)
[CLS]
(512)
embedded joints
(512)
embedded action sequence
(k x 512)

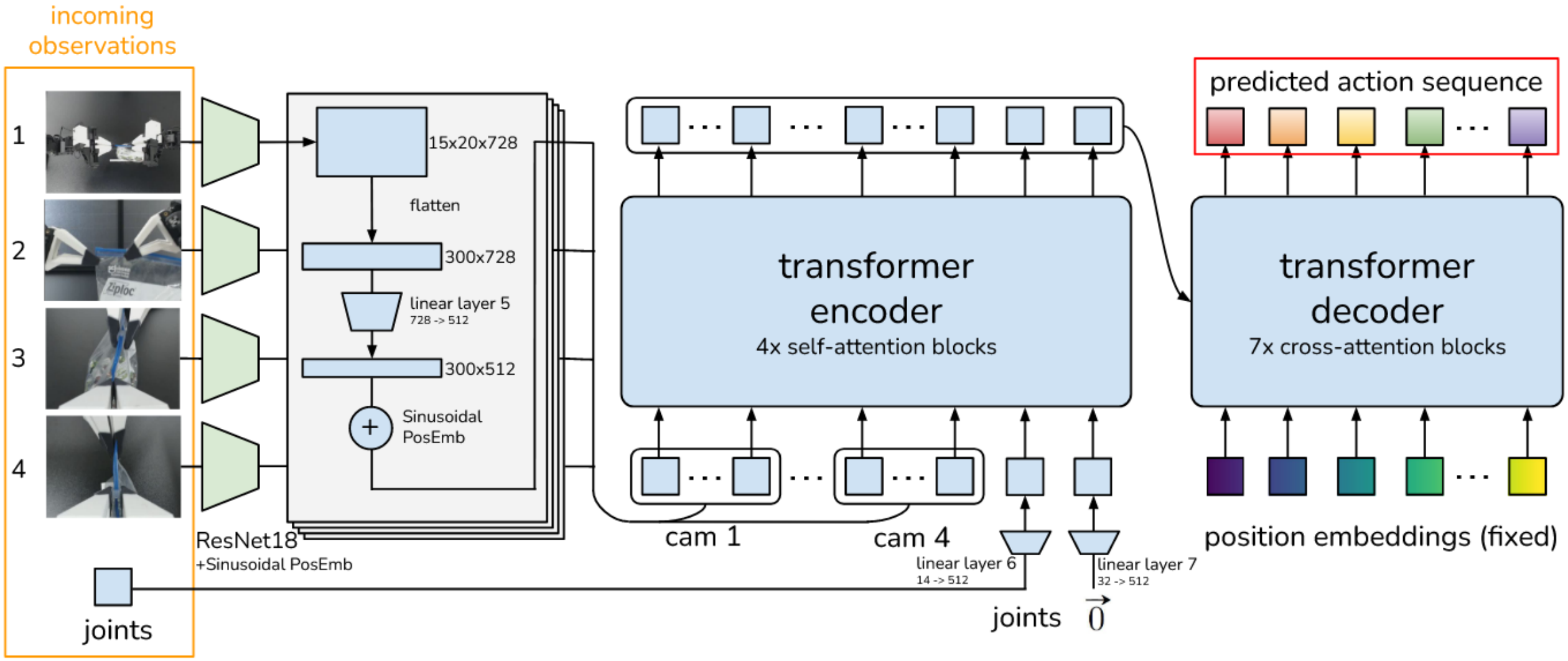
Model (train)

Step 3: predict action sequence

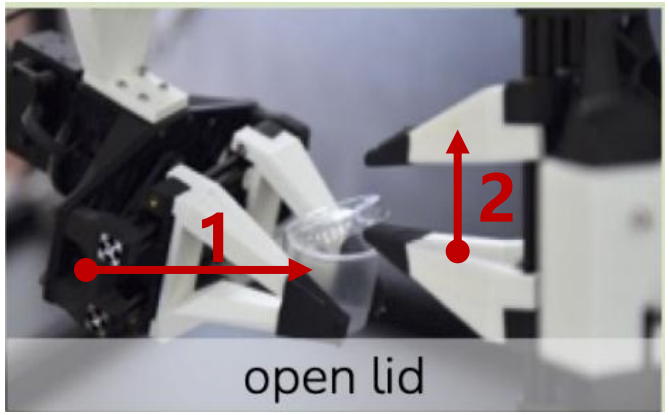


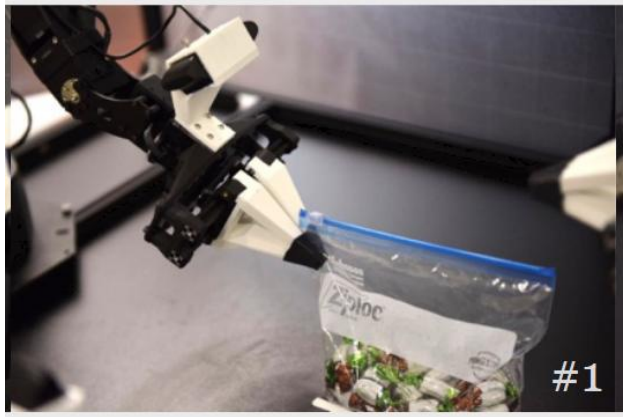
Model (test)

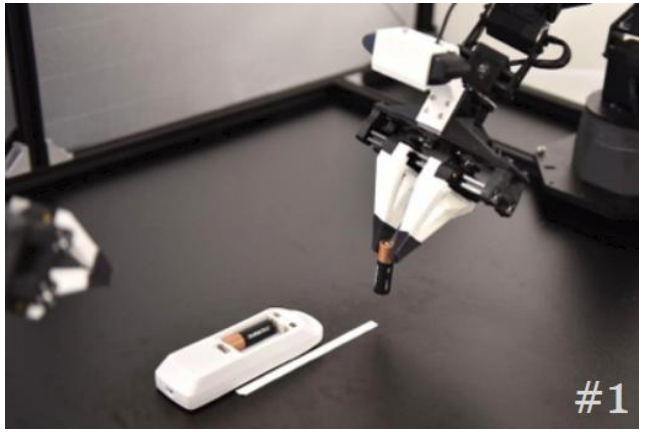
Testing

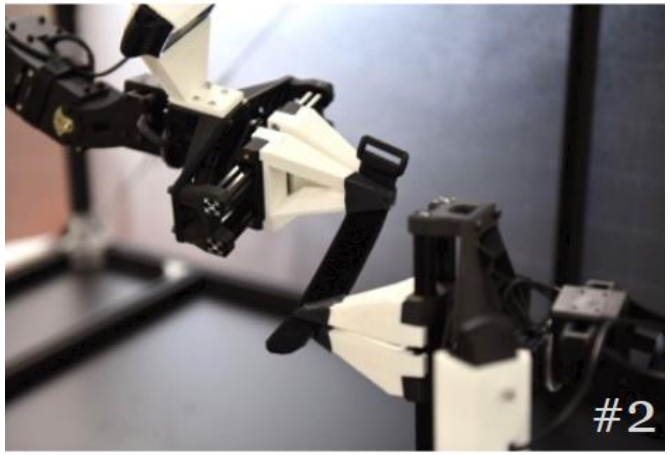


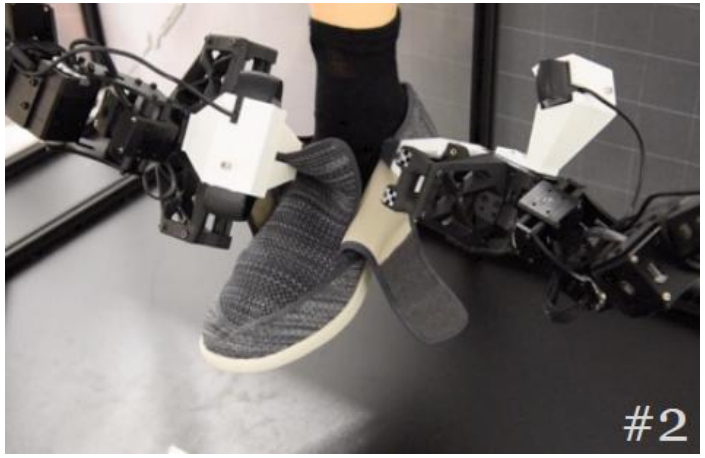
experiment











演示尽量随机，时长在10~20min

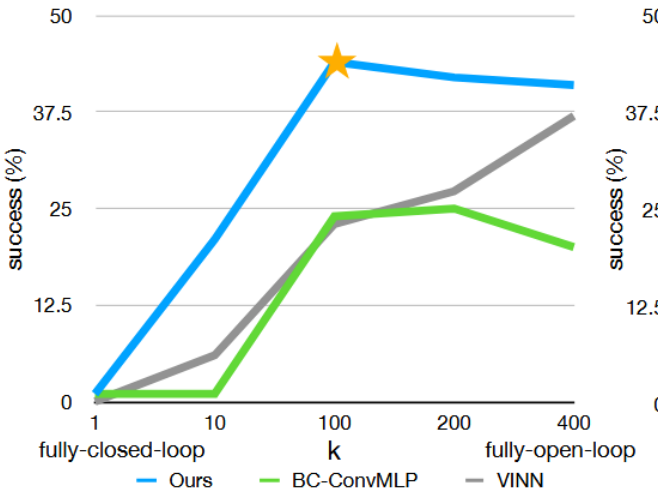
experiment

	Cube Transfer (sim)			**Bimanual Insertion** (sim)			**Slide Ziploc** (real)			**Slot Battery** (real)		
	Touched	Lifted	Transfer	Grasp	Contact	Insert	Grasp	Pinch	Open	Grasp	Place	Insert
BC-ConvMLP	34 | 3	17 | 1	1 | 0	5 | 0	1 | 0	1 | 0	0	0	0	0	0	0
BeT	60 | 16	51 | 13	27 | 1	21 | 0	4 | 0	3 | 0	8	0	0	4	0	0
RT-1	44 | 4	33 | 2	2 | 0	2 | 0	0 | 0	1 | 0	4	0	0	4	0	0
VINN	13 | 17	9 | 11	3 | 0	6 | 0	1 | 0	1 | 0	28	0	0	20	0	0
ACT (Ours)	**97 | 82**	**90 | 60**	**86 | 50**	**93 | 76**	**90 | 66**	**32 | 20**	**92**	**96**	**88**	**100**	**100**	**96**

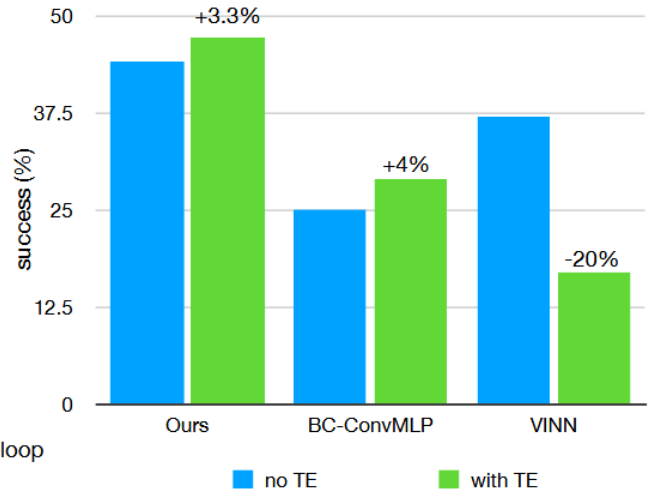
	Open Cup (real)			**Thread Velcro** (real)			**Prep Tape** (real)				**Put On Shoe** (real)			
	Tip Over	Grasp	Open Lid	Lift	Grasp	Insert	Grasp	Cut	Handover	Hang	Lift	Insert	Support	Secure
BeT	12	0	0	24	0	0	8	0	0	0	12	0	0	0
ACT (Ours)	**100**	**96**	**84**	**92**	**40**	**20**	**96**	**92**	**72**	**64**	**100**	**92**	**92**	**92**

感知系统不精确，带子体积小
黑色和桌面黑色对比度差

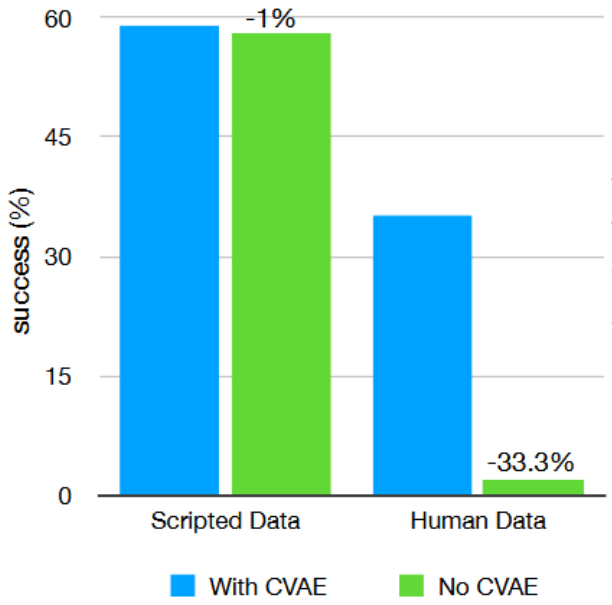
ablation



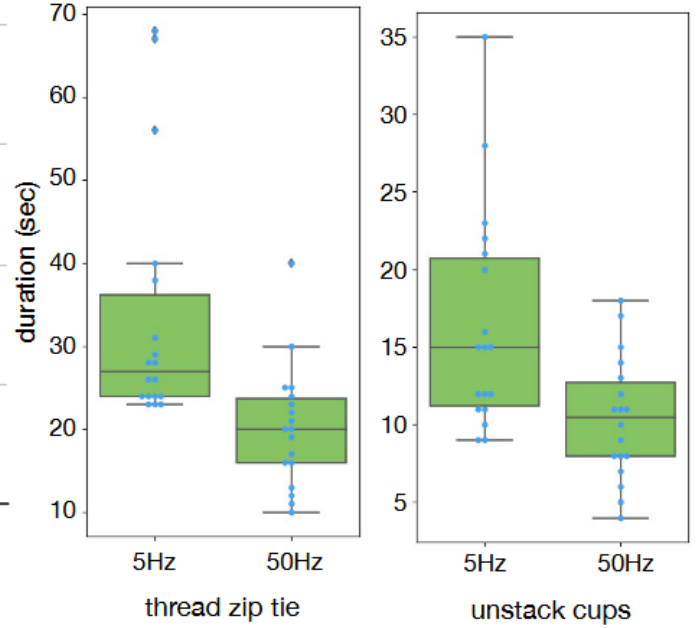
(a)



(b)



(c)



(d)

experiment

Why this for Generative model ?

Another challenge that arises is learning from noisy human demonstrations. Given the same observation, a human can use different trajectories to solve the task. Humans will also be more stochastic in regions where precision matters less [38]. Thus, it is important for the policy to focus on regions where high precision matters.

What is this mean?

Both BeT and RT-1 discretize the action space: the output is a categorical distribution over discrete bins, but with an added continuous offset from the bincenter in the case of BeT

Why set zero when testing?

At test time, we set z to be the mean of the prior distribution i.e. zero to deterministically decode.

What is wall-clock time?

experiment

What is the difference between VINN and the other models?

We notice a performance drop for VINN, a non-parametric method. We hypothesize that a temporal ensemble mostly benefits parametric methods by smoothing out the modeling errors. In contrast, VINN retrieves ground-truth actions from the dataset and does not suffer from this issue.

What is the position embedding?

What is the transformer block like?