OCTOBER 1 - 5, 2023

IEEE/RSJ International Conference on Intelligent Robots and Systems

Learning a Causal Transition Model for Object Cutting

Zeyu Zhang[1,2*], Muzhi Han[1,2*], Baoxiong Jia[1,2], Ziyuan Jiao[1,2], Yixin Zhu[4], Song-Chun Zhu[1,3,4], Hangxin Liu[1†]

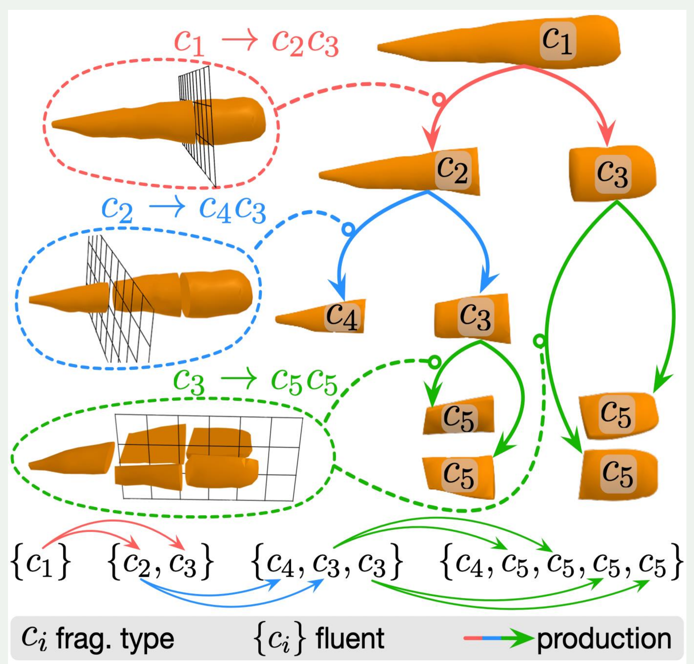
[1]National Key Laboratory of General Artificial Intelligence, Beijing Institute for General Artificial Intelligence (BIGAI). [2]Center for Vision, Cognition, Learning, and Autonomy, UCLA. [3]School of Intelligence Science and Technology, Peking University. [4]Institute for Artificial Intelligence, Peking University. *Equal contributions †Corresponding Author

Introduction

Motivation: Understand object fragmentation

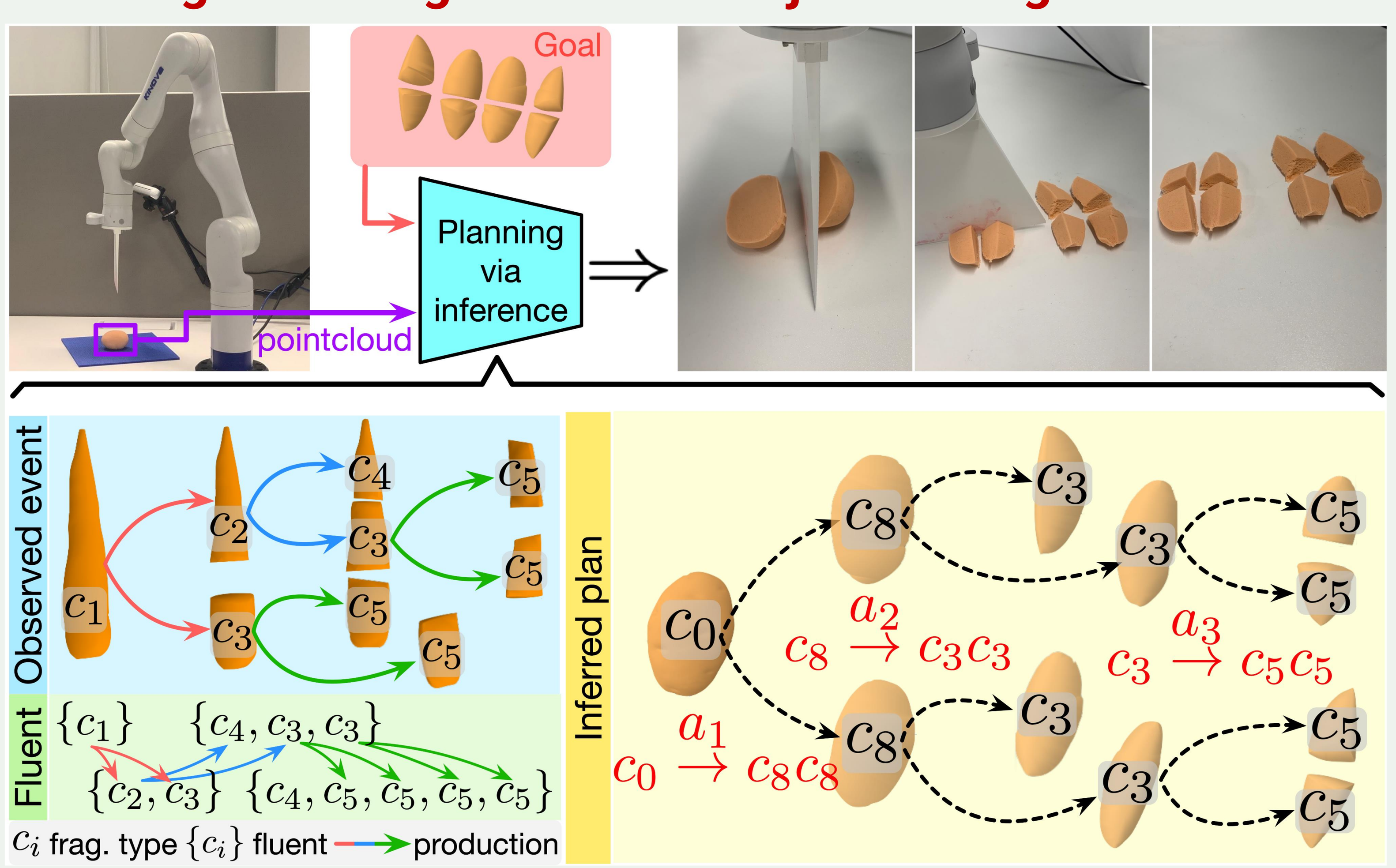
- Understand **fragments:**
  - Different fragments look alike, whereas some of them are pre-attentively different.
  - How to **properly discriminate** fragments?
- Understand **transitions** in object fragmentation:
  - Changing **instance number** and **shape.**
  - **Large state (i.e., fluent) space.**

Modeling fragmentation via attributed stochastic grammar



- We use a **grammar model** to define the **states** and **transitions**
  - **Nodes** represent fragment types.
  - **Production rules** define the one-to-many transitions.
  - A **parse tree** represents a specific fragmentation process.
  - **The set of terminal nodes** in a parse tree defines the state resulted from a fragmentation process.

Planning with the grammar for object cutting



- **Planning for object cutting** is equivalent to **inferring an optimal parse tree** of the grammar.
- The learned production rule can **generalize** to cutting unseen objects.

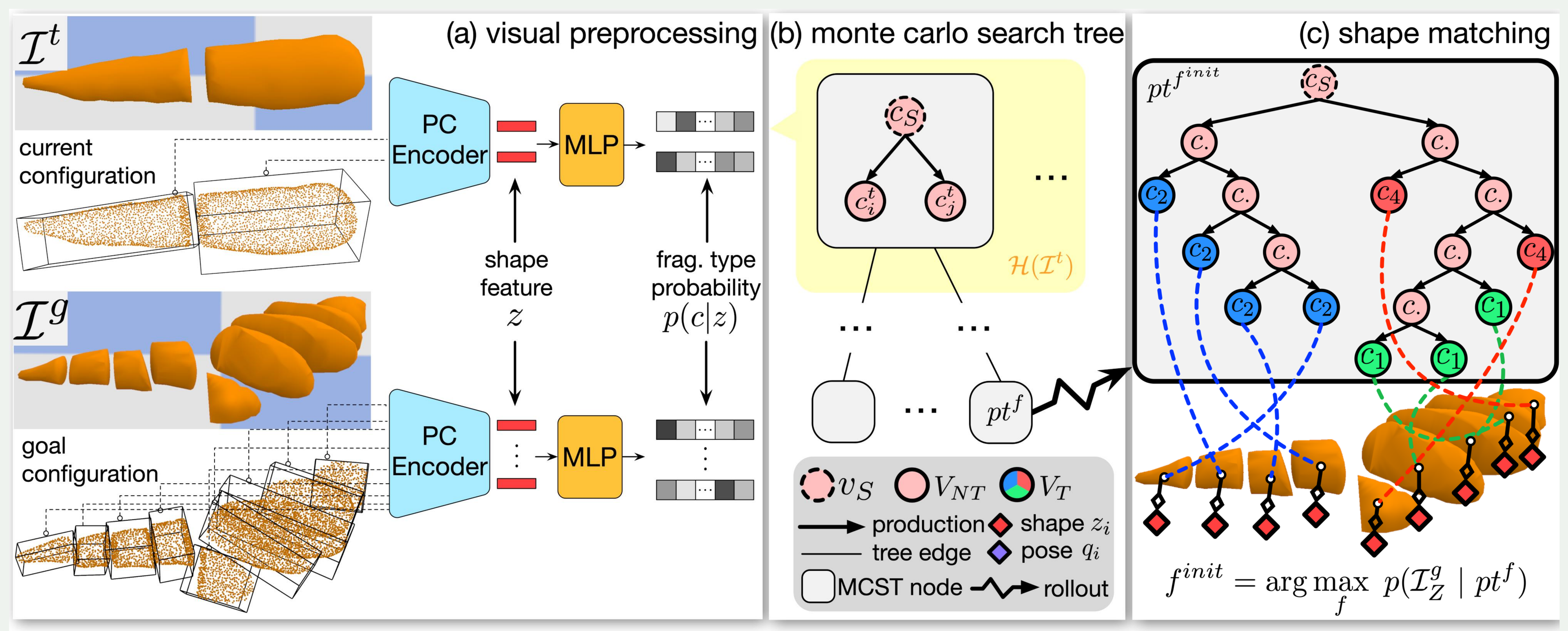
Framework Overview

Learning grammar from collected object cutting data

$$\mathcal{G}^* = \underset{\mathcal{G}^k}{\operatorname{argmax}}\ p(\mathcal{D}_c^k \mid \mathcal{G}^k)\, p(\mathcal{G}^k)$$
$$= \underset{\mathcal{G}^k}{\operatorname{argmax}} \underbrace{\prod_{(\alpha_i \rightarrow \beta_i) \in \mathcal{D}_c^k} p(\alpha_i \rightarrow \beta_i \mid \mathcal{G}^k) \cdot \underbrace{e^{\gamma |\mathcal{G}^k|}}_{\text{model prior}}}_{\text{data likelihood}},$$

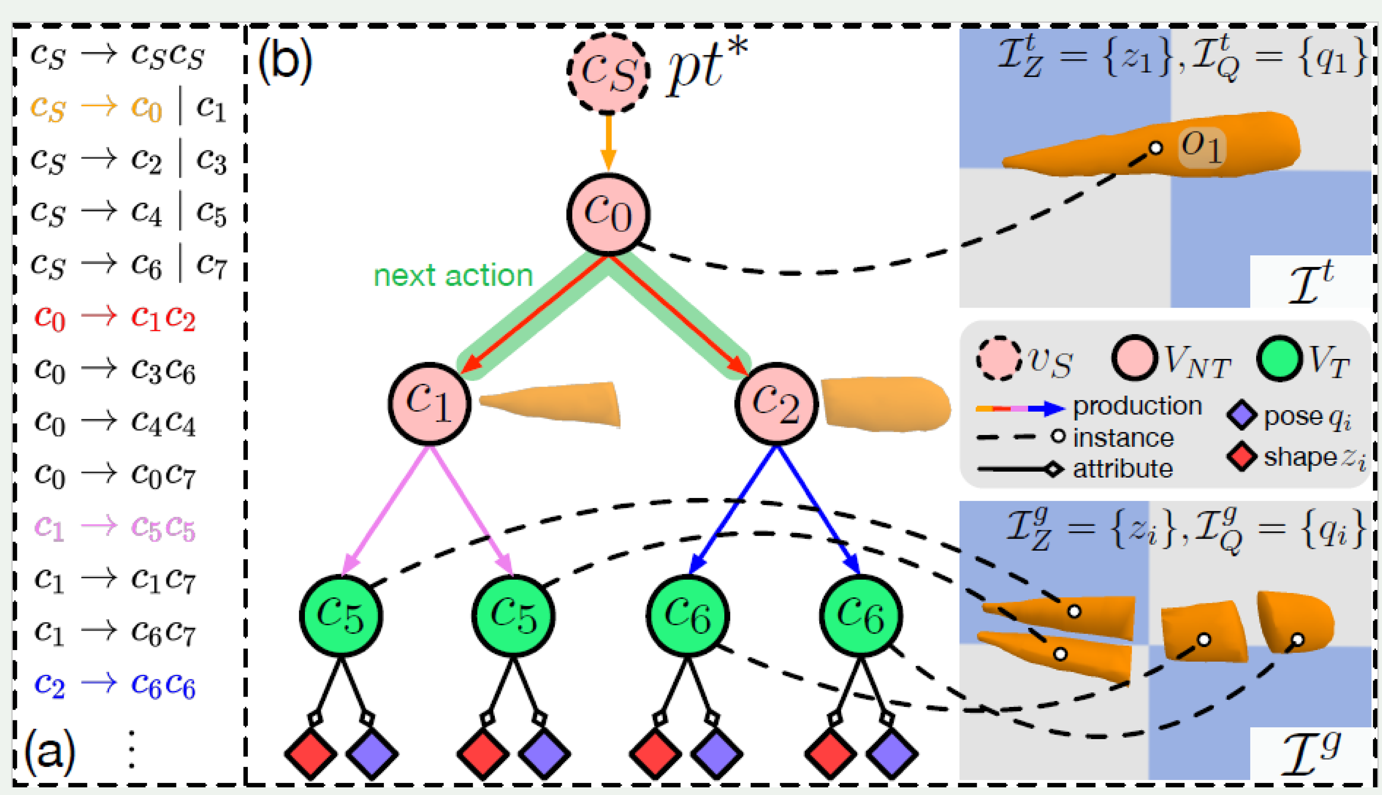
- We induce the grammar from human demonstrations of object cutting:
  - Extract **shape features** for each fragment, and cluster them into k **fragment types**.
  - Learn grammar from recorded transitions with a **MAP** objective.
    - **The objective balances the number of fragment types k and grammar complexity.**

Planning as Inference: Inferring an optimal parse tree



- Given the point clouds of the current and goal configuration, we use a pre-trained encoder and an MLP to predict the **fragment type probabilities**.
- We adopt **Monte-Carlo Tree Search** to find the optimal parse tree that transits the current configuration to the goal.

Bridging abstracted actions and continuous motion

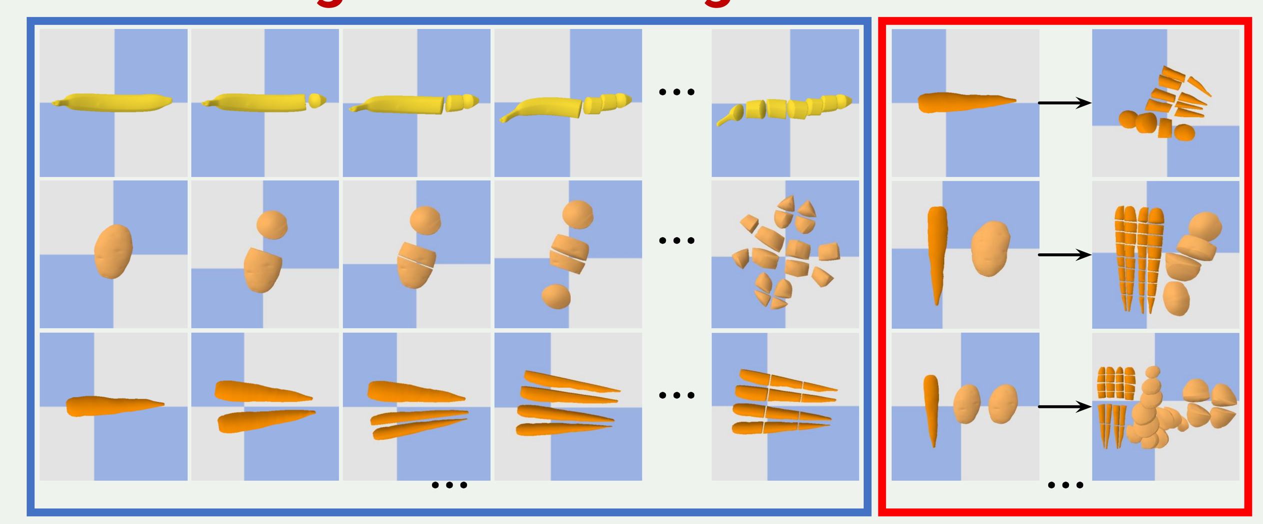


- Each **production rule** is associated with a **cutting action,** parameterized by a **cutting plane.**
- We fit a **GMM** to the cutting plane parameters using human cutting data.

$$p(\boldsymbol{\pi} \mid r, z) = \sum_{i=1}^{k} w_i \mathcal{N}(\boldsymbol{\mu}_i(r, z), \boldsymbol{\sigma}_i(r, z))$$

Evaluation

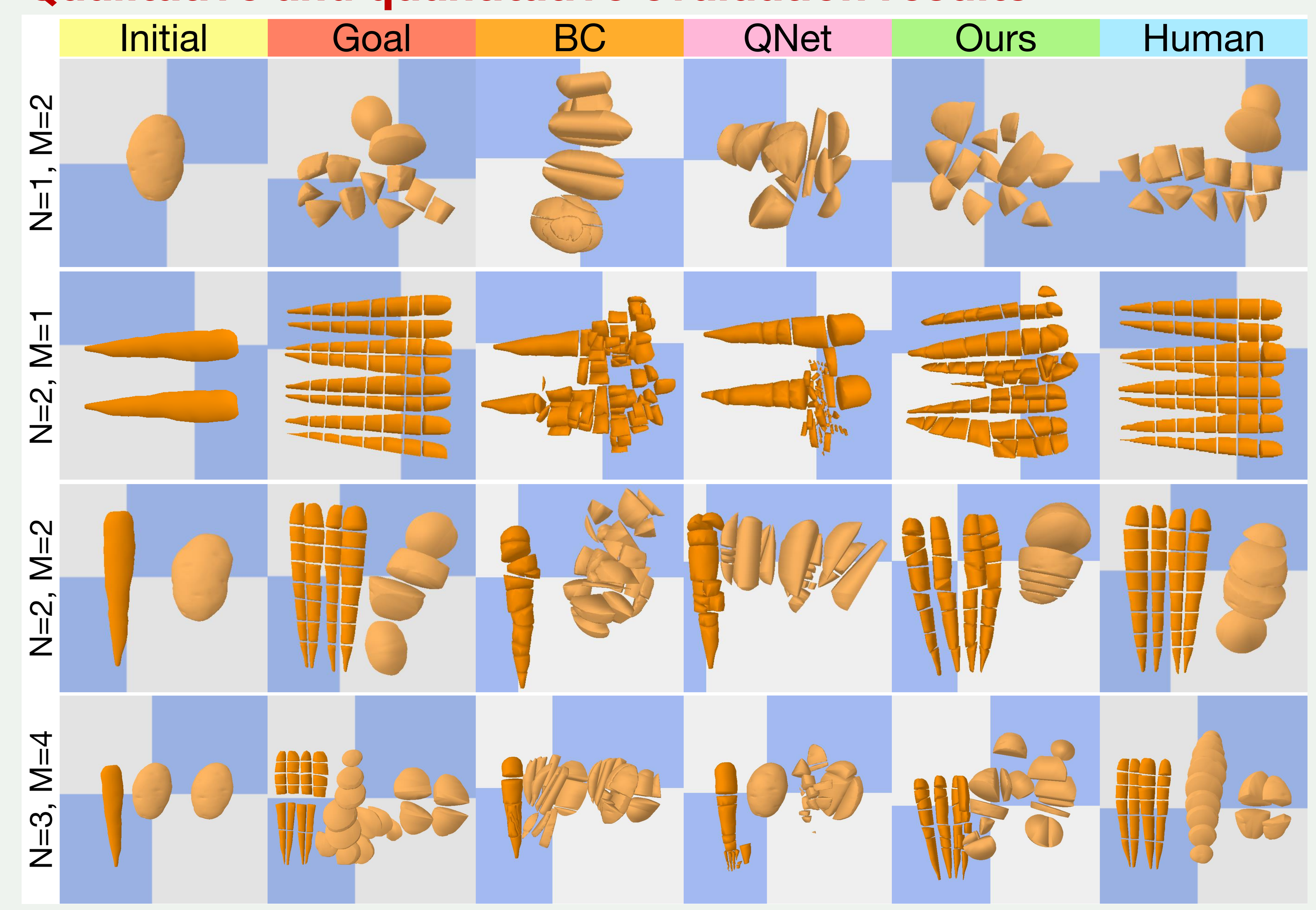
Partitioning of the training and test sets



Training:
$N = 1, M = 1$
Testing:
$N \geq 1, M \geq 1$

- N: initial number of objects.
- M: number of fragment types in goal.

Qualitative and quantitative evaluation results



Task Setup		BC		QNet		Ours		Human	
		IoU	HR	IoU	HR	IoU	HR	IoU	HR
Seen	N=1, M=1	0.37±0.11	2.19±1.07	0.40±0.16	2.14±1.21	**0.58**±0.08	**4.32**±0.77	0.57±0.03	4.48±0.96
Unseen	N=1, M=2	0.35±0.08	1.76±0.87	0.32±0.12	1.95±0.87	**0.49**±0.06	**3.60**±1.02	0.62±0.07	4.86±0.35
	N=2, M=1	0.44±0.08	1.64±0.65	0.34±0.16	1.19±0.39	**0.56**±0.03	**3.69**±0.89	0.62±0.09	4.83±0.37
	N=2, M=2	0.42±0.03	2.07±0.86	0.29±0.09	1.24±0.43	**0.52**±0.04	**3.74**±0.90	0.56±0.04	4.79±0.56
	N=2, M=3	0.38±0.03	1.73±0.99	0.28±0.09	1.52±0.92	**0.52**±0.03	**3.21**±0.86	0.60±0.04	4.81±0.55
	N=3, M=4	0.38±0.04	1.57±0.62	0.22±0.08	1.26±0.49	**0.52**±0.02	**3.21**±0.86	0.56±0.04	4.81±0.55

Real world object-cutting experiments