ConSOR: A Context-Aware Semantic Object Rearrangement Framework for Partially Arranged Scenes

Kartik Ramachandruni, Max Zuo, and Sonia Chernova

***Abstract*— Robots assisting users in the home will be performing various organizational tasks, such as putting groceries in the fridge, or restocking the kitchen pantry, which require the robot to identify the correct placement of objects in a complex environment. Prior work on object rearrangement has explored a diverse set of techniques to determine object placements, but require the user to explicitly communicate their organizational preference via instructions or task demonstrations. In this work, we argue that contextual cues from a partially arranged environment, such as a common household setting, provide sufficient context for robots to perform object rearrangement *without any explicit user goal specification*. We introduce ConSOR, a Context-aware Semantic Object Rearrangement framework that utilizes contextual cues from a partially arranged environment to complete the arrangement of new objects, without user instruction or demonstration. We demonstrate that ConSOR strongly outperforms two baselines in generalizing to novel object arrangements and unseen object categories.**

I. Introduction

Consider a service robot performing an organizational task, such as putting away newly delivered groceries, or tidying a living room. In both tasks, the environment is most likely already partially arranged, and that arrangement provides valuable clues for where new items should be placed. For example, the pantry may contain unfinished boxes of cereals and pasta on different shelves, indicating that a new box of oatmeal should be placed near the cereal instead of the pasta. Prior work in object rearrangement [1] has explored techniques to determine correct object placements in complex environments. These techniques require the user to either provide explicit instructions on arranging objects (e.g., logical predicates [2], goal images [3], natural language descriptions [4]) or demonstrate the rearrangement task for the robot [5]. Rather than burdening users with communicating their organizational preferences, we posit that contextual cues from partially arranged scenes (i.e., the placement of some number of pre-arranged objects in the environment) provide sufficient context to perform object rearrangement *without any explicit user goal specification*.

In this work, we introduce ConSOR, a Context-aware Semantic Object Rearrangement framework that utilizes contextual cues from a partially arranged initial state of the environment to complete the arrangement of new objects, without user instruction or demonstration. Our work makes the following contributions: i) a formalization of the object rearrangement problem in partially arranged environments, ii) a semantic object rearrangement framework that replaces human instruction with contextual cues from a partially arranged environment to infer the desired placement of objects, and iii) a novel dataset of 8k arranged goal scenes generated using 38 household object categories, with each goal scene belonging to one of four organizational schemas, and iv) an evaluation of ConSOR's performance when generalizing to unseen object arrangements with known objects and arrangements with novel object categories.

Georgia Institute of Technology, Atlanta, Georgia, United States. Contact: {kvr6,zuo,chernova}@gatech.edu

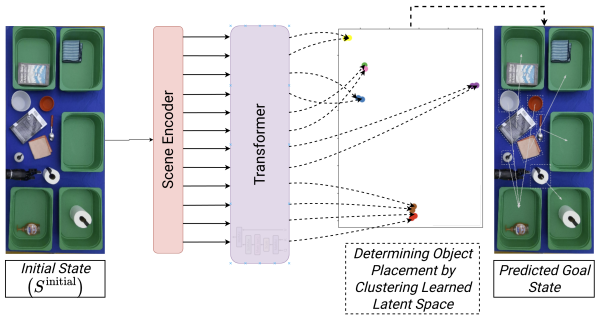

Fig. 1. Model architecture showing how ConSOR encodes a rearrangement scene into an object-centric representation (Scene Encoder block) and transforms it into a learned embedding space (Transformer block), which is then used to determine object placements in the predicted goal state.

II. ConSOR Object Rearrangement Framework

We introduce the Context-aware Semantic Object Rearrangement (ConSOR) framework for enabling robots to rearrange objects in partially arranged environments without goal specification. We define a partially arranged environment as consisting of a set of receptacles $\mathcal{R}$ in which objects can be placed and a set of prearranged objects $\mathcal{X}_P$ arranged within $\mathcal{R}$ to match the user's organizational style. The robot must infer the user's organizational style from $\mathcal{X}_P$ and $\mathcal{R}$ and accordingly determine the placement of a set of unarranged objects $\mathcal{X}_U$.

Figure 1 shows the various components of the ConSOR framework. First, ConSOR's scene encoder constructs an object-centric scene representation of the initial state by projecting each object instance to a higher-dimensional embedding space. Each object embedding consists of a pretrained ConceptNet embedding corresponding to the object's category, a positional encoding to indicate which receptacle the object lies in, and an instance-specific representation to distinguish between object instances of the same category. Second, ConSOR's transformer encoder produces a latent space of object embeddings from the scene representation. The encoder is trained with a triplet margin loss to mimic the

Method	Class Schema (F_{class})		Utility Schema ($F_{utility}$)		One-of-everything Schema (F_{OOE})		Affordance Schema ($F_{affordance}$)	
	M^{SR}	M^{NSED}	M^{SR}	M^{NSED}	M^{SR}	M^{NSED}	M^{SR}	M^{NSED}
ConSOR (**Ours**)	**99%**	**1.4 (SD=0.5)**	**99%**	**1.2 (SD=0.4)**	**100%**	-	**98%**	**1.0 (SD=0.0)**
Abdo-CF	89%	3.9 (SD=1.3)	93%	3.6 (SD=0.8)	0%	15.4 (SD=3.1)	90%	3.4 (SD=1.1)
GPT-3	36%	3.1 (SD=2.1)	41%	3.2 (SD=2.2)	4%	9.8 (SD=5.6)	40%	3.3 (SD=2.2)

TABLE I

EVALUATION RESULTS FOR EACH SCHEMA CALCULATED ON TEST DATA OF REARRANGEMENT SCENES WITH UNSEEN OBJECT ARRANGEMENTS AND KNOWN OBJECT CATEGORIES.

object grouping in the desired goal state by grouping latent embeddings of objects with the same receptacle in the goal state together. Finally, during inference, ConSOR places each unarranged object in the receptacle whose latent centroid has the highest cosine similarity with the latent embedding of the corresponding unarranged object.

III. DATASET OF ORGANIZATIONAL SCHEMAS FOR OBJECT REARRANGEMENT

To validate object rearrangement in partially arranged environments, we contribute a novel dataset of arranged scenes generated using household objects from the AI2Thor simulator. To generate the data, we defined four organizational schemas to determine how objects are grouped in the goal state: the *Class* schema which groups objects based on their semantic categories, the *Utility* schema which groups objects based on similar application, the *Affordance* schema which groups objects based on similar geometry, and the *One-of-Everything* schema which places objects of the same category in different receptacles. We then created a schema-balanced dataset by generating 2200 goal scenes from each of the four schemas using a set of 38 household object categories and grounded in WordNet. The dataset was split into arrangements with known object categories and arrangements with object categories unseen during training.

IV. EVALUATION RESULTS

We present the results of two generalization experiments, first evaluating generalization to previously unseen arrangements with known objects, and second evaluating zero-shot generalization to novel object categories.

We evaluate object arrangement performance by measuring the similarity between the predicted object arrangement and the schema-computed arrangement. This is quantified using the Scene Edit Distance (SED) measure – the minimum number of object displacements required to reach the goal arrangement from the predicted object arrangement. We derive two aggregate evaluation metrics: the Success Rate, M^{SR}, which corresponds to the fraction of goal states predicted correctly ($SED = 0$), and the Average Non-zero SED, M^{NSED}, or the average number of objects misplaced among incorrectly predicted goal states ($SED > 0$).

We compare ConSOR performance against two baselines. The first baseline, ***Abdo-CF***, is a collaborative filtering technique by Abdo et al. that learns pairwise object similarities for every user from preference rankings of different users [6]. The second baseline, ***GPT-3***, is a GPT-3 large language model prompted with few-shot demonstrations (one from each schema) of how to place unarranged objects in a partially arranged environment. Note that, unlike ConSOR and *GPT-3*, the *Abdo-CF* baseline requires the desired organizational schema label to be provided as input.

Table I presents a summary of evaluation results from testing on scenes with unseen object arrangements. ConSOR yields a higher M^{SR} than both *Abdo-CF* and *GPT-3* across all four schemas and is the only method to perfectly rearrange F_{OOE} scenes. Additionally, ConSOR has the least M^{NSED} score across all four schemas, indicating that, in the rare cases that errors occur, ConSOR's state predictions are closer to the true goal state than that of the baselines. In comparison, *Abdo-CF* has the second-best performance in 3/4 of the schemas while failing to successfully rearrange a single F_{OOE} scene, and *GPT-3* performs the worst on 3 schemas with a slightly higher M^{SR} in F_{OOE} than *Abdo-CF*. When evaluating generalization to scenes with novel object categories, ConSOR yields a success rate of 96.8% in comparison to *GPT-3* with a success rate of 37%. As the *Abdo-CF* baseline requires external semantic knowledge to rearrange novel object categories, and other methods lack this knowledge, we do not evaluate this baseline on novel object categories.

V. CONCLUSION

This work introduces ConSOR, a semantic reasoning framework for object rearrangement that attends to contextual cues from a partially arranged environment to infer the desired object arrangement without instruction or demonstration. ConSOR also leverages external commonsense knowledge from ConceptNet to perform zero-shot generalization to rearrange scenes with novel object categories. Our evaluation of ConSOR's generalization performance reveals that ConSOR outperforms both the *Abdo-CF* and *GPT-3* baselines across all tested conditions.

REFERENCES

[1] D. Batra *et al.*, "Rearrangement: A challenge for embodied AI," *arXiv*, 2020.

[2] C. Paxton *et al.*, "Predicting stable configurations for semantic placement of novel objects," in *CoRL*, 2022.

[3] A. Goyal *et al.*, "IFOR: Iterative flow minimization for robotic object rearrangement," in *CVPR*, 2022.

[4] W. Liu *et al.*, "StructDiffusion: Language-guided creation of physically-valid structures using unseen objects," in *RSS*, 2023.

[5] I. Kapelyukh and E. Johns, "My House, My Rules: Learning tidying preferences with graph neural networks," in *CoRL*, 2022.

[6] N. Abdo *et al.*, "Organizing objects by predicting user preferences through collaborative filtering," *IJRR*, 2016.