2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Transformer-Based Relationship Inference Model for Household Object Organization by Integrating Graph Topology and Ontology

Xiaodong Li[1] , Guohui Tian[2], *Member, IEEE*, Yongcheng Cui[3], Yu Gu[4]

Abstract*— In domestic environments, the conventional organization of objects by service robots often relies on the inherent properties of each object, such as placing fragile bowls in enclosed cupboards. However, this approach tends to overlook the importance of the orderly arrangement of objects, neglecting the specific placement order of bowls within the cabinet. In practice, effective object organization necessitates consideration of both individual properties and the relationships defined by these properties. In this paper, we have constructed a specialized dataset encompassing the ontological properties of household objects along with their relationships. Furthermore, we have introduced a graph-based model to explicitly represent these relationships and proposed a novel feature extraction technique that integrates the Graph Attention Network (GAT) with the BERT model to predict the relationships among objects. Subsequently, we utilized the Transformer framework to train a model, enabling it to infer relationships between objects. Experimental validation demonstrates the effectiveness of our approach in accurately predicting relationships between household objects, thus facilitating their orderly organization. Our approach significantly augments the object organization capabilities for service robots by accurately predicting the relationships among household objects. Our code is available at: *https://github.com/Li-XD-Pro/Household-Object-Organization

I. INTRODUCTION

In modern households, the effective organization of objects plays a vital role in maintaining an orderly living space and enhancing the usability of objects. The challenge in this domain lies in enabling robots to comprehend the intricate relationships among objects, which is essential for tasks like sorting kitchen supplies or arranging bookshelves. For these tasks, robots must understand spatial and functional relationships among objects, for instance, distinguishing between spices and tableware or organizing books by size and shape. Therefore, this paper focuses on improving robots' capabilities in managing these tasks by deeply analyzing object relationships in domestic environments. The process of organizing objects by the robot is shown in Figure 1.

Despite advances in spatial organization and robotic learning—such as Xu's [1] work on visual and semantic reasoning and DALL-E-Bot's [2] innovative use of diffusion models—there is still a gap in fully understanding intricate object relationships in home environments. Unlike methodologies like PARAGON [3], which translates language into structured graphs, or NeatNet [4], which personalizes organization

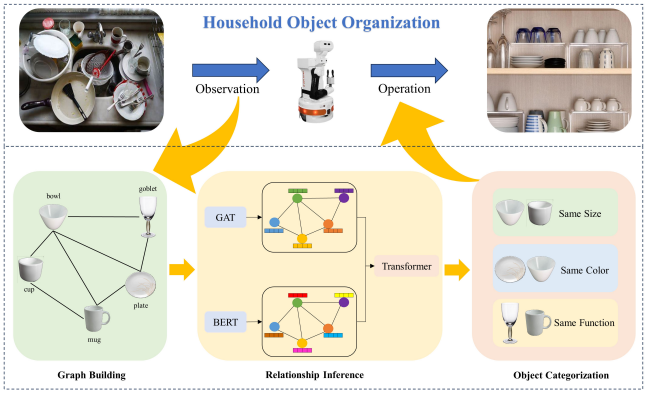

Fig. 1. Robotic organization of household objects. In the presence of a cluttered home setting, the robot constructs a visual scene graph, employs the proposed method to infer relationships among objects, categorizes them, and accordingly carries out the household object organization task.

based on preferences, our paper focuses on nuanced detection of object relationships within households. Effective home organization involves not just the ontological properties of objects, like color and size, but also their interrelationships, such as complementary functions and similar shapes. We address this by creating a comprehensive dataset that outlines both the ontological properties and the diverse relationships between objects, providing a solid foundation for robots to understand and organize home objects effectively.

The field of relationship prediction has advanced significantly, with notable works such as Mo's [5] on relation-aware Graph Convolutional Networks (GCN) and NC-KGE's [6] node-based learning in biomedicine, which underscore the importance of forecasting complex interactions. Building on these models, our study adapts them for the unique challenges of home settings. We use graph theory to model object relationships, leveraging the global perspective of graph structures to understand and predict interactions between household objects, extending beyond direct adjacencies. By combining graph topology with object ontological characteristics and employing GAT and BERT within a Transformer framework, our approach innovatively achieves accurate and efficient object relationship forecasting.

In summary, the contributions of this paper are as follows:

(1) Constructing a detailed dataset tailored for detecting relationships among household objects, including 105 objects, 13 ontological properties, and 11 relationship categories, servicing as a foundational resource for analyzing how objects relate to each other in a home environment.

(2) Introducing a feature extraction technique that merges

This work is supported by National Natural Science Foundation of China under Grant 62273203, the National Key R&D Program of China under Grant 2018YFB1307101, and the Taishan Scholars Program of Shandong Province(ts201511005). *(Corresponding author: Guohui Tian.)*

The authors are with the School of Control Science and Engineering, Shandong University, Jinan 250061, China (e-mail: lix994615@gmail.com; g.h.tian@sdu.edu.cn; cuiyc@mail.sdu.edu.cn; y.gu@mail.sdu.edu.cn)

graph topology and object ontology. By representing object relationships through a graph and integrating Graph Attention Networks (GAT) with BERT for feature extraction, we lay the groundwork for a Transformer-based model that can accurately predict the relationships between objects.

(3) Developing an inference procedure for the detection model, employing YOLOv8 for object detection and CLIP for property verification against a comprehensive ontology knowledge base, thereby facilitating precise reasoning about the relationships among household objects.

II. RELATED WORK

A. Object Organization

Object organization by service robots refers to the process of arranging objects within a space or scene in a manner that is logical, efficient, or aesthetically pleasing.The groundbreaking work by Xu [1] lays the foundation by merging visual and semantic commonsense reasoning, allowing robots to consider both aesthetics and functionality in their organization tasks, thus reducing dependence on human-labeled data. Building on this, the DALL-E-Bot [2] introduces web-scale diffusion models to robotics, enabling object rearrangement to achieve goal images generated from textual descriptions, pushing the boundaries of unsupervised learning. The PARAGON [3] project advances this by incorporating differentiable parsing and visual grounding of natural language instructions, transforming these instructions into structured, object-centric graphs for better handling of complex tasks. The utilization of the CLIP vision-language model for object matching [7] leverages semantic and visual information to improve environmental interactions, further enriched by the StructFormer [8]. This model integrates point cloud data with language instructions, enabling robots to rearrange objects according to semantic guidelines and facilitating the execution of more complex spatial arrangement tasks, signifying a significant leap in the capabilities of robotics in the domain of object organization.

For personalized organization, NeatNet [4] tailors robot organization to individual preferences through the use of Graph Neural Networks and a Variational Autoencoder, illustrating the adaptability of AI to unique spatial preferences. This feature is augmented by a sophisticated system that merges collaborative filtering with spectral clustering. It is designed to fine-tune shelf-tidying actions according to user-specific preferences, showcasing the flexibility of robots within residential spaces [9]. This comprehensive blend of technological advancements underscores a significant evolution in robotics, offering nuanced and user-centric solutions in the realm of object organization.

B. Relationship Prediction

Relationship prediction involves analyzing and forecasting the dynamics or connections between entities in a network or system. Mo [5] introduces the Relation-aware Heterogeneous Graph Convolutional Network (RHGCN), employing a novel algorithm to capture interaction dynamics in temporal heterogeneous networks, marking a step forward in understanding relationship evolution. This is further advanced by a method [10] that combines meta-path-based modeling with recurrent neural networks for predicting relationship formation timing in dynamic networks, refining the approach to capturing relationship dynamics.The Cross-Platform Social Relationship Prediction (CPSRP) [11] study innovates in predicting social relationships across various platforms by combining improved user data aggregation with advanced embedding techniques and sequential feature extraction networks, enhancing the precision of social relationship predictions. Additionally, the introduction of a hierarchical stacking relationship prediction network for robotic manipulation tasks [12] focuses on grasping multiple objects by understanding their hierarchical relationships, thereby optimizing the manipulation process.

In the medical domain, the NC-KGE [6] introduces a node-based contrastive learning method for biomedical relationship prediction, leveraging a multi-head attention mechanism to enhance relation semantics extraction among entities, thus improving prediction accuracy and efficiency. Finally, the RSG-Net [13] develops a graph convolutional network to model and predict semantic relationships in road scenes, accurately capturing and predicting complex semantic relationships in dynamic environments, leveraging comprehensive datasets for training and evaluation.

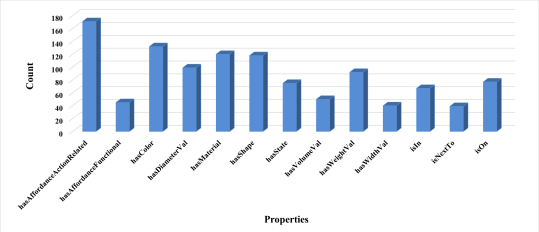

Fig. 2. Distribution of object properties

III. DATASET

A. Household Object Relationship Detection Dataset

This paper addresses the task of organizing household objects by analyzing their relationships to improve the efficiency and accuracy of organization. Due to the lack of datasets in this field, we developed the Household Object Relationship Detection Dataset (HORDD), focusing on the comprehensiveness and detail of object ontological properties. This approach ensures accurate representation of each object's characteristics and functions, and enhances the diversity and practicality of object relationships. The HORDD includes 105 common household objects, each described by 13 ontological properties. We also identified and categorized 11 types of relationships specific to the needs of household object organization. Table I presents examples from the HORDD, while Figures 2 and 3 detail the distribution of ontological properties and inter-object relationships, respectively.

TABLE I

EXAMPLE OF HOUSEHOLD OBJECT RELATIONSHIP DETECTION DATASET (HORDD).

Object1	Properties1	Object2	Properties2	Relationship
eggplant	hasWeightVal 163g hasShape cylinder hasColor purple hasAffordanceActionRelated movable isIn refrigerator hasAffordanceActionRelated graspable	chicken	hasWeightVal 1058g hasAffordanceActionRelated movable hasShape irregularshape hasColor red isIn refrigerator hasAffordanceActionRelated graspable	Action Association Spatial Proximity

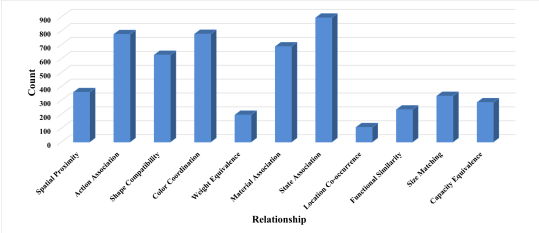

Fig. 3. Distribution of object relationships

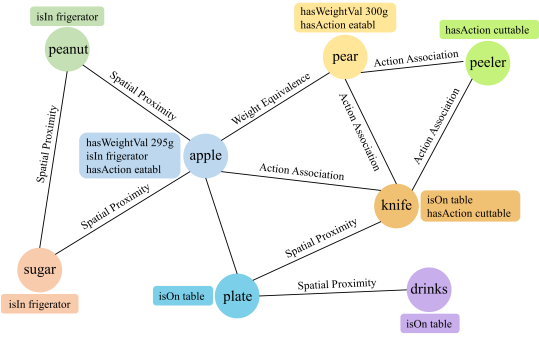

Fig. 4. Household object relationship graph (subgraph)

B. Household Object Relationship Graph

When predicting relationships between household objects, it is crucial to consider not only the objects' properties but also their relationships within the household scene. These relationships can be modeled using graph theory concepts, where objects are nodes and relationships are edges without feature vectors, as described by Scarselli et al. (2008) [14]. This paper uses the Household Object Relationship Detection Dataset (HORDD) to extract objects and their relationships, constructing a graph structure for analysis.

The resulting graph structure allows for the use of graph neural network to learn node feature vectors, thereby uncovering the graph's topological structure and the inter-object relationships. An example subgraph is shown in Figure 4. This graph-based framework integrates topological data with object properties to enhance the accuracy of predicting object relationships. Through this combined approach, we aim to significantly improve prediction accuracy.

IV. METHOD

This section aims to predict relationships between objects by leveraging both graph topological information and object ontological properties. We utilize the Graph Attention Network (GAT) [15] to train the household object relationship graph, obtaining feature vectors that reflect the topological structure. Concurrently, we encode the ontological properties of objects using the BERT [16] model to produce feature vectors that encapsulate ontological information. These vectors are then concatenated to form a comprehensive feature vector for each object. Using these feature vectors, we train an object relationship classification model with the Transformer [17] framework, which predicts the relationships between objects based on their combined features. Furthermore, this paper introduces an inference method that integrates YOLOv8, CLIP [18], an ontology knowledge base, and the classification model to accurately identify and infer relationships between household objects.

A. Graph Embedding

In this subsection, we discuss using the Graph Attention Network (GAT) to generate low-dimensional feature vectors for nodes in the household object relationship graph. GAT effectively captures the graph's topology and the intricacies of node relationships through an attention mechanism. This mechanism dynamically considers neighboring nodes' contributions during node feature updates, leading to richer and more accurate embeddings. These vectors include information about the nodes' properties and the impact of their neighborhood structure, providing essential support for analyzing complex interactions among household objects.

The core of the GAT model lies in its ability to dynamically adjust the influence of neighboring nodes on the current node by learning the attention weights between nodes. For node v_i in the household object relationship graph, GAT first calculates the attention coefficients between it and its neighboring node $v_i \in N(v_i)$, and then updates the feature vector of node v_i by weighting the aggregated neighboring node's features according to these coefficients. GAT updates the feature vector of node v_i to the next level $(l+1)$ by the following formula:

$$h_{v_i}^{(l+1)} = \sigma\Big(\sum_{v_j \in N(v_i) \cup v_i} \alpha_{ij}^{(l)} W^{(l)} h_{v_j}^{(l)}\Big) \tag{1}$$

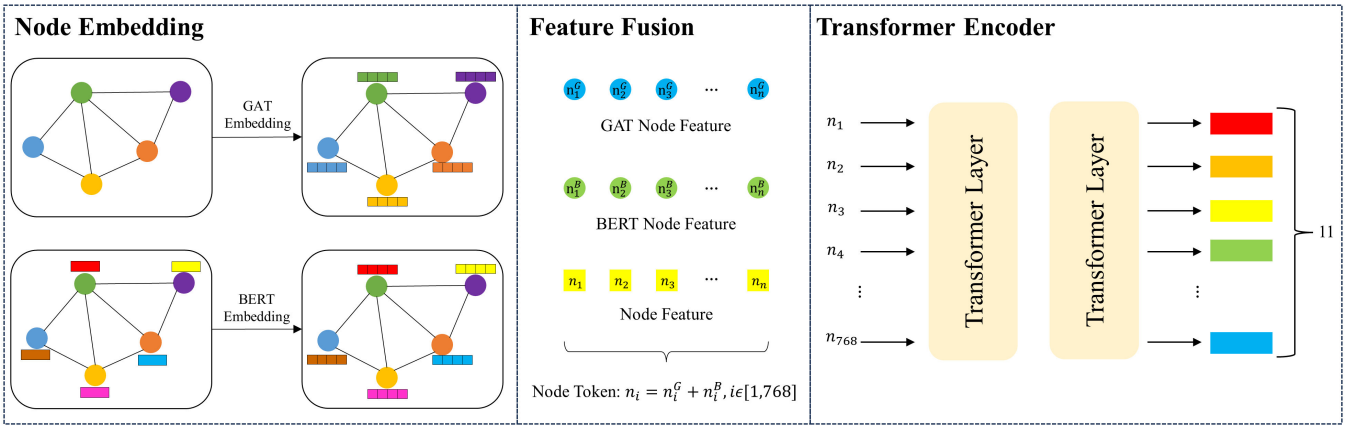

Fig. 5. Method overview: The workflow integrates GAT to extract topological features from the object relationship graph and BERT for ontological features of objects. These combined features are then fed into a Transformer to facilitate the training of the object relationship classification model.

Where $h_{v_i}^{(l+1)}$ represents the feature vector of node v_i in layer $l+1$, $\alpha_{ij}^{(l)}$ denotes the attentional weight between nodes v_i and v_j in layer l, $W^{(l)}$ is the weight matrix of layer l, and σ is the nonlinear activation function ReLU.

The computation of the attention weight $\alpha_{ij}^{(l)}$ is based on the feature vectors of nodes v_i and v_j, which are normalized in this paper using a neural network and a Softmax function:

$$\alpha_{ij} = \frac{exp(ReLU(a^T[Wh_{v_i}||Wh_{v_j}]))}{\sum_{k\in N(v_i)\cup v_i} exp(ReLU(a^T[Wh_{v_i}||Wh_{v_k}]))} \quad (2)$$

Here a represents the learned attention vector and $||$ represents the stitching of the vectors.

The main advantage of using GAT for graph embedding lies in its ability to dynamically learn weights between nodes, effectively capturing both node features and complex relationships. For the household object relationship graph, this allows the model to adjust the impact of interactions based on the actual strengths between objects, thereby providing a more accurate and enriched feature representation for tasks like object relationship prediction.

B. Feature Fusion and Relationship Classification

This section details how to effectively combine graph topology feature vectors with object ontological property feature vectors, using the Transformer model to accurately classify the fused vectors for object relationships.

First, we encode the ontological properties of objects semantically using the BERT model, a pre-trained deep bidirectional Transformer that captures the nuances in object property descriptions through its rich semantic capabilities, converting these descriptions into high-dimensional vectors. Next, we merge these ontological property vectors from BERT with graph topology vectors from GAT. The vectors are concatenated to create an extended, integrated feature vector. This fusion process is depicted in Formula (3).

$$F_{Final} = [F_{GAT}; F_{BERT}] \quad (3)$$

Where F_{GAT} and F_{BERT} represent the graph topology and object ontology property feature vectors, respectively, and F_{Final} is the fused feature vector.

This approach maintains the independence and integrity of the respective features while significantly enhancing the model's capability to understand and characterize complex object relationships by integrating diverse information sources.

Additionally, we employ the Transformer model architecture for classifying object relationships. Our Transformer model comprises an embedding layer, two Transformer encoding layers, and an output layer. The embedding layer processes the fused feature vectors, which are then refined in the encoding layers through self-attention mechanisms and positional encoding to capture the intricate dependencies and sequence information among features. The output layer predicts relationship classes based on these encoded vectors.

During training, we optimize the model parameters using the cross-entropy loss function and the Adam optimizer. The training dataset includes fused feature vectors and their corresponding relationship labels from the HORDD.

C. Object Relationship Reasoning

This section introduces a novel method for reasoning about object relationships, integrating YOLOv8, CLIP, an ontology knowledge base, and an object relationship classification model. This approach aims to accurately identify objects and their properties in domestic settings and infer the relationships between them based on these properties, as illustrated in Figure 6.

Initially, the YOLOv8 model identifies objects in a scene, represented as $O = o_1, o_2, ..., o_n$. These objects are then matched with categories in a custom-built ontology knowledge base to extract their properties, denoted as $P_i = p_1, p_2, ..., p_m$ for each object o_i, covering attributes like size, color, and function. To enhance property identification accuracy, the CLIP model is used to assess semantic similarities between visual and textual information, improving property selection to match detected objects through its capability

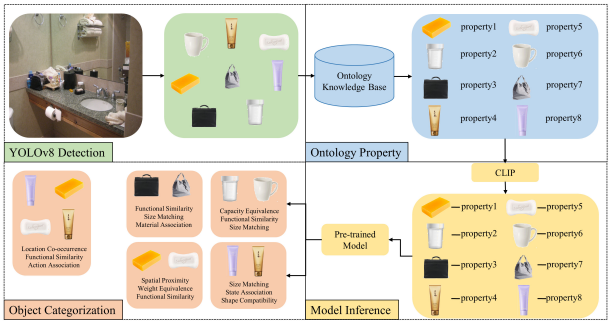

Fig. 6. Detailed process of object relationship inference, showing (1) object detection using YOLOv8, (2) property information retrieval via an ontology knowledge base, (3) property refinement through CLIP by evaluating semantic similarity, and (4) relationship inference between objects using the pre-trained classification model.

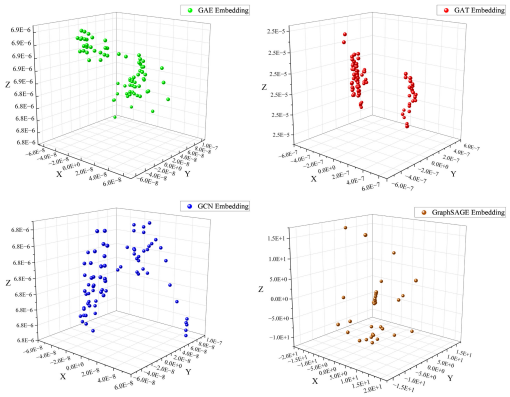

Fig. 7. Visualization of node feature vectors generated by graph embedding.

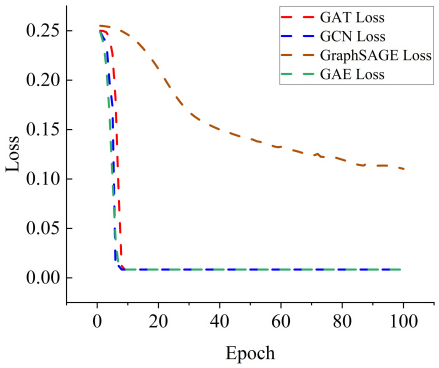

Fig. 8. Loss values during training of the graph embedding.

to link images with text. This process is depicted as a refinement function R, where $R(P_i)$ refines the properties to $P_i^{'}$.

The final step involves a pre-trained object relationship inference model, which uses the names and refined properties of objects to determine their relationships, such as spatial proximity or functional connections. This model inputs the refined properties $P_i^{'}$ and $P_j^{'}$ of any two objects o_i and o_j to predict their relationship r_{ij}. This method, integrating detailed property information, provides a robust framework for accurately determining real-world object relationships in any environment.

V. EXPERIMENT

In our experiments, we used a computing setup that included an Intel Core i9-9900X CPU at 3.50GHz and an NVIDIA GeForce RTX 3090 GPU. Our software environment ran on Ubuntu 18.04 and utilized the Pytorch framework for training and testing deep learning models. For the graph embedding model training, we maintained a learning rate of 0.001 across 100 epochs to ensure comprehensive learning and model stability. During the Transformer model training, we reduced the learning rate to 2e-5 and used two Transformer encoder layers with an 8-head attention mechanism to enhance the model's ability to discern and predict relationships within the graph structure. In real-world applications, we employed the Realsense D435i camera and the UR3 robotic arm for object manipulation.

A. Graph Embedding Evaluation

In this paper, we used four graph embedding methods to generate 768-dimensional feature vectors for nodes within our household object relationship graph, which we then reduced to three dimensions using PCA for visualization (Figure 7) and evaluated training performance through loss trends (Figure 8). The analysis revealed that the GraphSAGE model underperformed in loss reduction and feature vector dispersion post-reduction, making it unsuitable for feature extraction in our graph. In contrast, the GAT model demonstrated superior performance in categorizing feature vectors and capturing node categories, and was therefore selected for its robust feature extraction and relationship mapping capabilities.

TABLE II
PERFORMANCE COMPARISON OF GRAPH EMBEDDING ALGORITHMS FOR OBJECT RELATIONSHIP CLASSIFICATION

Model	Train Acc	Val Acc	Train Loss	Val Loss
GCN	94.79%	85.35%	**0.19**	0.87
$GraphSAGE$	94.34%	87.20%	0.20	0.76
GAE	94.76%	85.76%	0.20	0.84
GAT_{base}	**94.84%**	**87.62%**	0.21	**0.71**

Further comparative analysis highlighted the GAT model's superior capability in generating node feature vectors suitable for training an object relationship classification model, as shown in Table II. The implemented GAT model, referred to as GAT_{base}, includes two layers of GATConv and a single attention head, demonstrating its effectiveness in producing high-quality node feature vectors. This selection is based on the GAT model's significant performance advantages and its ability to capture complex node relationships, making it ideal for feature extraction within the household object relationship graph.

B. Relationship Classification Evaluation

Our experiment assessed the impact of varying GAT configurations on model performance. We altered the number of GATConv layers and attention heads, comparing a base model (2 layers, 1 head) with larger configurations (4 layers, with 1 or 4 heads) to gauge their effect on node feature vector generation, using accuracy and loss metrics for evaluation. The results are detailed in Table III.

TABLE III
IMPACT OF GATCONV LAYERS AND ATTENTION HEADS ON MODEL PERFORMANCE

Model	Layer	Head	Train Acc	Val Acc	Train Loss	Val Loss
GAT_{base}	2	1	**94.84%**	**87.62%**	**0.21**	**0.71**
GAT_{base}	2	4	94.60%	87.48%	**0.21**	0.74
GAT_{large}	4	1	94.35%	87.20%	0.22	0.80
GAT_{large}	4	4	94.00%	87.34%	0.22	0.75

Otherwise, we compared our method with prevalent NLP models—Long Short-Term Memory (LSTM) [19], Bi-directional Long Short-Term Memory (BiLSTM) [20], Text Convolutional Neural Network (TextCNN) [21], and Bidirectional Encoder Representations from Transformers (BERT) [16] —to evaluate their performance in text classification tasks across accuracy and loss. The result is shown in Table IV.

TABLE IV
COMPARISON OF MODEL PERFORMANCE ON TEXT CLASSIFICATION

Model	Train Acc	Val Acc	Train Loss	Val Loss
TextCNN [21]	23.82%	24.89%	2.88	2.90
LSTM [19]	37.73%	34.00%	2.31	2.69
BiLSTM [20]	45.79%	36.13%	1.87	2.65
BERT [16]	92.31%	83.47%	0.53	0.67
ours	**94.84%**	**87.62%**	**0.21**	**0.71**

The comparative evaluation showed that our GAT model, particularly the GAT_{base} configuration with its 2-layer, single-head architecture, outperformed both its more complex GAT_{large} variants and conventional NLP models across key performance metrics including training and validation accuracy, as well as loss. This demonstrates our model's consistent superiority, benefiting from incorporating topological structure information from the Household Object Relationship Graph, which allows a more accurate depiction of relationships among household objects. Notably, performance in graph neural network models like GAT can degrade with increased complexity due to overfitting or challenges in training deeper networks effectively. These findings confirm that our GAT-based approach, leveraging attention mechanisms and graph-based learning principles, not only improves text classification performance but also highlights the benefits of streamlined model architectures over more complex or traditional frameworks. This study underscores the strategic advantage of optimally configured GAT models, advocating their use in enhancing accuracy and efficiency in NLP tasks.

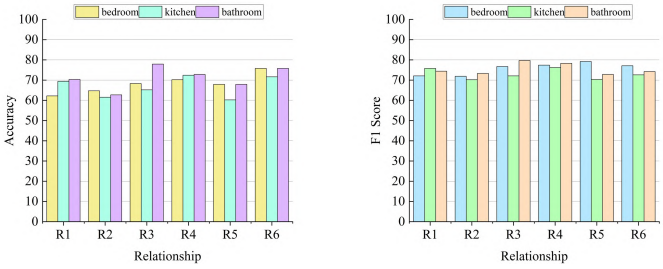

Fig. 9. Performance of the model in three environments

C. Cross-Environment Model Evaluation

To thoroughly evaluate our object relationship inference model, we conducted an extensive assessment across three domestic settings: kitchen, bedroom, and bathroom. This approach tested the model's ability to generalize across different environments. We used accuracy and F1 score as primary metrics to gauge the model's classification performance in these varied settings. The results, shown in Figure 9, focus on the 6 most common relationships identified in Figure 3.

D. Real Object Oiganization

In the experiment, we validate our model's capability to understand complex relationships among household objects. Figure 10 demonstrates the effectiveness of our approach. Using the object relation reasoning method outlined in section IV.C, we illustrate how the Shape Compatibility relationship categorizes objects on a table into two groups. Additionally, we conducted two experiments with different organizational strategies to showcase the model's versatility. The first experiment organized books based on shape and color in one group and size and shape in another, testing the model's visual and spatial handling (Figure 11). The second experiment organized household objects by spatial and functional properties in one group and by state and function in another, exploring how these factors influence storage logic and convenience (Figure 12). These tests aimed to validate our model's understanding of functional relationships and spatial arrangements in domestic environments, providing practical organizing insights.

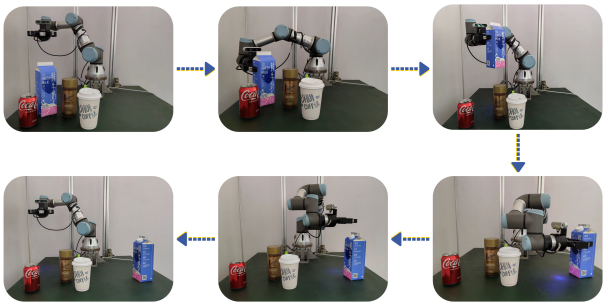

Fig. 10. Among the four objects, one is rectangular and the others are cylindrical. The UR3 robotic arm selects the rectangular object, successfully organizing the objects by shape.

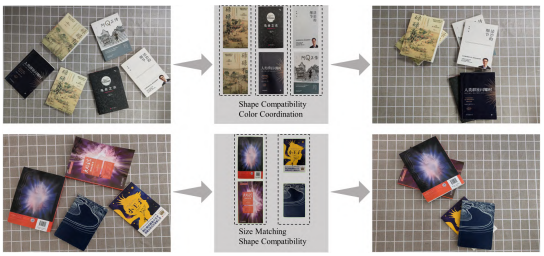

Fig. 11. Experiment on book organization. Books are organized into categories based on *Shape Compatibility*, *Color Coordination* and *Size Matching*.

Fig. 12. Experiment on common household object organization. Objects are organized into categories based on *Spatial Proximity*, *Function Similarity* and *State Association*.

The results highlighted our model's strong potential in interpreting and applying complex object relationships across diverse home environments. The experiments demonstrated its effectiveness in various scenarios and its ability to provide actionable organizational advice, presenting a novel approach to automated household organization.

VI. CONCLUSION AND FUTURE WORK

In this paper, we introduce an innovative method for organizing household objects using a Transformer-based model that integrates graph topology with object ontology to precisely identify object relationships. We start by creating a comprehensive dataset specifically designed for detecting household object relationships. Next, we develop a novel feature extraction technique combining Graph Attention Network (GAT) with BERT to merge graph topology and object ontology. Finally, we present an inference method that integrates YOLOv8, CLIP, and a carefully constructed knowledge base. Extensive experiments demonstrate the method's effectiveness and reliability, significantly enhancing service robots' ability to predict and organize object relationships with high accuracy.

For future work, we plan to expand our dataset to include a wider range of household objects from various environments and integrate multimodal data sources, such as visual, auditory, and textual inputs. This will improve the model's adaptability and robustness, offering a more comprehensive understanding of object relationships and providing practical, user-friendly organization strategies for service robots.

REFERENCES

[1] Yiqing Xu and David Hsu. How to tidy up a table: Fusing visual and semantic commonsense reasoning for robotic tasks with vague objectives. *arXiv preprint arXiv:2307.11319*, 2023.

[2] Ivan Kapelyukh, Vitalis Vosylius, and Edward Johns. Dall-e-bot: Introducing web-scale diffusion models to robotics. *IEEE Robotics and Automation Letters*, 2023.

[3] Zirui Zhao, Wee Sun Lee, and David Hsu. Differentiable parsing and visual grounding of natural language instructions for object placement. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11546–11553. IEEE, 2023.

[4] Ivan Kapelyukh and Edward Johns. My house, my rules: Learning tidying preferences with graph neural networks. In *Conference on Robot Learning*, pages 740–749. PMLR, 2022.

[5] Xian Mo, Rui Tang, and Hao Liu. A relation-aware heterogeneous graph convolutional network for relationship prediction. *Information Sciences*, 623:311–323, 2023.

[6] Zhiguang Fan, Yuedong Yang, Mingyuan Xu, and Hongming Chen. Node-based knowledge graph contrastive learning for medical relationship prediction. *arXiv preprint arXiv:2310.10138*, 2023.

[7] Walter Goodwin, Sagar Vaze, Ioannis Havoutis, and Ingmar Posner. Semantically grounded object matching for robust robotic scene rearrangement. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 11138–11144. IEEE, 2022.

[8] Weiyu Liu, Chris Paxton, Tucker Hermans, and Dieter Fox. Structformer: Learning spatial structure for language-guided semantic rearrangement of novel objects. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6322–6329. IEEE, 2022.

[9] Nichola Abdo, Cyrill Stachniss, Luciano Spinello, and Wolfram Burgard. Robot, organize my shelves! tidying up objects by predicting user preferences. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 1557–1564. IEEE, 2015.

[10] Sina Sajadmanesh, Sogol Bazargani, Jiawei Zhang, and Hamid R Rabiee. Continuous-time relationship prediction in dynamic heterogeneous information networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(4):1–31, 2019.

[11] Hanwen Liu, Lianyong Qi, Shigen Shen, Arif Ali Khan, Shunmei Meng, and Qianmu Li. Microservice-driven privacy-aware cross-platform social relationship prediction based on sequential information. *Software: Practice and Experience*, 54(1):85–105, 2024.

[12] Zewen Wu, Jian Tang, Xingyu Chen, Chengzhong Ma, Xuguang Lan, and Nanning Zheng. Prioritized planning for target-oriented manipulation via hierarchical stacking relationship prediction. *arXiv preprint arXiv:2303.07828*, 2023.

[13] Yafu Tian, Alexander Carballo, Ruifeng Li, and Kazuya Takeda. Rsg-net: Towards rich sematic relationship prediction for intelligent vehicle in complex environments. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 546–552. IEEE, 2021.

[14] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

[15] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[19] Long Short-Term Memory. Long short-term memory. *Neural computation*, 9(8):1735–1780, 2010.

[20] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

[21] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.