Interactive Visual Task Learning for Robots

Weiwei Gu, Anant Sah, Nakul Gopalan

School of Computing and Augmented Intelligence, Arizona State University
{weiweigu, asah4, ng}@asu.edu

Abstract

We present a framework for robots to learn novel visual concepts and tasks via in-situ linguistic interactions with human users. Previous approaches have either used large pre-trained visual models to infer novel objects zero-shot, or added novel concepts along with their attributes and representations to a concept hierarchy. We extend the approaches that focus on learning visual concept hierarchies by enabling them to learn novel concepts and solve unseen robotics tasks with them. To enable a visual concept learner to solve robotics tasks one-shot, we developed two distinct techniques. Firstly, we propose a novel approach, Hi-Viscont(HIerarchical VISual CONcept learner for Task), which augments information of a novel concept to its parent nodes within a concept hierarchy. This information propagation allows all concepts in a hierarchy to update as novel concepts are taught in a continual learning setting. Secondly, we represent a visual task as a scene graph with language annotations, allowing us to create novel permutations of a demonstrated task zero-shot in-situ. We present two sets of results. Firstly, we compare Hi-Viscont with the baseline model (FALCON) on visual question answering(VQA) in three domains. While being comparable to the baseline model on leaf level concepts, Hi-Viscont achieves an improvement of over 9% on non-leaf concepts on average. Secondly, we conduct a human-subjects experiment where users teach our robot visual tasks in-situ. We compare our model's performance against the baseline FALCON model. Our framework achieves 33% improvements in success rate metric, and 19% improvements in the object level accuracy compared to the baseline model. With both of these results we demonstrate the ability of our model to learn tasks and concepts in a continual learning setting on the robot.

Introduction

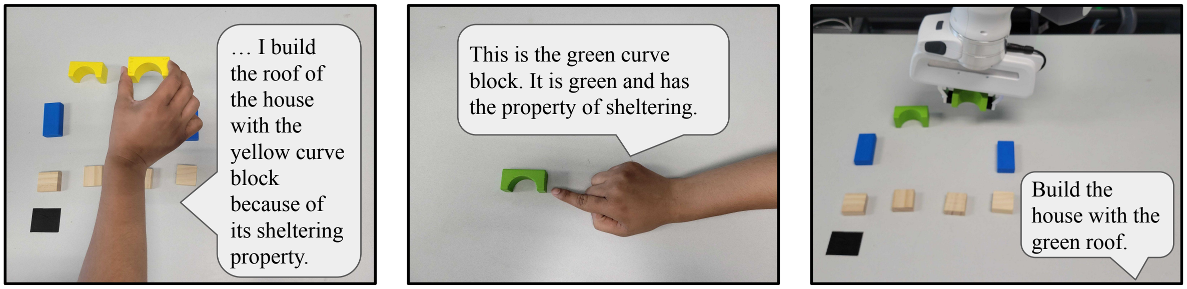
Robots in a household will encounter novel objects and tasks all the time. For example, a robot might need to use a novel vegetable peeler to peel potatoes even though it has never seen, let alone used such a peeler before. Our work focuses on teaching robots novel concepts and tasks one-shot via human-robot interactions, which include demonstrations and linguistic explanations. We then want the robot to generalize to a similar but unseen visual task. A robotic system that can learn generalizable tasks and concepts from few natural interactions from a human-teacher would represent a large leap for robotics applications in everyday settings. In this work we aim to take a step in the direction of generalizable interactive learning as demonstrated Fig. 1.

Previously, large image and language models have been extended to robotics to manipulate novel objects, and create visual scenes (Shridhar, Manuelli, and Fox 2021; Brohan et al. 2023a). These methods recognize novel objects by using their underlying large language and visual models to extract task-relevant knowledge. However, they are not capable of learning to create a novel visual scene from in-situ interactions with a human user. There is also significant work in few-shot learning of visual concepts in computer vision (Mei et al. 2022; Snell, Swersky, and Zemel 2017; Vinyals et al. 2017; Sung et al. 2018; Wang, Ye, and Gupta 2018; Tian et al. 2020), albeit without extensions to robotics domains. These approaches focus on learning novel concepts for image classification, but ignore the fact that the novel concepts also bring new information to update our understanding of concepts already known to the robot. The reverse path of knowledge propagation, that is, from novel concepts to previously known concepts is equivalently important in performing tasks in the real-life scenarios, especially when the agent has little knowledge of the world and needs to continually add information to known concepts.

In this work, we propose a novel framework, Hi-Viscont, that enables robots to learn visual tasks and visual concepts from natural interactions with a human user. We learn the task type and concepts from users one-shot, and then generalize to novel task variants within the task type zero-shot. We do this by connecting our insights on *one-shot visual concept learning* and the use of *scene graphs*. The robot learns the structure of a visual task by converting linguistic interactions with a human user into a scene graph with language annotations. Moreover, Hi-Viscont updates parental concepts of the novel concept being taught. Such updates allow us to generalize the use of the novel concepts in to solve novel tasks.

The contribution of this work is three-fold:

1. We present concept learning results on VQA tasks that are comparable to the state-of-the-art FALCON model. More specifically, Hi-Viscont improves on FALCON on all non-leaf concepts across all domains with significance.
2. We enable the robot agent to learn a visual task from in-situ interactions with a scene graph, allowing zero-


… I build the roof of the house with the yellow curve block because of its sheltering property.
This is the green curve block. It is green and has the property of sheltering.
Build the house with the green roof.

(a) (b) (c)

Figure 1: This figure demonstrates how Hi-Viscont learns from users interactively. (a) First, the user demonstrates a structure, say a "house," with its sub-components such as a "roof" and the concepts used to make the "roof" such as a "yellow curve block". (b) The user then teaches a novel concept such as a "green curve block" and describes its properties. (c) The user can now ask the robot to create a new structure ("house with green roof") zero-shot with the taught component without explicitly asking for the object of interest.

shot generalization to an unseen task of the same type, as demonstrated in Fig. 1.

3. Finally, we conduct a human-subjects experiment to show that our system is able to learn visual tasks and concepts from in-situ interactions with human users. Hi-Viscont achieves a 33.33% improvement in Success Rate when completing the users' requests compared to FALCON ($p = 0.014$).

Related Work

Language conditioned manipulation. Significant work has been performed in learning concepts and tasks for robots in interactive settings (Gopalan et al. 2018, 2020; Tellex et al. 2020) even with the use of dialog (Chai et al. 2018; Matuszek et al. 2012). Our work differs from previous works as it learns visual concepts for manipulation one-shot, and improves generalization by updating other known concepts. Moreover, our approach can learn a concept hierarchy starting from zero known concepts, displaying the adaptability of our model under a continual learning setup. Previous work has focused on language conditioned manipulation (Shridhar, Manuelli, and Fox 2021; Liu et al. 2021; Brohan et al. 2023b,a). Shridhar, Manuelli, and Fox 2021 computes a pick and place location conditioned on linguistic and visual inputs. Liu et al. 2021 focuses on semantic arrangement on unseen objects. Other works train on large scale linguistic and visual data and can perform real-life robotic task based on language instructions (Ahn et al. 2022; Brohan et al. 2023b,a). Our work focuses on interactive teaching of tasks and concepts instead of focusing on the emergent behaviors from large models. Daruna et al. 2019 learns a representation of a knowledge graph by predicting directed relations between objects allowing a robot to predict object locations. To the best of the author's knowledge, our work is the first that learns concepts and tasks one-shot to generalize to novel task scenarios on a robot, making our contributions significant compared to other related works.

Visual reasoning and visual concept learning. Our work is related to visual concept learning (Mei et al. 2022; Mao et al. 2019; Yi et al. 2019; Han et al. 2020; Li et al. 2020) and visual reasoning (Mascharka et al. 2018; Hu et al. 2018; Johnson et al. 2017; Hudson and Manning 2018). To perform the visual reasoning task, traditional methods (Mascharka et al. 2018; Hu et al. 2018; Johnson et al. 2017; Hudson and Manning 2018) decompose the visual reasoning task into visual feature extraction and reasoning by parsing the queries into executable neuro-symbolic programs. On top of that, many concept learning frameworks (Mei et al. 2022; Mao et al. 2019; Yi et al. 2019; Han et al. 2020; Li et al. 2020) learn the representation of concepts by aligning concepts onto objects in the visual scene. As far as we know, Mei et al. 2022's FALCON is the most similar work to our work in this line of research. However, when introducing a new concept, our work continually updates the representation of all related concepts, whereas Mei et al. 2022 does not, which makes it ill-suited for continual learning settings. Our work is also related to the area of few-shot learning (Snell, Swersky, and Zemel 2017; Tian et al. 2020; Vinyals et al. 2017), which learns to recognize new objects or classes from only a few examples but does not represent a concept hierarchy which is useful in robotics settings.

Scene graph. Scene graphs are structural representations of all objects and their relationships within an image. The scene graph representation (Chang et al. 2023) of images is widely used in the visual domains for various tasks, such as image retrieval(Johnson et al. 2017), image generation(Johnson, Gupta, and Fei-Fei 2018), and question answering(Teney, Liu, and van den Hengel 2017). This form of representation has also been used in the robotics domains for long-horizon manipulation (Zhu et al. 2021).

Methods

We first present the baseline FALCON model and then introduce our Hi-Viscont model. Our model is based on concept learners as they learn concepts few shot, and can reason over the attributes of chosen (and their parent) concept classes. FALCON is a State-of-the-art (SOTA) concept learner which learns novel concepts one-shot.

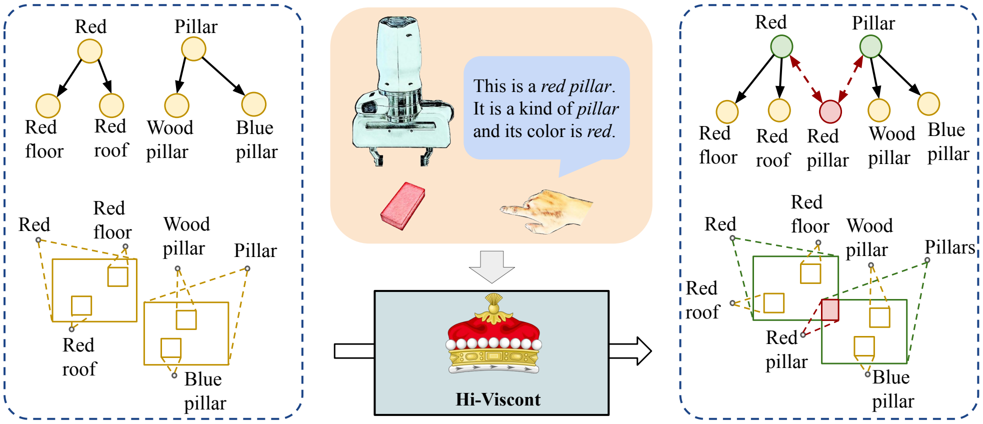


Figure 2: We demonstrate the updates to the box embedding space and the parent concepts when a novel concept is taught to our robot using Hi-Viscont. Existing approaches only edit the leaf nodes as those represent novel concepts.

FALCON

Mei et al. (2022) developed FALCON, a meta-learning framework for one-shot concept learning in visual domains. FALCON learns a new visual concept with one or a few examples, and uses the learned concept to answer visual reasoning questions on unseen images. There are three components for the model: a visual feature extractor that extracts the object-centric features for the input image, a graph neural network (GNN) based concept learner, and a neuro-symbolic program executor that executes the input neuro-symbolic program.

Natural language sentences describing objects and their queries are represented as structured neuro-symbolic programs. FALCON learns novel concepts by interpreting the images presented and the relationships between known concepts and the unknown concept being learned using a neuro-symbolic program. After learning, the model performs reasoning over questions, executed as neuro-symbolic programs to answer questions about images.

A pre-trained ResNet-34 is used as visual feature extractor for the model. The visual feature extractor computes a feature for each object in a scene separately, which can be used for downstream visual reasoning. FALCON uses a box embedding(Vilnis et al. 2018) to represent concepts and their object visual features.

Finally, the concept learning module is composed of two separate Graph Neural Networks(GNNs), the Relation GNN and the Example GNN. To compute a representation for a novel concept c, FALCON starts with an embedding that is randomly sampled from a prior Dirichlet distribution. Then, the model updates the representation of c by computing messages from parent nodes based on their factor weights or relationship and also computing a message from the visual feature (represented as a node within the Example GNN) for the concept being learned. This representation for the novel concept c will then be used for downstream VQA tasks.

There are two major issues to directly use FALCON for interactive task learning on the robot. Firstly, the model lacks scene information to solve tasks. We address this in our work. Secondly, the model assumes concepts are learned perfectly and do not need to be updated as it learns more concepts. For example, when we teach the model the concept of “container” with an image of a “cup,” FALCON cannot update the features of the “container” concept when the concept of “bowl” is taught as a child to the “container” concepts. This allows FALCON to learn that all “containers” have handles which is untrue.

Hi-Viscont

We present our concept net model, Hi-Viscont (HIerarchical VISual CONcept learner for Task), which actively updates the related known concepts when we introduce the novel concept to improve upon FALCON’s generalization capabilities. We adopted several modules from the framework of FALCON, including the visual feature extractor, the neuro-symbolic program executor, the box embedding space, and the novel concept learner. Moreover, we introduce an additional GNN module, Ancestor Relational GNN (ARGNN), that updates the related known concepts as a novel concept is introduced. ARGNN predicts a new embedding for the related known ancestor concepts to the novel concept. To do this update we compute a message from the visual feature of novel concept’s instance to the embedding of the related nodes using the relations between the parent concepts and the novel concept.

When a novel concept c is inserted to Hi-Viscont, the extracted visual feature o_c of concept c and its relations with known concepts R_c are fed to Hi-Viscont as input. Each relation $rel = (c', c, r)$, where c' denotes the related concept, and r describes its relationship with c. We compute an embedding e_c for novel concept c using the same method as FALCON. Using the additional ARGNN, we predict a new embedding for each related concept c' by computing a message from the visual feature o_c to the current embedding of the related concept $e^0_{c'}$ using the same relationship rel. The formula for this update is denoted as follows:

$$e^1_{c'} = \text{ARGNN}(o_c, rel, e^0_{c'})$$

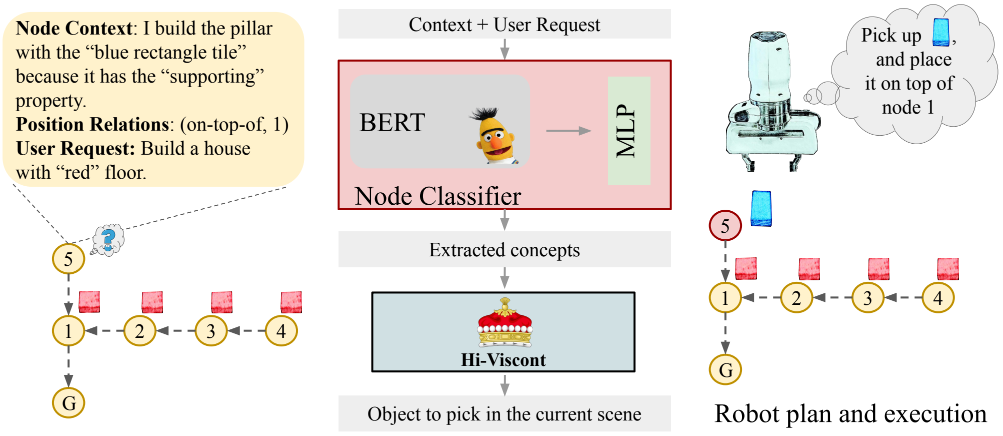

Node Context: I build the pillar with the "blue rectangle tile" because it has the "supporting" property.
Position Relations: (on-top-of, 1)
User Request: Build a house with "red" floor.
5
1
2
3
4
G
Context + User Request
BERT
MLP
Node Classifier
Extracted concepts
Hi-Viscont
Object to pick in the current scene
Pick up, and place it on top of node 1
5
1
2
3
4
G
Robot plan and execution

Figure 3: Our pipeline decides objects for each node in the scene graph one at a time. The node's context and the request phrase are fed into a node classifier, which is composed of a BERT encoder and an multilayer perceptron, to decide the concepts applicable for the current node. Hi-Viscont then decides the object to pick in the current scene based on the extracted concepts. In this example, the object chosen for Node 5 is "blue rectangular tile" as it existed in the original demonstration and was not changed given the novel task's linguistic request. Notice that the new structure has red floor tiles which were never demonstrated to the robot.

The resulted embedding $e^1_{c'}$ will be used as the representation for concept c' for future task or updates.

To provide gradient flow to train ARGNN, we extended the concept learning task proposed by FALCON by adding validation questions for each related concept, that is when a new concept is added all concepts in the concept net are tested for accuracy over the novel concept. For example, from our previous discussion the newly inserted "bowl" concept's object instance is checked with the "container" parent to see if the presented "bowl" also tests as a "container." A more detailed description of our training pipeline and methodology can be found in the appendix. While FALCON was evaluated solely on the newly inserted concept, we evaluate all concepts (leaf and parent nodes) of our model on unseen images. Such an evaluation ensures consistency between parent and child concepts which is a necessity in continual learning settings. This evaluation mechanism allows us to evaluate the quality for the embedding of all concepts in the resulting knowledge graph, which is closer to how these knowledge are used in the real world setting.

Learning Visual Task via Scene Graph

Figure 1 illustrates how our pipeline learns a visual task from a single in-situ interaction with a human user. The user's demonstration(Fig. 1.a) is first converted into an initial scene graph. Each node of the initial scene graph corresponds to an object that the user placed, and it contains the bounding box information of the object and the user's linguistic description of the object. We also store the positional relations with respect to other nodes for each node, which allows for object placements when reconstructing the scene. A fixed location on the table is marked with black tape as the origin, which is treated as the zeroth object. All other objects placed by the user will be to the top right of the origin.

Based on the initial scene graph and the user's linguistic request for the desired variant of the visual scene, we infer a goal scene graph modelled as a node-wise classification task as shown in Figure 3. Since the variant of the visual task from the user request shares the same structure as the demonstration, the goal scene graph has the same number of nodes as the initial scene graph. We take the user's description of the corresponding node of the initial scene graph t_i and the user's request of the variant of the structure q as inputs, and perform a two-step inference: First we decide if the node in the goal graph is different from the one in demonstration; Subsequently if the node is different we decide with classification which object satisfies the node location.

To decide whether the concept of a node within the scene has changed given the user's description of the node and the user's current request q, we perform a binary classification at each node. The result of this classification decides if we are changing a node's concept or not. We use a pretrained $\text{BERT}_{\text{base}}$ model to encode the context request pair, which is then fed into a multi-layer perceptron (MLP) with a Cross-Entropy loss. The second step of the inference extracts the related concepts from the context if the node's concept needs to be changed as per the request. We convert the concept extraction problem into a classification problem by providing concept candidates as a part of the input again with BERT model and an MLP with a Cross-Entropy loss. The related concepts of each node are fed as input for the concept net model to decide the object to pick, and the positional relations with other nodes are used to compute the placement location. The robot reconstructs the scene following the order of the nodes. Pairing the concept net model with scene graph, the robot is able to learn the arrangement of a scene in one single demonstration and perform variants of the scene without demonstration. We allow FALCON access to our Scene Graph classifiers to have a valid baseline to compare against.

Robotics Setup for User Study

We integrate our visual task learning and concept learning model with a Franka Emika Resarch 3 arm (FR3). This pipeline allows us to show the generalizability with which Hi-Viscont learns visual concepts when compared to Falcon (Mei et al. 2022) in learning and solving novel tasks. To set this demonstration up we use a Franka Emika Research 3 arm (FR3), two calibrated realsense D435 depth cameras, and a mono-colored table to allow for background subtraction. We use the SAM (Segment Anything Model) (Kirillov et al. 2023) to separate the foreground and the background and get individual bounding boxes for each of the blocks on the table. For pick and place, we initially experimented with Transporter networks (Zeng et al. 2021) but finally used a simpler Visuo-Motor Servoing mechanism for reliability. We expected users to maintain about an inch of space between each object in the scene to allow the robot to pick objects without collisions and for SAM to segment objects from the background accurately. In the process of picking and placing, the robot autonomously recovers if an error is made. Once an object is grasped, we place it into the Task scene based on the position calculated relatively with respect to the previously placed object nodes or the zeroth origin object. This process is done iteratively until the whole scene graph is completed.

Human-Subjects Experiment

Study Design We conduct a 1 × 2 within-subject experiment to measure the framework's ability to learn visual task and visual concepts from in-situ interaction. We extend the FALCON model with our scene graph module and use it as a baseline to compare against because we could not find any equivalent prior work. Both concept net models are trained with the same split of concepts and the same training data for the same number of steps. Through this experiment, we aim to demonstrate that our framework achieves better performance than FALCON model because of the continual update for the known knowledge. Participants for our experiment interact with the robot in three phases. For each interactive phase, the participants only interact with the robot once, and the interaction is recorded as the input for both systems. After the interactive phases, the participant observes the two systems construct the scene requested by the participant. Half the participants observe FALCON first and other half observe Hi-Viscont system first to avoid ordering confounds.

Human Subjects Experiment Domain We evaluate our approach with the human-subjects experiment in a 2-D object rearrangement domain, which is a problem commonly used in language grounding and HRI research (Liu et al. 2021; Shridhar, Manuelli, and Fox 2021). The domain we choose for this study is the House-Construction domain which we introduce in Domains Section. We designed this domain as the users have the ability to create complex types of structures with different object classes. Building blocks from children's toys were used as objects in this domain as they are varied and easy to grasp by the robot, as grasping is not a focus of our work.

Metrics The objective metrics we collect for the human-subjects experiment are as follows. We measure the success rate (SR) of completing the user's request with complete accuracy, and the node level accuracy of each scene graph for both systems. Both metrics are used to measure each system's ability to actually complete the visual task objectively. The success rate metric gives us the insight of system's ability of completing the whole task, while the node level accuracy metric provides a more fine-grained result on few-shot object recognition. In the post-study survey for each system, we administer the Perceived Intelligence and Anthropomorphism sub-scales of the Godspeed Questionnaire Series (Bartneck et al. 2009), Trust in Automated systems questionnaire (Jian, Bisantz, and Drury 2000), System Usability Scale (SUS)(Brooke 1995). In addition, we handcrafted a direct comparison survey for preference between Hi-Viscont and the FALCON model.

Procedure This study was approved by our university's Institutional Review Board (IRB). We recruited all participants through on campus advertisements. The study took under 90 minutes with voluntary participation. The participants were not compensated for their efforts. The procedure of the study is as follows. Participants first fill out consent form and then the pre-study survey. After the pre-study survey, we hand out a general introduction for the experiment of the study. Then, we guide the participant through the task teaching phase, the concept teaching phase, and the request phase sequentially as described below. Before each phase, we provide a demonstration video and the instructions of the corresponding phase to the participants. The instruction videos is demonstrating a different task (bridge making) with different objects (foam blocks) than the actual task the participant is teaching. The anonymized instruction manual and the link to these videos are provided in the Appendix and the associated webpage [1].

Task Teaching Phase - In the task teaching phase, the participants teach the robot a visual task by demonstrating the scene with its constituent structures one by one. The participants also describe the structures and the objects used to build the structure with natural language. For example, a participant might build a house, with floors, pillars and a roof. While building the roof the participant might say "This a roof which I build with the curved blue tile because of it's sheltering capability." The participants demonstrate each of the structures with their chosen language commands one after another to build a house. We record all descriptions in audio and convert them into text using audio to text tools.

Concept Teaching Phase - In this phase, the participants teach a novel concept of their choice to both systems by showing the object to the camera, and describing the concept's properties, such as the color and functional characteristics of the object, in natural language. The description to the novel object concept is converted to neuro-symbolic programs, which are given to both models for updates as described in the Methods section.

Request Phase - In the request phase, the participants are asked to provide a request in natural language for a novel

[1] https://sites.google.com/view/ivtl

Method	CUB-200-2011	House Construction	Zoo
Hi-Viscont	74.39±7.04	86.41±5.28	83.50±8.44
FALCON	73.40±5.77	87.17±4.17	85.12±6.64

Table 1: The average F1 score and standard deviation of Hi-Viscont and FALCON on the test concepts across all the three domains.

scene that they did not demonstrate in the task teaching phase. They are instructed to use the object taught in the concept teaching phase in the request. The requested task still needs to be a house which is the same task type as the demonstration, but not a house that the models have seen.

After completing the three phases, the participants watch the two systems construct their requested scene in real time with a robot in a randomized order. As each system finishes the construction, the participants are asked to fill a post-survey for that system. After both systems finish the construction, the participants also fill out a comparative survey.

Hypotheses We have the following hypotheses:

Hypothesis 1 - Our framework will have a higher success rate in completing the user's request without any errors. We hypothesize that our framework will have a higher success rate than FALCON model because of its update to the related known concepts, which could be used in requests that point to the same object indirectly.

Hypothesis 2 - Our framework will have a higher node level accuracy. We hypothesize that our framework will achieve a higher accuracy in the node level than the FALCON model as Hi-Viscont can correctly predict the object in a node without direct queries specifying the type of object. FALCON does not handle these indirect queries well as it does not update parent concepts with knowledge about novel leaf level concepts.

Hypothesis 3 - We hypothesize that our framework will achieve higher ratings on subjective metrics compared to the baseline because of its higher accuracy and competency in completing the requests.

Results

In this section, we present two sets of results. We first present experiment results on concept learning on the visual question answering task on three different domains. The experiment results demonstrate that our concept net model learns better representation for concepts than the baseline model, and is more robust for continual learning across all domains. Secondly, we present the results for the human-subjects experiments. The results for the human-subjects study demonstrate that our framework Hi-Viscont can learn visual task via an in-situ interaction with human user with more accuracy and usability than the baseline model.

VQA Experiment Setup

Domains. We first present experimental results on VQA tasks for three domains: the **CUB-200-2011 dataset**, a **custom house-construction domain** with building blocks, and a **custom zoo domain** with terrestrial and aquatic animals. We created the House-Construction and Zoo domains because they allow us to construct arbitrarily hard tasks with different types of objects that a robot can grasp. For each domain, we introduce additional general concepts on top of the existing concept classes to construct a concept hierarchy. The detailed descriptions and the statistics of the datasets can be found in the Appendix.

Data Creation Protocol. Following FALCON's data creation protocol, we procedurely generate training and testing examples for each domain. We generate descriptive sentences and questions based on the ground truth annotations of images and external knowledge, which is the relationship between concepts. For all the descriptive sentences and the questions, we also generate the corresponding neural-symbolic programs.

Experiment Configuration. We directly compare Hi-Viscont with FALCON on all the three domains. To demonstrate that Hi-Viscont is better for continual learning, we compare these models with no pre-trained concepts. We present the mean and standard deviation of the F1 metric across the three datasets as our major results. Each of these results is obtained from five trials with different splits of concepts and image. We evaluate the question-answer pairs for all concepts for all the three domains on images that are not shown in the pre-train or the train phase. Images used for testing are never seen by the model in any phase of training for both train concepts and test concepts. Additional statistics(precision, recall, and t-test) and a more detailed analysis can be found in the Appendix.

Mtd.	Species	Genera	Family	Order	Class
HV	87.1±2.0	**90.4±0.6**	**90.7±1.7**	**92.0±0.8**	95.9±8.2
FCN	86.5±1.4	88.2±1.0	84.3±1.4	84.3±3.2	99.3±1.0

Table 2: The average F1 score and standard deviation of Hi-Viscont (HV) and FALCON (FCN) on the CUB dataset by the depth of concepts in the hierarchy.

VQA Results

In Table 1, we present the results on the VQA task for test concepts. Our model, Hi-Viscont achieves comparable results to the baseline state-of-the-art FALCON model on test concepts in all three domains. Given that in a concept network there are fewer parent concepts than leaf concepts, the performance of both models is comparable in such a general test case. However, when we split the concepts by their depth in the hierarchy, Hi-Viscont shines and achieves a significantly better performance with the parental nodes, which will be discussed by each domain separately.

CUB dataset: We present our results for concepts by their level in the taxonomy in Table 2. Hi-Viscont is better with significance for concepts in the level of Genera($p < 0.001$), Family($p = 0.001$), and Order($p = 0.001$) according to paired t-tests. Species are the leaf level concepts where the models again perform comparably as expected. This is because the leaf level updates of Hi-Viscont and FALCON do not differ significantly. As there is only one highest level ancestor for the Class with CUB there is no negative example

Method	Object	Color	Affordance
Hi-Viscont	88.46±1.58	**99.24±0.70**	**89.86±9.12**
FALCON	89.28±0.93	87.27±5.83	57.35±9.23

Table 3: The average F1 score and standard deviation of Hi-Viscont and FALCON on the house construction domain by type of concepts.

for it in the dataset leading to similar performance by both models as the answer is always `True`.

House construction domain: In this domain, the Color and Affordance concepts are non-leaf nodes in the hierarchy, whereas the object concepts are the leaf nodes. Following expectations, as demonstrated in Table 3, Hi-Viscont has a comparable performance to FALCON in the leaf node object concepts, while achieving significant improvements in both Color ($p = 0.005$) and Affordance (non-leaf) concepts ($p = 0.002$) according to the pairwise t-tests.

Zoo Domain In the zoo domain, leaf concepts are not at equivalent depths from the root node forcing us to analyze the performance crudely with respect to leaf and non-leaf nodes in Table 4. Similarly, Hi-Viscont achieves a comparable performance at leaf level concepts, but becomes significantly better than FALCON in the non-leaf concepts ($p = 0.001$).

Human Subjects Study

We conducted a human-subjects study with 18 participants (22.22% female, mean age = 25.36, standard deviation = 3.49). To design our study we conducted pilot studies with 10 participants. Each participant completed three phases of interaction with the robot on the house construction domain and filled out all surveys. Our experiment results with respect to each hypothesis are as follows:

Hypothesis 1: - Hi-Viscont achieves a 33.33% improvement in success rate(SR) compared to FALCON. Results from the Wilcoxon signed-rank test indicate that Hi-Viscont's SR is significantly better than FALCON ($Z = 0.0$, $p = 0.014$).

Hypothesis 2: - Hi-Viscont achieves a 19.44% improvements in accuracy at node level compared to FALCON. Results from the Wilcoxon signed-rank test indicate that Hi-Viscont's node level accuracy is significantly better than FALCON($Z = 1.5, p = 0.005$).

Hypothesis 3: - Hi-Viscont achieves higher ratings on subjective metrics than FALCON. Users prefer Hi-Viscont in all the scales that we measured with significance: Trust($t = 2.325$, $p = 0.016$, $df = 17$), SUS($t = 2.428$, $p = 0.013$, $df = 17$), Perceived Intelligence($t = 2.591$, $p = 0.010$, $df = 17$), and Anthropomorphism ($t = 2.924$, $p = 0.005$, $df = 17$), suggested by paired t-test. Additionally, results from Wilcoxon signed-rank test suggest Hi-Viscont is significantly preferred over FALCON($Z = 0.0$, $p < 0.001$).

Method	Leaf	Non-leaf
Hi-Viscont	87.93±3.40	**85.84±5.79**
FALCON	88.99±3.75	66.15±5.34

Table 4: The average F1 score and standard deviation of Hi-Viscont and FALCON on the zoo domain by type of concepts.

Metrics	Hi-Viscont	FALCON
Success Rate(%)	50.00 ± 51.45	16.67 ± 38.25
Node Accuracy(%)	81.25 ± 21.11	61.81 ± 23.67
Comparative	5.44 ± 2.68	0.39 ± 0.78
Trust	58.56 ± 11.60	51.94 ± 13.91
SUS	46.72 ± 10.33	43.33 ± 11.26
Intelligence	33.00 ± 5.90	28.61 ± 7.37
Natural	13.89 ± 4.19	12.11 ± 4.10

Table 5: The results of the human-subjects experiment. Success Rate and Node Accuracy are measured in percentage. Hi-Viscont is better than FALCON on all metrics with significance as described in the Human Subjects Study Section.

Limitations

There are a few clear limitations of our approach. Firstly, although that we tested Hi-Vicont on a large VQA dataset, we conducted our robotics study of visual task learning only on the House domain, which contains a small number of objects. We would like to increase the task complexity and the number of objects available in the domain in the future. Secondly, the proposed method does not generalize across different domains automatically akin to a foundation model. Using this method on a completely new domain requires us to train the concept net model from scratch. Thirdly, the interaction between users and the robots is controlled without being completely open and dynamic. Even though a fixed template for their language is not required, the users have to follow specific turn-taking rules. Lastly, our study uses college-age human subject's and we would like a wider sample of the population using our system.

Conclusion

In conclusion, we present Hi-Viscont, a novel concept learning framework that actively updates the representations of known concepts which is essential in continual learning settings such as robotics. Hi-Viscont achieves comparable performance to SOTA FALCON model on VQA task across three domains in leaf level concepts, and is significantly better on non-leaf concepts. Moreover, Hi-Viscont enables robots to learn a visual task from in-situ interactions by representing visual tasks with a scene graph. This approach allows zero-shot generalization to an unseen task of the same type. Finally, we conducted a human-subjects experiment to demonstrate Hi-Viscont's ability to learn visual tasks and concepts from in-situ interactions from participants that have no domain knowledge in the real world.

References

Ahn, M.; Brohan, A.; Brown, N.; Chebotar, Y.; Cortes, O.; David, B.; Finn, C.; Fu, C.; Gopalakrishnan, K.; Hausman, K.; Herzog, A.; Ho, D.; Hsu, J.; Ibarz, J.; Ichter, B.; Irpan, A.; Jang, E.; Ruano, R. J.; Jeffrey, K.; Jesmonth, S.; Joshi, N. J.; Julian, R.; Kalashnikov, D.; Kuang, Y.; Lee, K.-H.; Levine, S.; Lu, Y.; Luu, L.; Parada, C.; Pastor, P.; Quiambao, J.; Rao, K.; Rettinghouse, J.; Reyes, D.; Sermanet, P.; Sievers, N.; Tan, C.; Toshev, A.; Vanhoucke, V.; Xia, F.; Xiao, T.; Xu, P.; Xu, S.; Yan, M.; and Zeng, A. 2022. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. arXiv:2204.01691.

Bartneck, C.; Kulić, D.; Croft, E. A.; and Zoghbi, S. 2009. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, 1: 71–81.

Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Chen, X.; Choromanski, K.; Ding, T.; Driess, D.; Dubey, A.; Finn, C.; Florence, P.; Fu, C.; Arenas, M. G.; Gopalakrishnan, K.; Han, K.; Hausman, K.; Herzog, A.; Hsu, J.; Ichter, B.; Irpan, A.; Joshi, N.; Julian, R.; Kalashnikov, D.; Kuang, Y.; Leal, I.; Lee, L.; Lee, T.-W. E.; Levine, S.; Lu, Y.; Michalewski, H.; Mordatch, I.; Pertsch, K.; Rao, K.; Reymann, K.; Ryoo, M.; Salazar, G.; Sanketi, P.; Sermanet, P.; Singh, J.; Singh, A.; Soricut, R.; Tran, H.; Vanhoucke, V.; Vuong, Q.; Wahid, A.; Welker, S.; Wohlhart, P.; Wu, J.; Xia, F.; Xiao, T.; Xu, P.; Xu, S.; Yu, T.; and Zitkovich, B. 2023a. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. arXiv:2307.15818.

Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Dabis, J.; Finn, C.; Gopalakrishnan, K.; Hausman, K.; Herzog, A.; Hsu, J.; Ibarz, J.; Ichter, B.; Irpan, A.; Jackson, T.; Jesmonth, S.; Joshi, N. J.; Julian, R.; Kalashnikov, D.; Kuang, Y.; Leal, I.; Lee, K.-H.; Levine, S.; Lu, Y.; Malla, U.; Manjunath, D.; Mordatch, I.; Nachum, O.; Parada, C.; Peralta, J.; Perez, E.; Pertsch, K.; Quiambao, J.; Rao, K.; Ryoo, M.; Salazar, G.; Sanketi, P.; Sayed, K.; Singh, J.; Sontakke, S.; Stone, A.; Tan, C.; Tran, H.; Vanhoucke, V.; Vega, S.; Vuong, Q.; Xia, F.; Xiao, T.; Xu, P.; Xu, S.; Yu, T.; and Zitkovich, B. 2023b. RT-1: Robotics Transformer for Real-World Control at Scale. arXiv:2212.06817.

Brooke, J. 1995. SUS: A quick and dirty usability scale. *Usability Eval. Ind.*, 189.

Chai, J. Y.; Gao, Q.; She, L.; Yang, S.; Saba-Sadiya, S.; and Xu, G. 2018. Language to Action: Towards Interactive Task Learning with Physical Agents. In *IJCAI*, 2–9.

Chang, X.; Ren, P.; Xu, P.; Li, Z.; Chen, X.; and Hauptmann, A. 2023. A Comprehensive Survey of Scene Graphs: Generation and Application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 1–26.

Daruna, A.; Liu, W.; Kira, Z.; and Chernova, S. 2019. RoboCSE: Robot Common Sense Embedding. arXiv:1903.00412.

Gopalan, N.; Arumugam, D.; Wong, L. L.; and Tellex, S. 2018. Sequence-to-Sequence Language Grounding of Non-Markovian Task Specifications. In *Proceedings of Robotics: Science and System*.

Gopalan, N.; Rosen, E.; Konidaris, G.; and Tellex, S. 2020. Simultaneously learning transferable symbols and language groundings from perceptual data for instruction following. In *Proceedings of Robotics: Science and System*.

Han, C.; Mao, J.; Gan, C.; Tenenbaum, J. B.; and Wu, J. 2020. Visual Concept-Metaconcept Learning. arXiv:2002.01464.

Hu, R.; Andreas, J.; Darrell, T.; and Saenko, K. 2018. Explainable Neural Computation via Stack Neural Module Networks. *CoRR*, abs/1807.08556.

Hudson, D. A.; and Manning, C. D. 2018. Compositional Attention Networks for Machine Reasoning. *CoRR*, abs/1803.03067.

Jian, J.-Y.; Bisantz, A. M.; and Drury, C. G. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, 4(1): 53–71.

Johnson, J.; Gupta, A.; and Fei-Fei, L. 2018. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1219–1228.

Johnson, J.; Hariharan, B.; van der Maaten, L.; Hoffman, J.; Fei-Fei, L.; Zitnick, C. L.; and Girshick, R. B. 2017. Inferring and Executing Programs for Visual Reasoning. *CoRR*, abs/1705.03633.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment Anything. arXiv:2304.02643.

Li, Q.; Huang, S.; Hong, Y.; and Zhu, S.-C. 2020. A Competence-aware Curriculum for Visual Concepts Learning via Question Answering. arXiv:2007.01499.

Liu, W.; Paxton, C.; Hermans, T.; and Fox, D. 2021. StructFormer: Learning Spatial Structure for Language-Guided Semantic Rearrangement of Novel Objects. arXiv:2110.10189.

Mao, J.; Gan, C.; Kohli, P.; Tenenbaum, J. B.; and Wu, J. 2019. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. In *International Conference on Learning Representations*.

Mascharka, D.; Tran, P.; Soklaski, R.; and Majumdar, A. 2018. Transparency by Design: Closing the Gap Between Performance and Interpretability in Visual Reasoning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE.

Matuszek, C.; FitzGerald, N.; Zettlemoyer, L.; Bo, L.; and Fox, D. 2012. A joint model of language and perception for grounded attribute learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 1435–1442.

Mei, L.; Mao, J.; Wang, Z.; Gan, C.; and Tenenbaum, J. B. 2022. FALCON: Fast Visual Concept Learning by Integrating Images, Linguistic descriptions, and Conceptual Relations. In *International Conference on Learning Representations*.

Shridhar, M.; Manuelli, L.; and Fox, D. 2021. CLIPort: What and Where Pathways for Robotic Manipulation. arXiv:2109.12098.

Snell, J.; Swersky, K.; and Zemel, R. S. 2017. Prototypical Networks for Few-shot Learning. arXiv:1703.05175.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H. S.; and Hospedales, T. M. 2018. Learning to Compare: Relation Network for Few-Shot Learning. arXiv:1711.06025.

Tellex, S.; Gopalan, N.; Kress-Gazit, H.; and Matuszek, C. 2020. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3: 25–55.

Teney, D.; Liu, L.; and van den Hengel, A. 2017. Graph-Structured Representations for Visual Question Answering. arXiv:1609.05600.

Tian, Y.; Wang, Y.; Krishnan, D.; Tenenbaum, J. B.; and Isola, P. 2020. Rethinking Few-Shot Image Classification: a Good Embedding Is All You Need? arXiv:2003.11539.

Vilnis, L.; Li, X.; Murty, S.; and McCallum, A. 2018. Probabilistic Embedding of Knowledge Graphs with Box Lattice Measures. arXiv:1805.06627.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Kavukcuoglu, K.; and Wierstra, D. 2017. Matching Networks for One Shot Learning. arXiv:1606.04080.

Wang, X.; Ye, Y.; and Gupta, A. 2018. Zero-shot Recognition via Semantic Embeddings and Knowledge Graphs. arXiv:1803.08035.

Yi, K.; Wu, J.; Gan, C.; Torralba, A.; Kohli, P.; and Tenenbaum, J. B. 2019. Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. arXiv:1810.02338.

Zeng, A.; Florence, P.; Tompson, J.; Welker, S.; Chien, J.; Attarian, M.; Armstrong, T.; Krasin, I.; Duong, D.; Sindhwani, V.; et al. 2021. Transporter networks: Rearranging the visual world for robotic manipulation. In *Conference on Robot Learning*, 726–747. PMLR.

Zhu, Y.; Tremblay, J.; Birchfield, S.; and Zhu, Y. 2021. Hierarchical Planning for Long-Horizon Manipulation with Geometric and Symbolic Scene Graphs. arXiv:2012.07277.