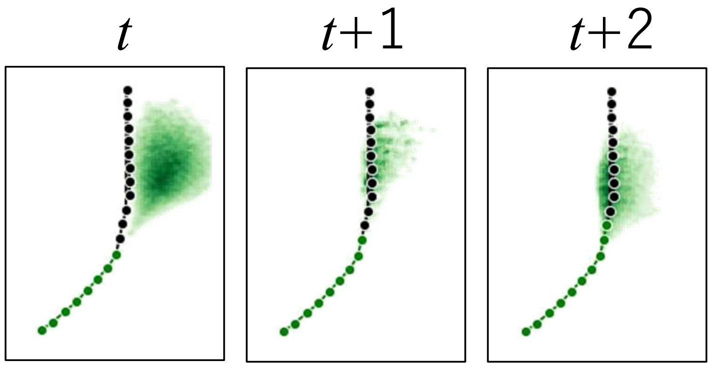
# Human-like Stochastic Motion Generation and Prediction

Norimichi Ukita
Toyota Technological Institute, Japan

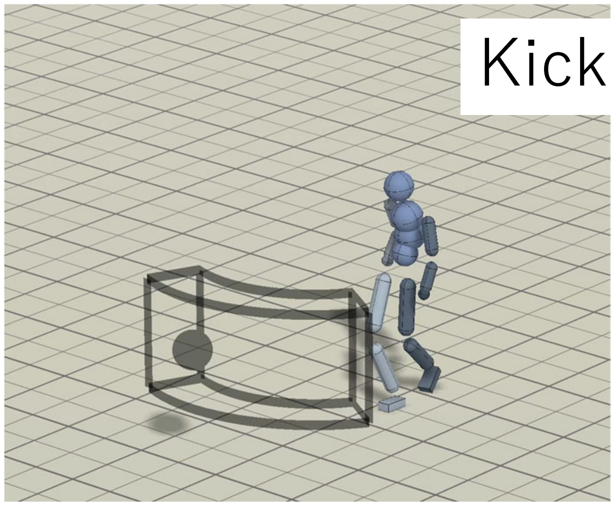
# Motivation – Why stochastic motions? –

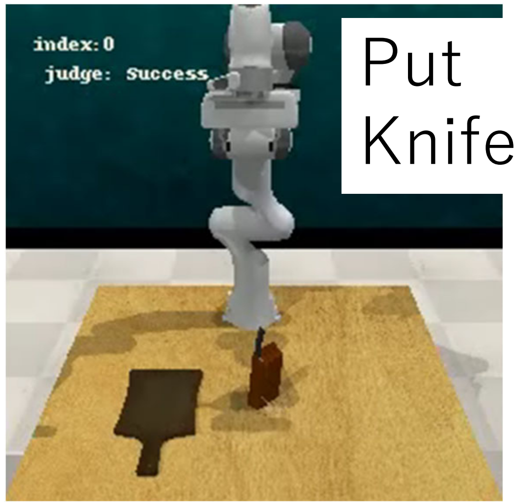
- Human motions are high-dimensional, complex, diverse, and stochastic.
- Deterministic models are not appropriate for representing such complex and stochastic motions.

Predictions at




Walking trajectory



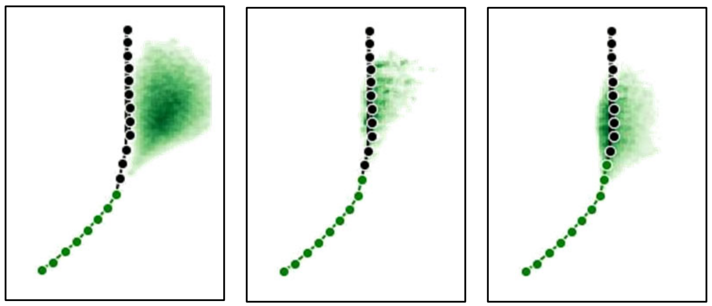



# Today’s Topics

1. **<u>Super-fast</u>** task-agnostic probabilistic prediction
2. **<u>Physically-constrained</u>** human motion generation
3. **<u>Task-achievable</u>** robot motion planning by refining retrieved motions

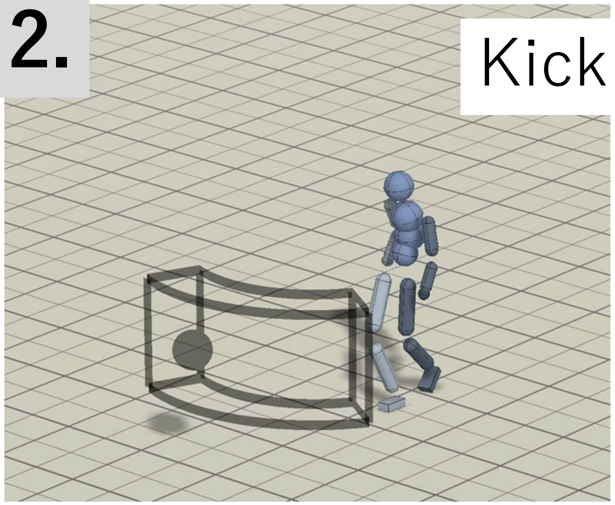
**1.** Predictions at

$t$ $t+1$ $t+2$

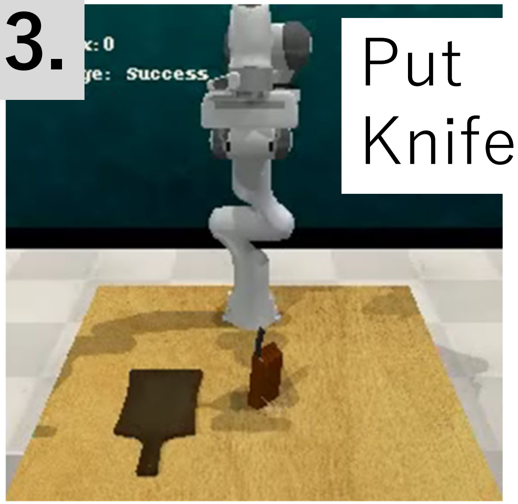


Walking trajectory

**2.**



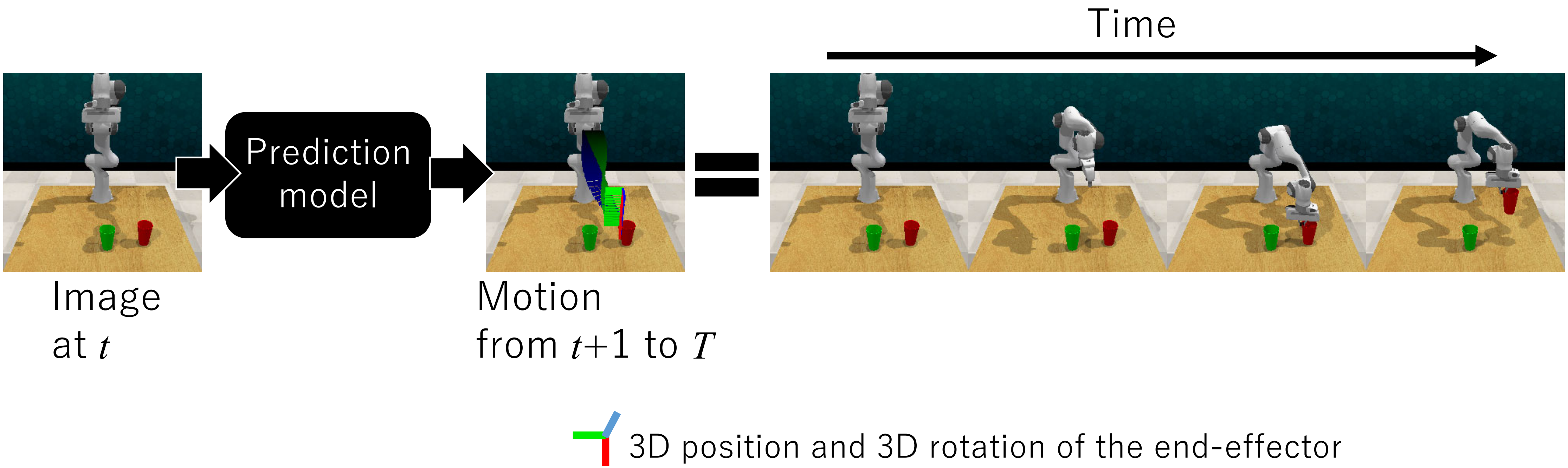

**3.**



# 3. Task-achievable Robot Motion Planning by Refining Retrieved Motions

**Takeru Oba and Norimichi Ukita**

# Task: Motion Prediction from an Image



# Difficulty in Robot Motion Learning

- Stochasticity
  - Not only one but also several motions can achieve each task.
- Controllability
  - Complexity in articulated joint control
    - Similar motions can or cannot be achieved due to the limited range of joint motions.
- High precision
  - Small motion difference may disturb a task.
- Small number of training samples
  - Image generation >> Real robot motion planning
    - Robot motions are collected by manually controlling robots.

# Difficulty in Robot Motion Learning

- Stochasticity
  - Probabilistic models
    - Representing multiple task-achievable motions
- Controllability
  - Retrieval-based motion planning
    - Motion optimization/refinement from real controllable motions
- High precision
  - High-fidelity motion refinement
    - Refinement in a high-resolution refinement space
- Small number of training samples
  - Generative models
    - Successful in-distribution sampling from a limited number of samples

TOYOTA TECHNOLOGICAL INSTITUTE 豊田工大

Intelligent
Information
Media

PDF Code

# Data-Driven Stochastic Motion Evaluation and Optimization with Image by Spatially-Aligned Temporal Encoding

Takeru Oba and Norimichi Ukita
ICRA2023

# Our Solutions for Robot Motion Learning

- Probabilistic models
  - Representing multiple task-achievable motions

➡ Energy-Based Models (EBM)

- Retrieval-based motion planning
  - Motion optimization/refinement from real controllable motions

➡ Refining real samples in a supervised manner

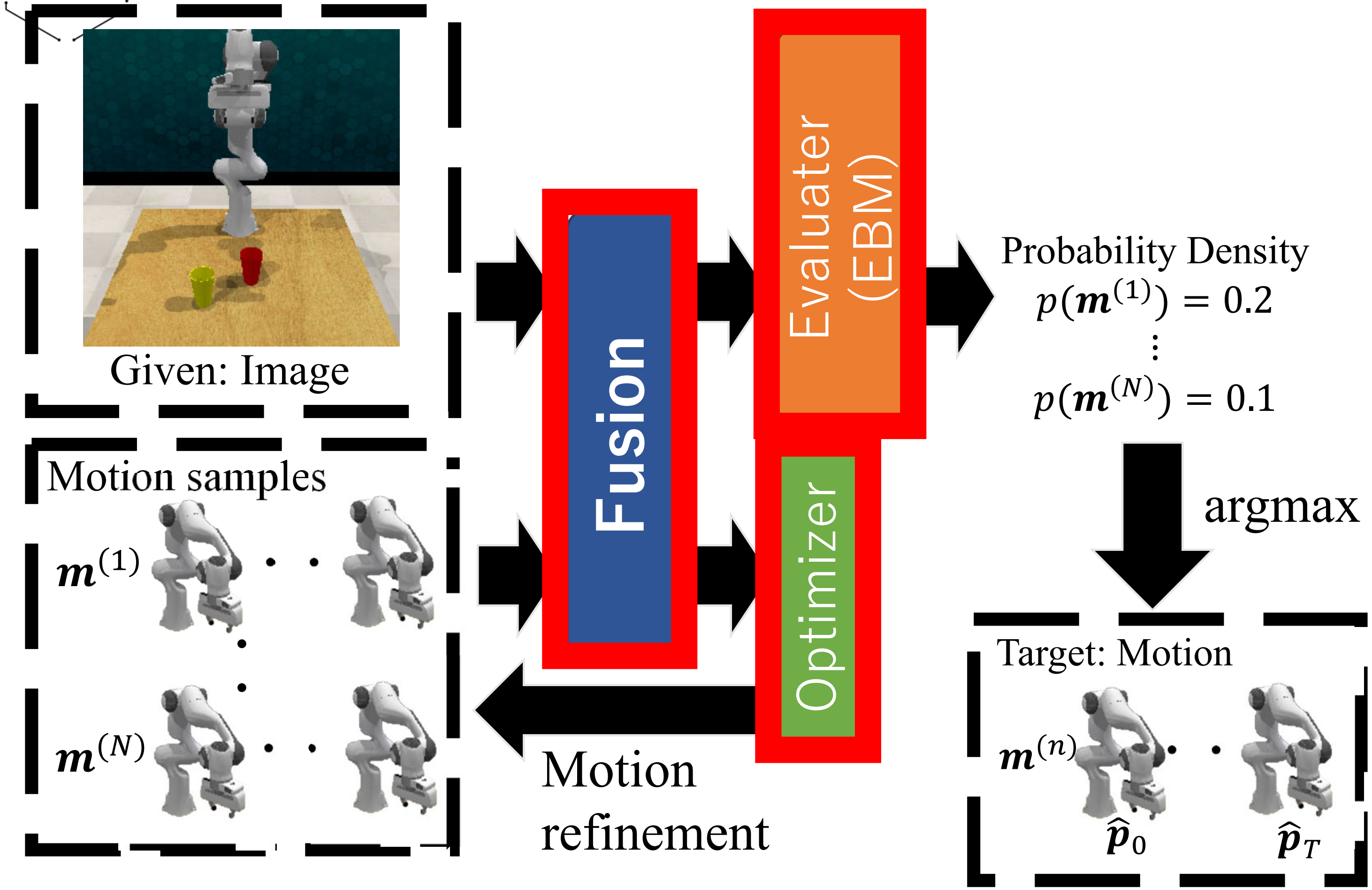
- High-fidelity motion refinement
  - Refinement in a high-resolution refinement space

➡ HR feature space by Spatially-aligned Temporal Encoding (STE)

- Generative models
  - Successful in-distribution sampling from a limited number of samples

➡ EBM augmented by VAE

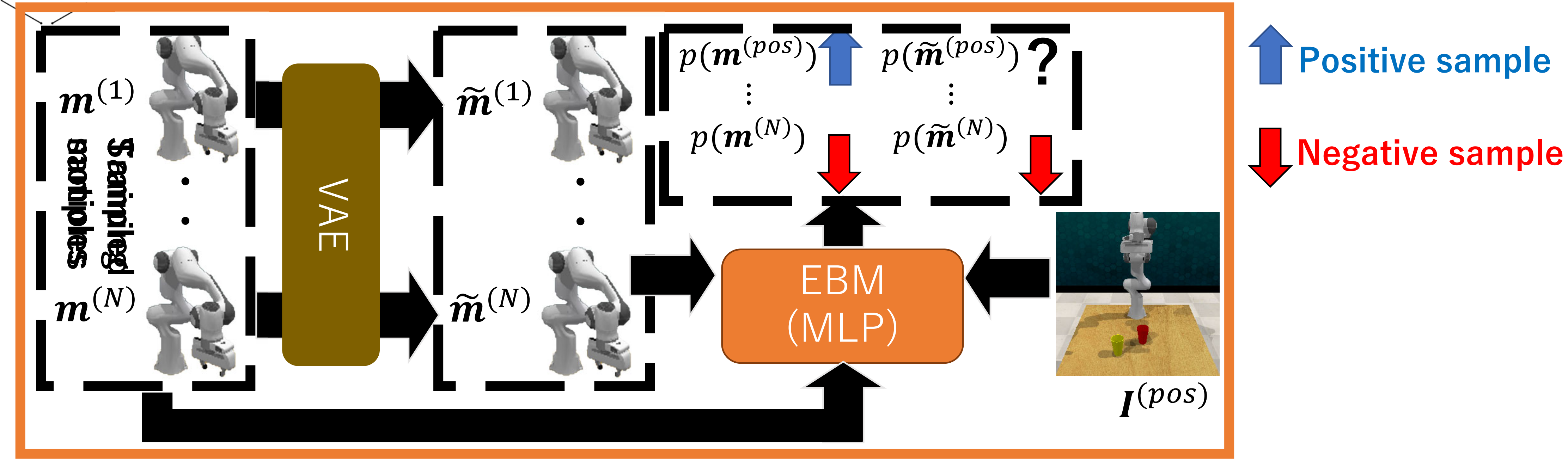
# Overview: EBM + Optimizer + Fusion




**Goal**

1. **Evaluate a consistency between each optimized motion and the given image probabilistically.**
2. **Optimize each motion for the environment expressed in the image.**
3. **Fuse synchronized image and motion data in a high-dimensional feature space for consistency evaluation.**

# 1. EBM Training with Real Samples and Samples Augmented by VAE



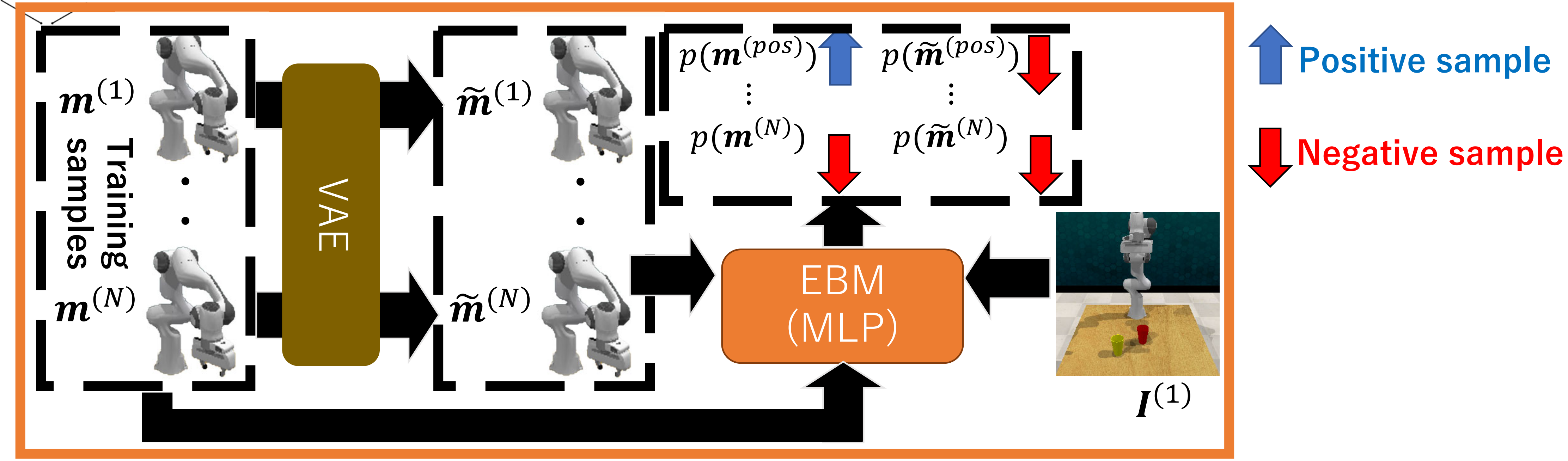

- For training, the gradient of the EBM loss is expressed with motions sampled based on $p(m|I)$.

$$\mathcal{L}_{EBM}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(-E_\theta(I^i, P^i) - \log Z_\theta(I^i)\right)$$

- This difficult sampling in the high-dimensional space is avoided by using real motion samples.
- These motion samples are augmented from real training samples by VAE.

# Why Traditional Models Fail to Grasp?

# 1. EBM Training with Real Samples and Samples Augmented by VAE

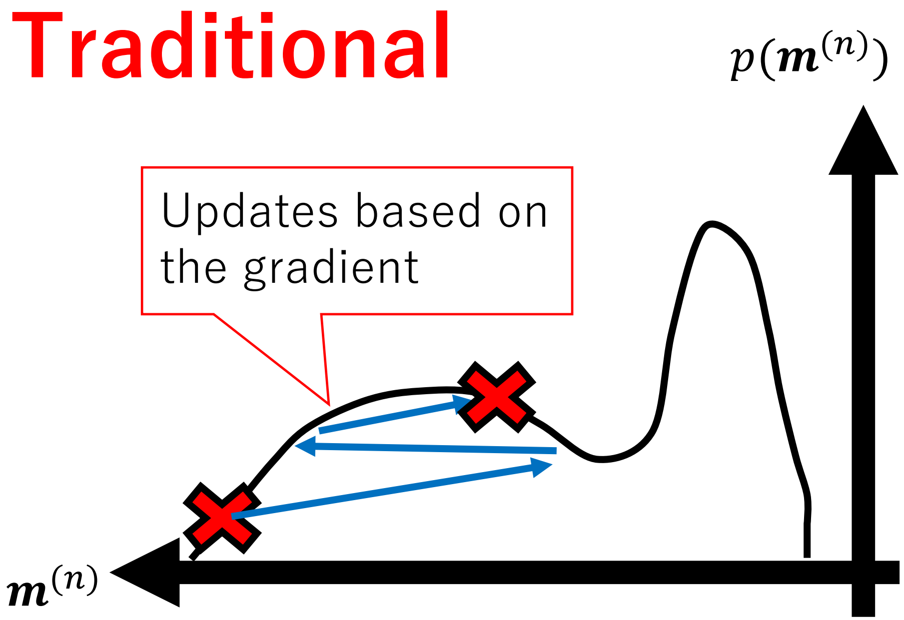


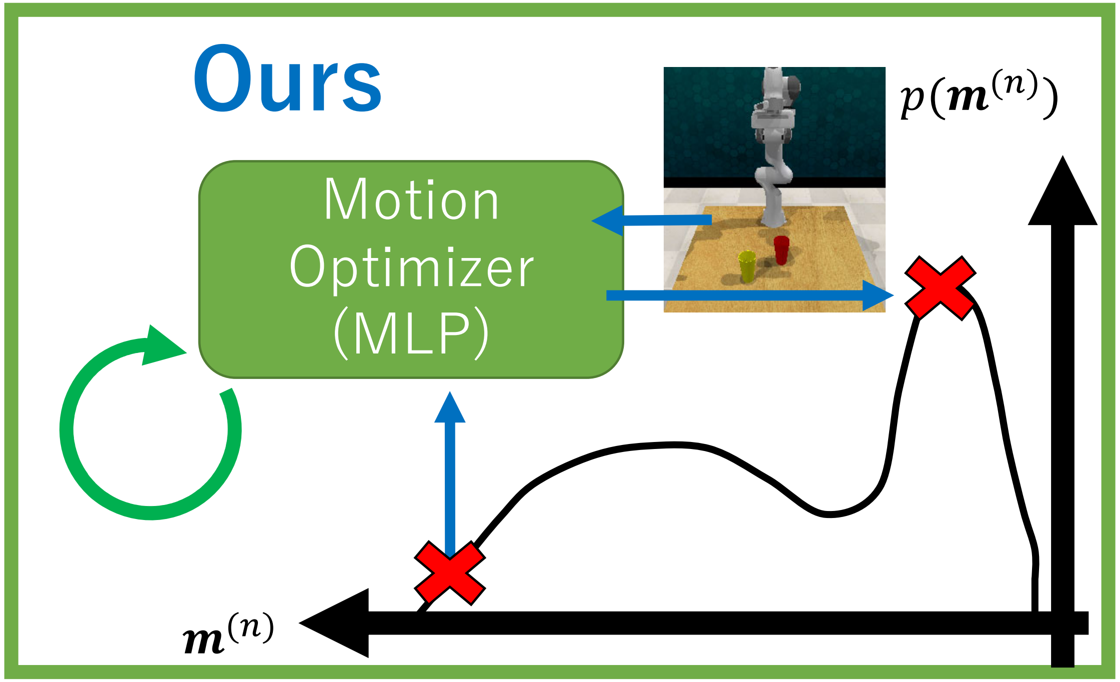

- For training, the gradient of the EBM loss is expressed with motions sampled based on $p(m|I)$.

$$\mathcal{L}_{EBM}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(-E_\theta(I^i, P^i) - \log Z_\theta(I^i)\right)$$

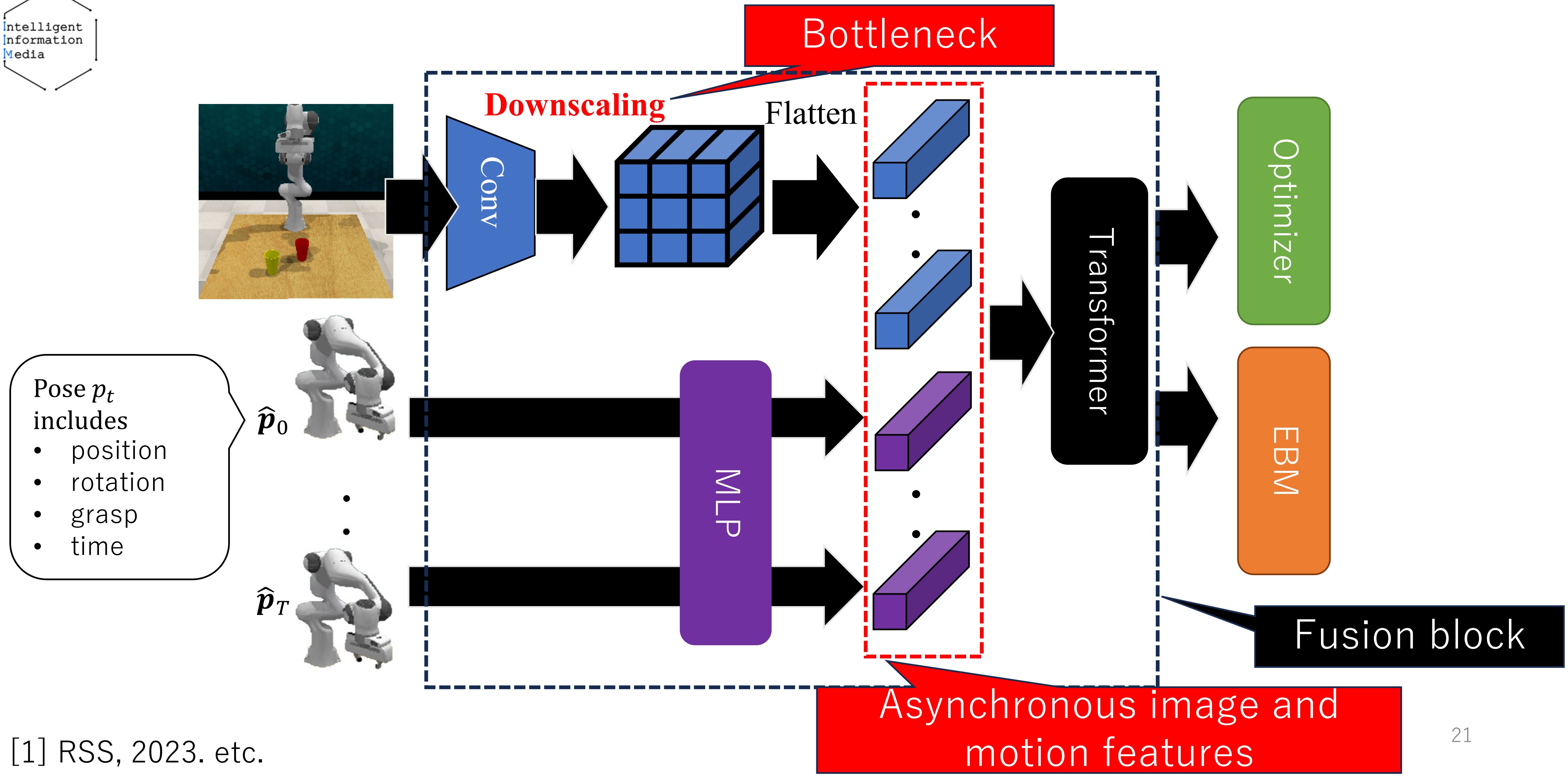
- This difficult sampling in the high-dimensional space is avoided by using real motion samples.
- These motion samples are augmented from real training samples by VAE.

# 2. Optimizer w/o Gradient Descent in Supervised Manner



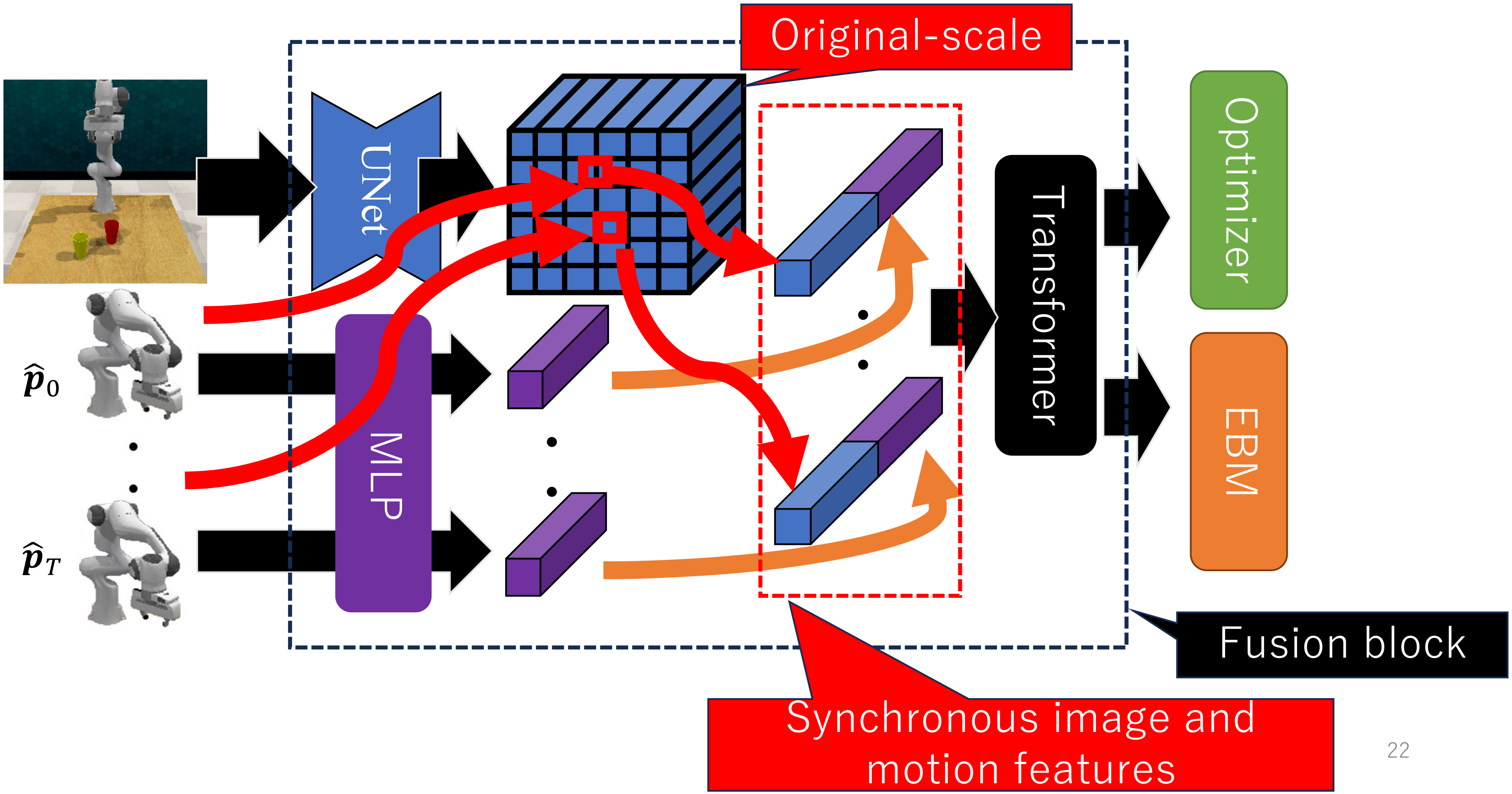



# 3. Traditional Fusion [1]

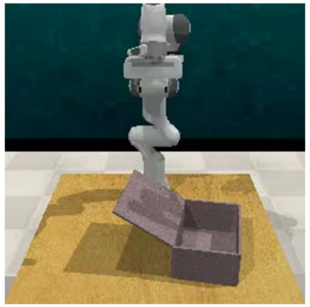



[1] RSS, 2023. etc.

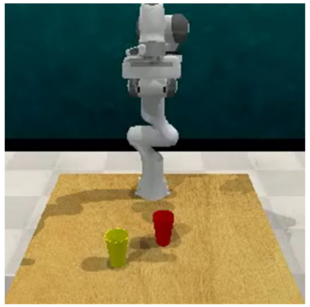
# 3. Proposed Fusion: Spatially-aligned Temporal Encoding (STE)



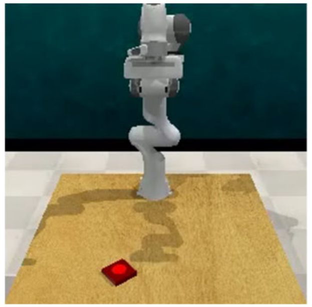
# Experiments: Tasks



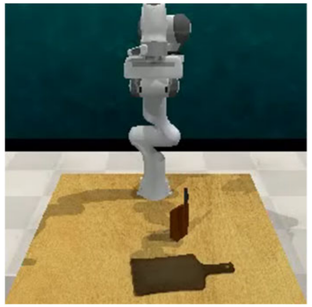
Close Box (CB)



Pick up Cup (PC)



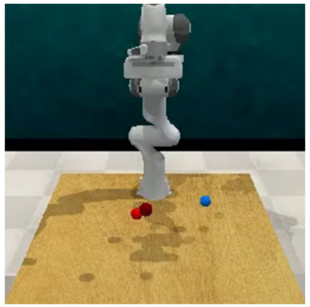
Push Button (PB)



Put Knife (PK)

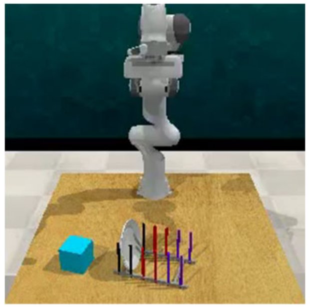
Put Rubbish (PR)

Put Umbrella (PU)



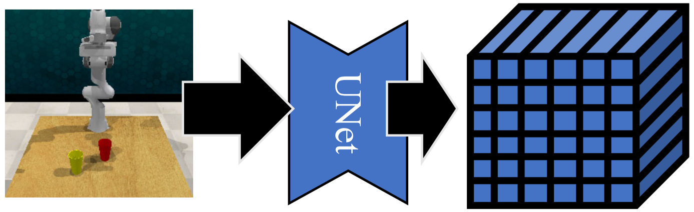
Reach Target (RT)

Stack Wine (SW)



Take Plate (TP)

# Results: Task Success Rates



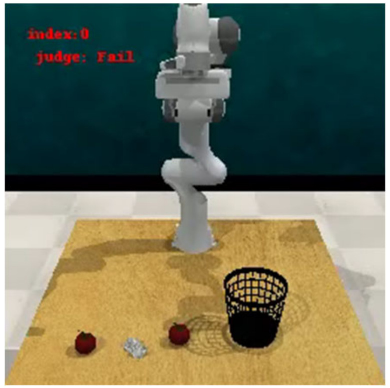

| | CB | PC | PB | PK | PR | PU | RT | SW | TP |
|---|---|---|---|---|---|---|---|---|---|
| (a) **Ours** | **97** | **89** | **85** | **51** | **85** | **25** | **32** | **74** | **94** |
| (b) VAEBM [1] | 64 | **88** | 81 | **27** | 35 | 10 | **32** | **48** | 77 |
| (c) Ours w/Langevin [2] | **71** | 77 | **83** | 0 | **46** | **11** | 0 | 18 | 77 |
| (d) Ours w/ GD [3] | 54 | 78 | 61 | 18 | 40 | 5 | **36** | **48** | **81** |
| (e) Ours w/ GAP [4] | 5 | 3 | 0 | 4 | 1 | 0 | 3 | 5 | 0 |
| (f) Ours w/ ViT (Traditional) | 0 | 0 | 0 | 1 | 1 | 1 | 4 | 0 | 2 |

[1] ICLR, 2021. [2] Bernoulli, 1996. [3] Neural networks , 2003. [4] CORL, 2021.

# Results: Visual Comparison

Put Rubbish

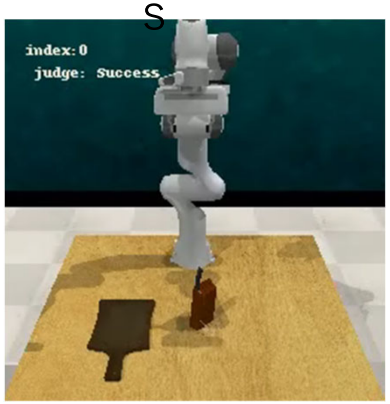


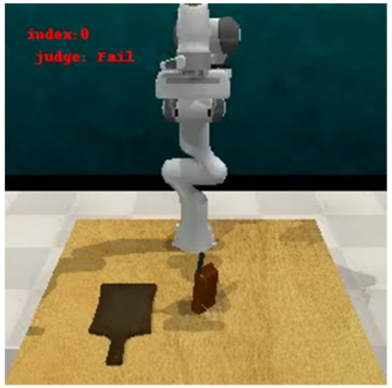


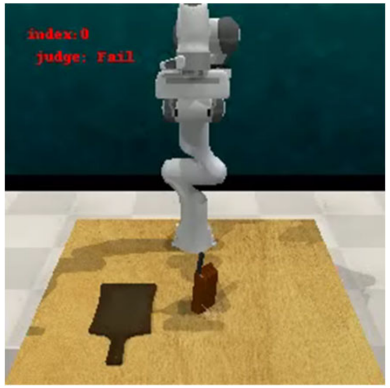


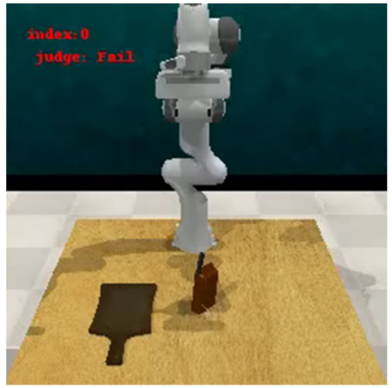
Ours | Gradient Descent | Langevin MCMC | VAEBM

Put Knife

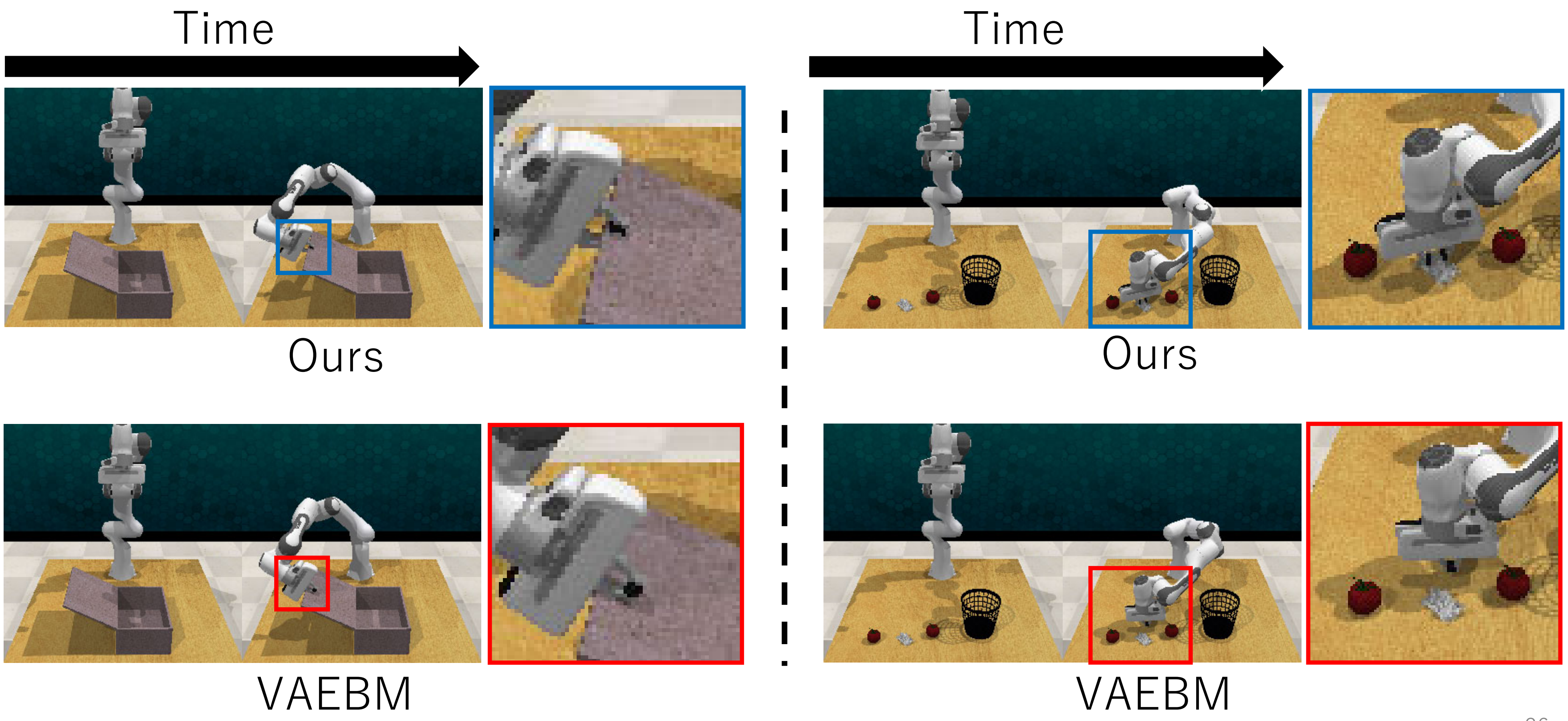


# Results: Detailed Visual Comparison



Intelligent
Information
Media

# Concluding Remarks

# Summary and Future Work

- Summary
  1. Super-fast task-agnostic probabilistic prediction
  2. Physically-constrained human motion
  3. Robot motion planning with the initial state presented by an image
- Future Work
  1. Extension to High-dimensional data
  2. End-to-end network with differentiable physics simulator
  3. For physically-realistic motion planning
     1. Physical & other constraints in optimization
     2. Domain gap between simulation and real data: Cyber-Physical systems