Learning Concept-Based Causal Transition and Symbolic Reasoning for Visual Planning

Yilue Qian[1,2,3], Peiyu Yu[4], Ying Nian Wu[4], Yao Su[1], Wei Wang[1], Lifeng Fan[1]

[1] State Key Laboratory of General Artificial Intelligence, BIGAI [2] Institute for Artificial Intelligence, Peking University
[3] Yuanpei College, Peking University [4] Department of Statistics, University of California, Los Angeles (UCLA)

https://fqyqc.github.io/ConTranPlan

Motivation

Motivation: Human's step-by-step planning process

- Human planning involves proposing a sequence of actions to achieve goals. It requires foresight, causality, and imagination to reason through actions and consequential intermediate states before reaching the final goal.

Motivation: Bi-level planning, concept and symbol

- Changes to objects are limited to specific attributes. Understanding these conceptual changes beyond mere pixels enhances generalizability and interpretability. Further abstracting concepts to symbols improves reasoning, planning, and overall comprehension.

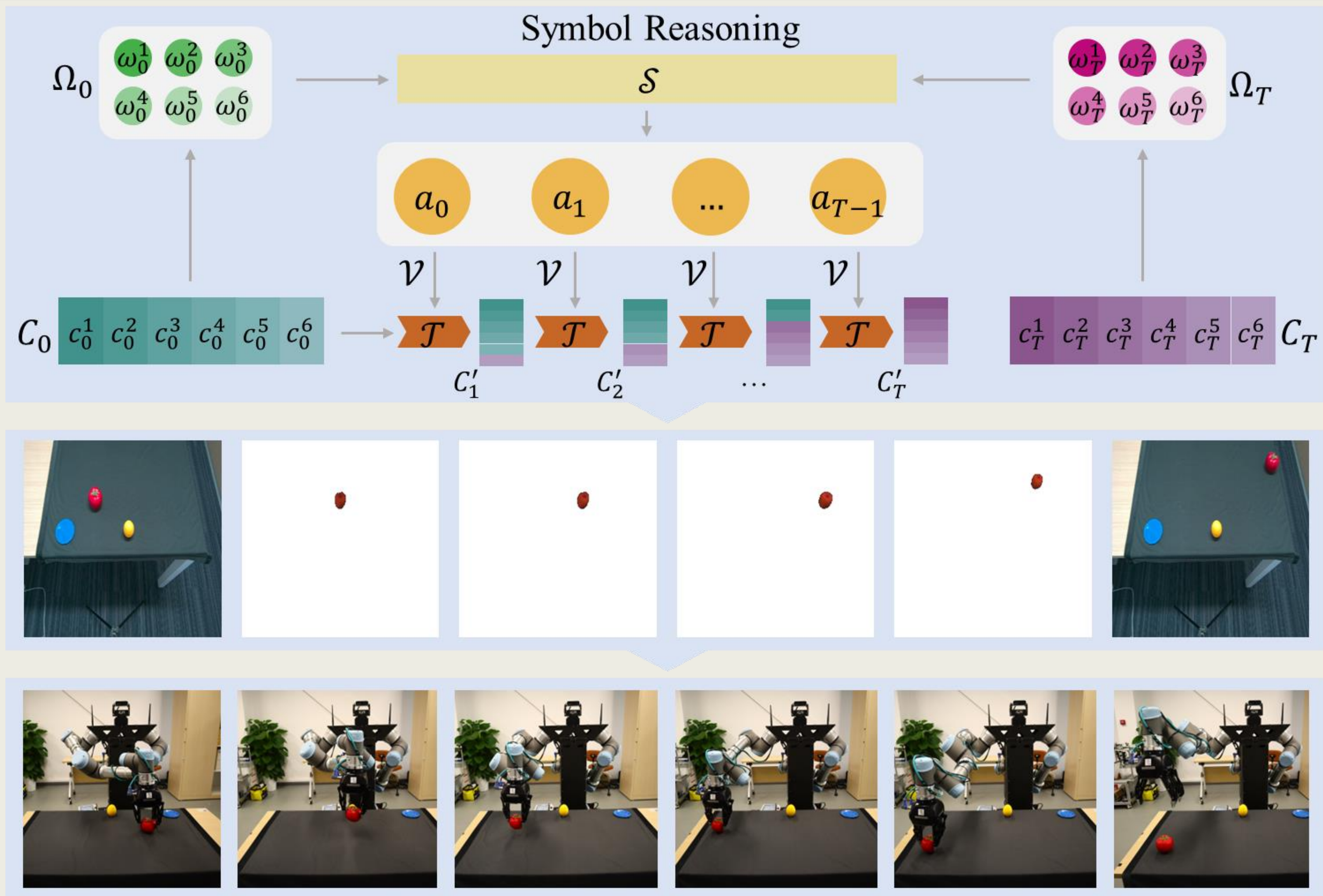
Contribution

Proposing an interpretable and generalizable visual planning framework with the following highlights:

- Disentangled concept-level representation
- Concept-level visual causal transition with discrete symbolic abstraction and reasoning (Bi-level planning)
- Explicit intermediate states awareness
- Generalizable and Interpretable causal chains

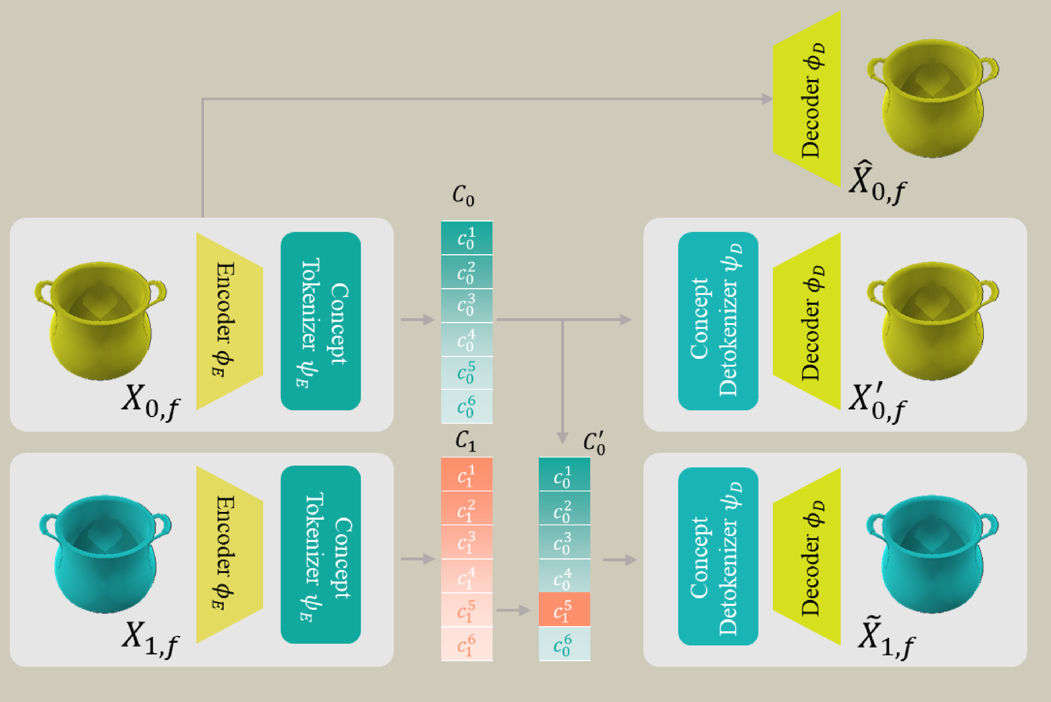
Also, announcing dataset *CCTP*, including concept learning, visual planning, and real-world planning tasks.

Task & Method



Overview

- Given an initial state and a goal state, we aim to predict the intermediate states (2nd row) that will guide a robot to manipulate the target objects (3rd row).
- The disentangled concept representation C, abstracted symbol representation Ω, their corresponding causal transition $\mathcal{T}$ and symbol reasoning $\mathcal{S}$, are effectively combined into a bi-level planning framework for better generalization (1st row).

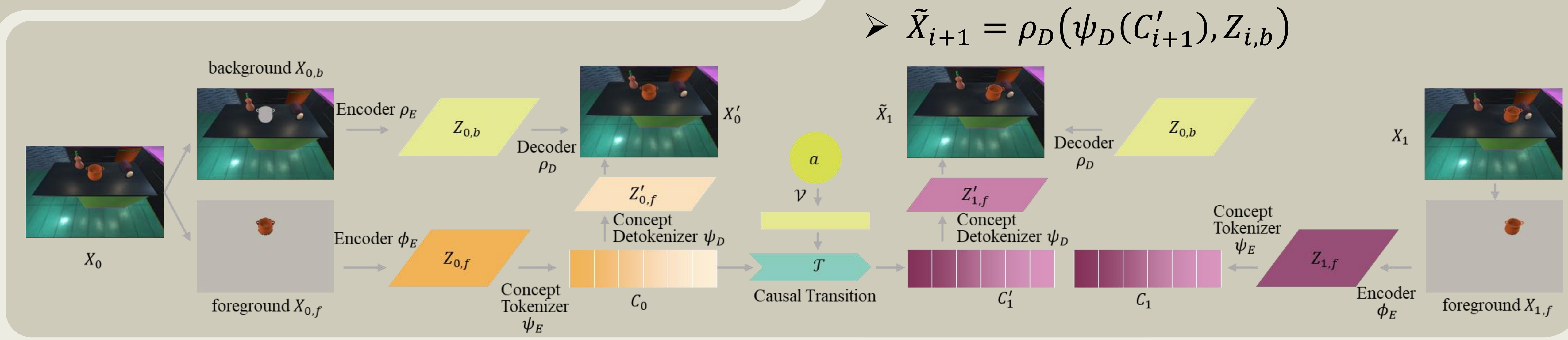


Substitution-based Concept Learner (SCL)

- $C_i = \{c_i^k\}_{k=1,\dots,6} = \psi_T(Z_{i,f}) = \psi_T\left(\phi_E(X_{i,f})\right)$

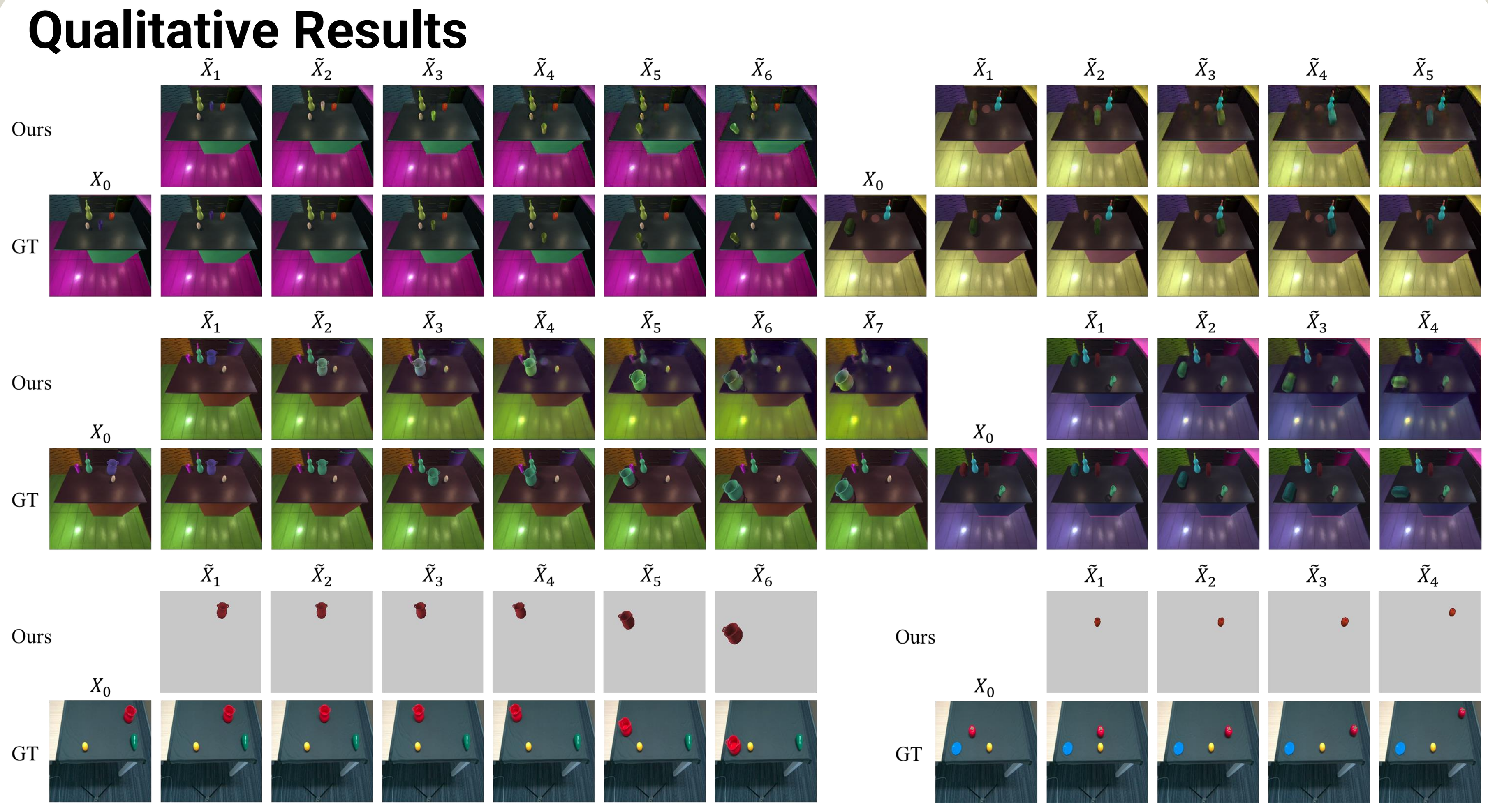
Visual Causal Transition model (ViCT)

- $C'_{i+1} = \mathcal{T}(C_i, \mathcal{V}(a_i))$
- $\tilde{X}_{i+1} = \rho_D(\psi_D(C'_{i+1}), Z_{i,b})$

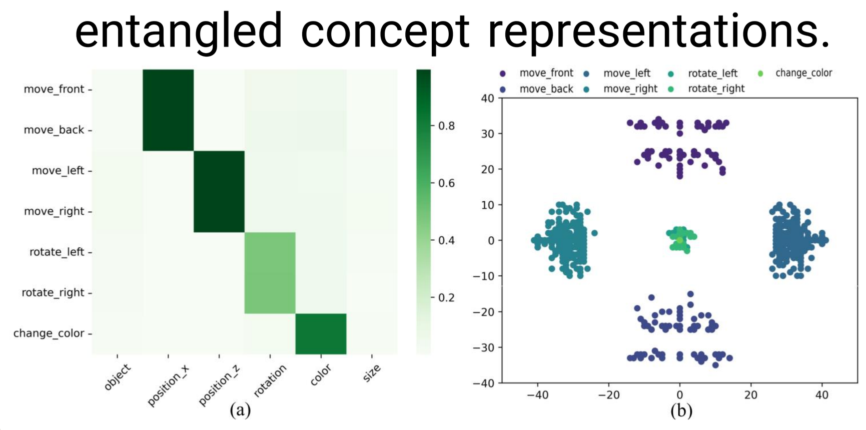


Results & Conclusion

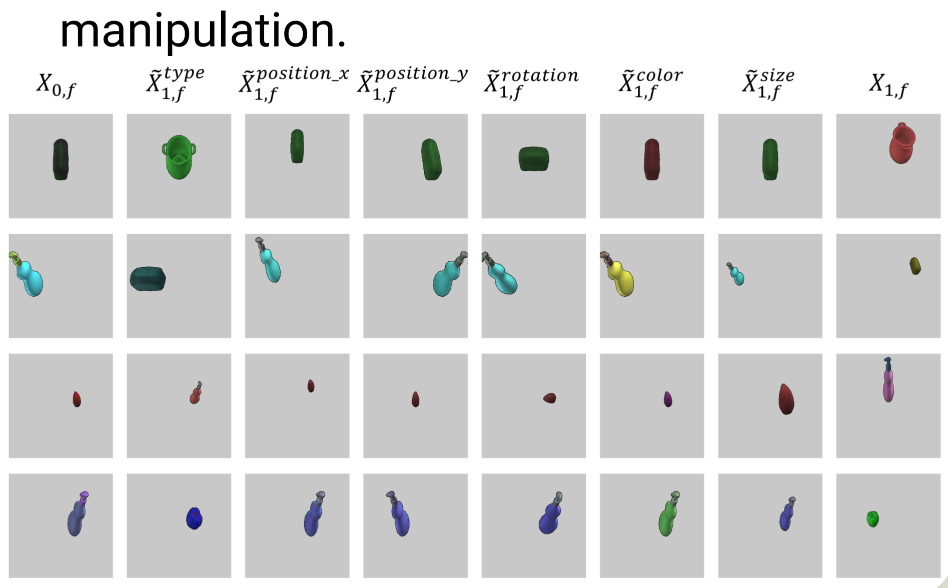
Qualitative Results



Qualitative results of our visual planning model.



Action effects on the learned disentangled concept representations.



Fine-grained attribute-level concept manipulation.

Quantitative Results

- **Best Path Accuracy:** The highest ASAcc and FSD by large margins.
- **Competitive Efficiency:** The ASE of our models exceeds that of all models with reasonable ASAcc.
- Maintains strong performance when encountering harder tasks.
- **Generalizability:** Robust to *Unseen Object* and *Unseen Task* Tests.
- **Real-world Tests:** Successfully recognition of objects' attributes, and best performance compared to all comparison models.

	Model Name	ASAcc.(%)(↑) Top-1	ASAcc.(%)(↑) Top-5	ASE(↑)	FSD(↓)	ASAcc.(%)(↑) Top-1	ASAcc.(%)(↑) Top-5	ASE(↑)	FSD(↓)
CCTP Main Dataset		Dataset level-1				Dataset level-2			
	Chance	1.3	7.3	-	3.139	0.4	2.2	-	3.499
	PlaTe [4]	38.9	-	-	-	15.3	-	-	-
	Ours w/ β-VAE [39]	0.5	3.0	0.970	3.220	0.0	3.5	-	3.670
	Ours w/ VCT [31]	54.1	60.6	0.972	1.483	1.6	4.9	0.988	1.294
	Ours w/o symbol	65.8	76.9	0.983	1.197	41.0	52.6	0.962	1.627
	Ours w/o concept	56.9	77.6	0.986	1.644	-	-	-	-
	Ours w/o causal	1.4	-	-	3.326	0.3	-	-	3.419
	Ours w/ RL	29.7	35.1	**0.991**	2.418	2.5	6.0	**1.000**	3.150
	Ours	**97.9**	**99.2**	0.971	**0.025**	**99.4**	**99.6**	0.981	**0.013**
		Dataset level-3				Dataset level-4			
	Chance	0.0	0.4	-	3.513	0.1	0.4	-	3.147
	PlaTe [4]	0.7	-	-	-	0.4	-	-	-
	Ours w/ β-VAE [39]	0.0	0.5	-	3.596	0.0	0.0	-	3.107
	Ours w/ VCT [31]	0.7	1.2	0.968	3.442	0.2	0.3	1.000	3.193
	Ours w/o symbol	15.4	24.1	0.970	2.278	9.8	14.0	0.981	2.149
	Ours w/o causal	0.0	-	-	3.691	0.0	-	-	3.201
	Ours w/ RL	3.0	3.9	**1.000**	3.030	2.8	3.5	**1.000**	2.498
	Ours	**86.5**	**87.0**	0.966	**0.037**	**55.1**	**76.7**	0.978	**0.003**
Unseen Object		Dataset level-1				Dataset level-2			
	Chance	0.6	4.7	-	3.203	1.1	3.2	-	3.591
	PlaTe [4]	18.5	-	-	-	9.7	-	-	-
	Ours w/o symbol	44.0	59.9	0.968	1.507	29.0	43.8	0.986	1.880
	Ours w/o concept	37.1	60.5	0.950	1.319	-	-	-	-
	Ours w/o causal	1.7	-	-	3.233	0.2	-	-	3.563
	Ours w/ RL	30.2	35.9	**0.989**	1.887	2.2	6.1	**1.000**	3.549
	Ours	**72.4**	**97.2**	0.987	**0.470**	**73.2**	**93.6**	0.978	**0.491**
		Dataset level-3				Dataset level-4			
	Chance	0.0	0.0	-	3.544	0	0.1	-	3.518
	PlaTe [4]	0.6	-	-	-	0.8	-	-	-
	Ours w/o symbol	12.6	22.5	0.990	2.710	6.9	11.7	0.972	2.917
	Ours w/o causal	0.0	-	-	3.467	0.0	-	-	3.183
	Ours w/ RL	1.9	5.3	**1.000**	3.484	1.4	4.9	**1.000**	3.370
	Ours	**61.8**	**66.9**	0.960	**0.307**	**29.1**	**43.9**	0.954	**0.424**
Unseen Task		Dataset level-1				Dataset level-2			
	Chance	0.4	2.1	-	3.550	0.1	0.3	-	3.513
	PlaTe [4]	1.4	-	-	-	0.5	-	-	-
	Ours w/o symbol	63.1	78.0	0.974	1.022	40.0	51.9	0.980	1.407
	Ours w/o concept	42.7	70.7	0.971	1.485	-	-	-	-
	Ours w/o causal	0.0	-	-	3.536	0.0	-	-	3.525
	Ours w/ RL	26.3	30.1	**0.994**	2.159	2.8	7.0	**1.000**	3.417
	Ours	**98.7**	**99.3**	0.985	**0.015**	**98.2**	**99.4**	0.991	**0.019**
Real-world Data		Dataset level-1				Dataset level-2			
	Chance	2.0	5.0	-	3.261	1.0	2.0	-	3.370
	PlaTe [4]	12.0	-	-	-	5.0	-	-	-
	Ours	**52.0**	**71.0**	**0.980**	**1.341**	**36.0**	**47.0**	**0.987**	**1.765**
		Dataset level-3				Dataset level-4			
	Chance	0.0	1.0	-	3.498	0.0	0.0	-	3.552
	PlaTe [4]	1.0	-	-	-	1.0	-	-	-
	Ours	**21.0**	**27.0**	**0.993**	**1.436**	**11.0**	**15.0**	**1.000**	**1.735**

Conclusion

- We propose a novel visual planning model involving concept-based disentangled representation learning, symbolic reasoning, and visual causal transition modeling.
- In the future, we plan to extend our model to more complex planning tasks with diverse concepts and actions, assisting robots in real downstream application tasks.