Guiding Multi-Step Rearrangement Tasks with Natural Language Instructions

Anonymous Author(s)
Affiliation
Address
email

Abstract: Enabling human operators to interact with robotic agents using natural language would allow non-experts to intuitively instruct these agents. Towards this goal, we propose a novel Transformer-based model which enables a user to guide a robot arm through a 3D multi-step manipulation task with natural language commands. Our system maps images and commands to masks over grasp or place locations, grounding the language directly in perceptual space. In a suite of block rearrangement tasks, we show that these masks can be combined with an existing manipulation framework without re-training to greatly improve learning efficiency. Our masking model performs well with just a few hundred real training images, which is several orders of magnitude fewer examples than the millions of examples typically required for training Transformer models. Our modular design allows us to leverage supervised and reinforcement learning, providing an easy interface for experimentation with different architectures. We first test our model on a block manipulation task with synthetic commands, where we see a 530% improvement in task completion rate over a UNet-based baseline. We demonstrate that our model not only learns to localize actions correctly, but also learns a mapping of symbols to perceptual input that supports compositional reasoning. We then port an existing 3D block dataset with crowdsourced language to a simulated environment, and extend our approach to this domain, where we demonstrate our model's ability to handle syntactic and lexical variation, observing a 25.3% absolute improvement in identifying the correct block. This dataset provides a valuable resource for 3D instruction following research, specifically for manipulation.

Keywords: Instruction following, object grasping and manipulation, multimodal fusion, computer vision for robotic applications

1 Introduction

Reinforcement Learning (RL) is a powerful tool for developing perception-driven robotic systems. Most RL algorithms train an agent to perform a single task based on an given reward function. However, what if we would like to adapt or extend the learned policy of a robot to novel or more specialized tasks by providing instructions? Could we exploit previously learned RL policies for a robot to solve novel tasks *now* specified using natural language? The ability to adapt learned policies through natural language to perform novel tasks would open the door to a modular integration of perception, language, and action, avoiding the complexities of joint training over all modalities simultaneously. For example, in our experiments, we constrain a general-purpose block manipulation policy using natural language, without needing any language data to train the policy itself.

To achieve this modular learning, we need the ability to inject guidance from language into the RL-driven action policy while at the same time addressing with the challenges of computationally understanding natural language. For example, natural language semantics is often underspecified, especially with respect to a physical environment. Given a command, such as, "Move the green block to the left of the right block", there is a large space of possible actions satisfying the truth conditions stipulated by the command; correctly choosing one involves pragmatic and common-sense reasoning. Furthermore, in order to map from language to action, we need to address the symbol

Submitted to the 5th Conference on Robot Learning (CoRL 2021). Do not distribute.

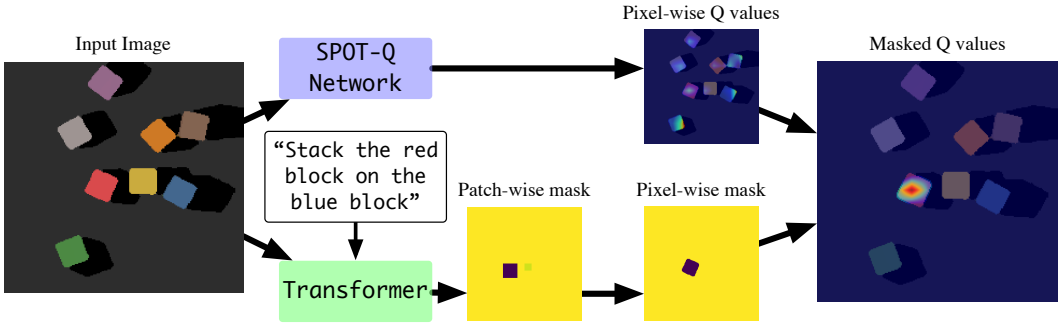


Figure 1: Schematic of our model. From an input image, pixel-wise Q-values are generated using a pre-trained network. Simultaneously, a Transformer-based model produces language-conditioned masks, which are heuristically turned into pixel-wise masks over the image. These masks are combined with the Q-values to find the highest Q-value that respects the language constraints.

grounding problem [1], i.e. how does one map symbols (e.g. "left", "right", "block", "red") to the physical world? Modeling this noisy mapping typically requires a sensor or data-fusion strategy [2, 3] to combine perceptual information about the world (e.g. images) with language. Previous work [4, 5, 6, 7, 8] has manually defined a fusion strategy in the architecture. In contrast we leverage the self-attention of Transformer encoders to *learn* a strategy from the data.

In our work, we factorize language-driven action into learning *how to act* and learning *where to act*, similar to Misra et al. [4]. Learning how to act is accomplished by the SPOT-Q algorithm [9], which teaches an agent to re-arrange colored blocks into stacks and rows by grasping and placing them. We use SPOT-Q models pre-trained without any language data for our manipulation experiments, demonstrating that our method lets us leverage existing manipulation models without additional training. Our model produces masks over the action space, conditioned on an image of the state as well as the language that segment the perceptual space into valid action regions for a set of possible actions. These masks, which encode valid action regions for grasping and placing, are intersected with a set of pixel-wise Q-values from the SPOT-Q RL framework to obtain the best action given a state and command. A schematic of our model is shown in Figure 1.

Factorizing the learning process this way provides several advantages. Firstly, it offers a more efficient path to learning. Separating the language grounding model from the reinforcement learning model, leads to a less complex RL model that receives more direct feedback for tasks. Moreover, the separation of action and language understanding allows us to learn the language component via supervised learning, which is typically more sample efficient. Secondly, the modular nature of the model allows us to swap components easily: we can train a masking model on simulated images *or* real images (cf. Section 5.1), allowing us to use the same action model for both simulated and real tasks without re-training. In an end-to-end model, a change to the input space would likely require re-training from scratch. Finally, the factorized model output is more interpretable; when an action is unsuccessful, we can inspect our model to localize the point of failure, be it a lack of comprehension or a mistake while acting. This may not be evident from the agents action alone.

In summary, our contributions are:

1. We introduce a novel Transformer-based model for mapping language commands to masks over an action space. Our approach's modular nature enables us to seamlessly integrate an existing Q-learning framework proposed by Hundt et al. [9] for block manipulation tasks.
2. We show our model succeeds in a variety of simulated and real data settings with orders of magnitude less data than typically available to Transformers, furthering the potential use of Transformers in the low-data regimes commonly found in robotics.
3. We demonstrate this model's utility on an existing dataset (introduced by Bisk et al. [10]) involving a challenging block re-arrangement task with *crowdsourced* natural language commands containing syntactic diversity as well as complex natural language phenomena.
4. We release our real and simulated datasets, plus a filtered version of the Bisk et al. [10] dataset containing the subset of examples feasible under a realistic model of physics.

2 Related Work

We describe related work in the areas of natural language instruction following on robots, Q-Learning for multistep manipulation tasks, and Transformer models for Computer Vision.

Robots and Natural Language Commands – We follow the Problem-Solution Sequences (PSS) paradigm [11], wherein a multi-step manipulation task (e.g. "build a stack") is framed as a sequence of incremental images or state representations. These may be paired with a instruction on how to accomplish the intermediate goal (e.g., "stack the red block on the blue block"). Several PSS datasets have been proposed [11, 10] and modeled in 2D [4, 12] and 3D [10, 3]. We translate the 3D dataset proposed by Bisk et al. [10] to a simulated environment in Section 5.2. Our experiments on this dataset and in simulation involve retrieving and placing blocks; Nguyen et al. [13] use natural language constraints to guide object retrieval, but do not model placement. Imitation learning [14] and semantic parsing [15] have also been used for block manipulation tasks; our work uses Q-learning for policies, and does not specify an ontology or semantic parse. TransporterNets [16] imitates multi-step tasks from demonstration, without natural language input. Concept2Robot [17] is among the most relevant work since it completes language-described single-step tasks via a combination of learning from demonstration and RL, however, it does not consider several key aspects of this work: we evaluate challenging multi-step tasks with the possibility of *progress reversal*[9]; we succeed on much smaller datasets; we consider crowdsourced commands; and we evaluate on real robot data.

A large body of work focuses on navigation tasks given natural language instructions [18, 4, 8, 7, 5]. Misra et al. [4] introduce the visual goal prediction paradigm for language-conditioned navigation. Visual goal prediction (VGP) frames the problem of choosing an action region as a semantic segmentation task, treating the classification of each pixel as a separate binary decision: act or do not act. This paradigm has been used to direct a simulated [5] and real quadcopter [8, 7]. We follow the same VGP paradigm, implementing a version of UNet[19] used by [4, 5, 7, 8] which we contrast with a novel Transformer-based architecture. While language has also been used for reward-shaping [20, 21, 22], we separate the RL from the natural language constraints, using the language-agnostic task-specific reward shaping implemented by SPOT-Q.

Q-learning for Multi-Step Tasks – Q-Learning for robotics is an active area of research into algorithms and supporting datasets [3], including various RL mechanisms for tabletop clutter removal [23, 24, 25, 26], but NLP is not addressed in these works. "Good Robot!" [9] was the first to complete multi-step tasks with consideration of *progress reversal*, a failure case where actions undo previous progress such that $\Omega(n)$ additional actions are required to recover, *e.g.* toppling the tower in a partially completed stacking task. Mohammed et al. [27] reviews RL grasping techniques.

Transformer Encoders for Vision – Though Transformer encoders, which build representations using self-attention, have become standard in NLP, computer vision tasks are still dominated by convolutional models, of which UNet [19] is a paradigmatic instance. The use of Transformers in vision tasks has been held back by the quadratic complexity of the Transformer's self-attention layer; an image input treated as a sequence of pixels would quickly become computationally intractable [28]. Dosovitskiy et al. [28] introduce a minimal modification to perform image recognition using Transformers, tiling the image into square patches, which they concatenate into a sequence and augment with learned linear positional embeddings. Though the model presented by Dosovitskiy et al. performs image-level recognition well, it cannot perform pixel-level tasks (e.g., semantic segmentation) as-is. Ranftl et al. [29] present an extension to this model for pixel-level prediction.

3 Models

We propose a Transformer-based model to convert image and natural language inputs into separate masks for grasping and placing, thereby allowing users to guide multi-step manipulation tasks.

We begin by tiling the $n \times n$ birds-eye image of the scene with C colored blocks into $k \times k$ patches, whose RGB pixel values are concatenated into one vector, resulting in patch vectors $I = i_1, \ldots, i_L$. The resulting vectors are augmented with a fixed sinusoidal positional encoding [30]. Similarly, each token in the command string $W = w_1, \ldots w_m$ is embedded into a continuous space resulting in vectors $e_1, \ldots, e_m$, and those vectors are augmented with positional encodings as well.

The goal is to train a model that produces a pair of language-defined spatial masks $\hat{M}^{\text{place}} \in \mathbb{R}^{n \times n}$ and $\hat{M}^{\text{grasp}} \in \mathbb{R}^{n \times n}$, and an optional pair of state reconstructions $\hat{S}^{\text{grasp}}, \hat{S}^{\text{place}} \in \mathbb{R}^{C \times n \times n}$ given by:

$$\hat{M}^t = R\Big(\sigma\big(H_m^T f_t\big(f_{\text{shared}}([I;W])_{0:k}\big)\Big),\ \hat{S}^t = R\Big(\sigma\big(H_s^T f_t\big(f_{\text{shared}}([I;W])_{0:k}\big)\Big) \tag{1}$$

where $f_.$ are parameterized by Transformer encoders with output size k, $H_m \in \mathbb{R}^{k \times 2}$, $H_s \in \mathbb{R}^{k \times C}$, σ is the softmax function, and $t \in \{\text{grasp, place}\}$. The reconstruction function R up-samples k

patches by copying the patch value across each of its constituent pixels, resulting in an $n \times n$ image. Note that structurally, the "shared," "grasp," and "place" Transformers are identical, but while the "shared" Transformer is used both for grasp and place prediction, the others are task-specific.

We train the masking model separately from the manipulation model. For the latter model, we use a pre-trained SPOT-Q model from Hundt et al. [9], which produces pixel-wise Q-values for picking and placing, making it easy to integrate with our approach. Furthermore, pixel-wise Q-values facilitate potential future work on extending our method to non-cubic shapes. All masking models are trained with a multi-task loss, composed of a pixel-wise binary cross-entropy loss for the masks, and an optional auxiliary reconstruction loss:

$$\mathcal{L}_t = -\frac{1}{N^2}\left(\sum_{i,j} M^t_{i,j} \log(\hat{M}^t_{i,j}) + \lambda \sum_{i,j} S^t_{i,j} \log(\hat{S}^t_{i,j})\right),\ \mathcal{L} = \mathcal{L}_{grasp} + \mathcal{L}_{place} \tag{2}$$

The reconstruction loss here forces the model to semantically segment the image by identifying pixel with its block ID, or labeling it as background. As in Nguyen and Salazar [31], we use a pre-norm Transformer layer with scaled initialization. Our learning rate schedule is based on Vaswani et al. [30], with a varied number of warmup steps as in Nguyen and Salazar. A comprehensive report of all hyperparameters and a figure of our architecture is given in Appendix A.

Q-Learning – Our model relies on a learned function from states and actions to expected rewards, $Q(s_t, a)$, which we learn via the SPOT-Q algorithm [9], which introduces novel reward shaping processes for improving efficiency and sim-to-real transfer. More details are given in Appendix A.3. Given a pixel-wise $n \times n$ state space of an image and two actions, grasp and place, the output of this function can be expressed as a tensor $Q \in \mathbb{R}^{n \times n \times 2}$.

4 Datasets and Resources

4.1 Simulation Experiments with Synthetic Commands

To train our masking model and demonstrate its performance in simulated environment on a robot arm, we create a dataset of commands and grasp/place masks. Using trajectories from Hundt et al. [9], we annotate successful grasp-and-place action pairs with synthetic commands generated from a template, and use simulation state information to construct the reference masks M^t and state information S^t. While synthetic language makes simplifying assumptions about the syntactic and lexical diversity of natural language, it provides a controllable test-bed for symbolic reasoning, and is often used to study compositionality [32, 33, 34], as we do in Section 5.1. For row-making, the template is "move the `color_a` block next to the `color_b` block", and for stacking the template is "stack the `color_a` block on the `color_b` block". We train supervised VGP models (cf. Section 3) on these commands, and combine the produced masks with the Q-values from the SPOT-Q framework in a simulated environment to execute the actions in order. The visual input to our models is the concatenation of a birds-eye color image and depth image. The training data is augmented by flipping each image across 4 axes, manipulating the command accordingly, as well as by the addition of pixel-wise Gaussian noise drawn from $\mathcal{N}(0, 0.05)$. This results in train/dev/test splits of $5520\,/46\,/46$ and $8970\,/74\,/76$ for row-making and stacking, respectively, where each sample is an input image and command paired with a reference place and grasp mask. Experiments and results on VGP metrics and in the simulation for this data are given in Section 5.1.

4.2 Transfer to Real Images

In order to be useful in real-life settings, our method must also be applicable to real data. However, unlike simulated data, data from a real robot is expensive and time-consuming to collect. Therefore, it is natural to ask: how much real data is needed to train an adequate VGP model?

The Hundt et al. [9] real robot trial logs include 1496 actions each for row-making and stacking. We split these into an 80/10/10 train/dev/test split and augment the examples by flipping, producing $5980, 747, 748$ datapoints. Image collection details are given in Appendix B.2. Given the small size of the real dataset, we pre-train a model on simulated images and fine-tune it with the real images.

4.3 Compositionality

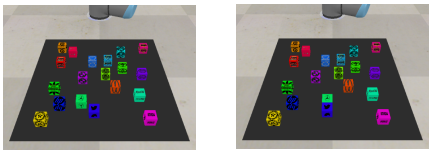
In linguistics, compositionality is typically taken to mean that the meaning of a whole utterance is *composed* of the meanings of its parts. For example, the meaning of "red block" is composed of the

meanings of "red" and "block" combined. Compositionality allows speakers to re-combine known symbols in novel ways, and underpins human language use. In our task, a compositional reasoner should correctly execute the instruction "stack the red block on the blue block" without ever having seen that particular combination of blocks; it should be able to localize red and blue blocks from other color contexts (e.g. red/green, blue/yellow, etc.). However, there is no explicit constraint during training to enforce this; the model could learn to interpret each of the 56 possible color combinations atomically. We test our model's compositional reasoning by examining the ability of our system to generalize to unseen color combinations. In this setting, we hold out one color pair during training, and then test exclusively on that pair. For example, we train a masking model with all color combinations except red and blue, and then test only on "stack the {red,blue} block on the {red,blue} block" examples. We iteratively hold-out each possible color pairing, training and evaluating a model as described above, average the metrics across all color pairs, and contrast this with a baseline of random train and test splits, where all color pairs are seen at train time. For efficiency, we do not augment the data with Gaussian noise and we train on each task separately.

4.4 Naturalistic Language Commands with Complex Scenes

Unlike language generated from a template, language sourced from real speakers displays syntactic and lexical diversity, mistakes, mis-spellings, and other idiosyncrasies that make modeling it far more challenging. While analyzing synthetic language showcases the ability of our model to integrate discrete symbolic commands with visual perception and action, and allows us to analyze our model's ability to generalize in a compositional fashion, it does not imply that our model generalizes to the complexities of real language.

To examine how our model handles these complexities, we use a 3D block manipulation dataset with natural language instructions introduced by Bisk et al. [10] (cf. Fig. 2). This dataset is set apart from other block-world data by its focus on complex instructions. Rather than using synthetic utterances, instructions were sourced from crowdworkers (9 per example), who were shown a pair of images and asked to provide instructions to a robot, with the goal of manipulating the blocks in the first image such that they result in the second image. This collection method result in more syntactically diverse utterances with complex linguistic concepts, such as "mirroring" and "balancing".



"Mercedes Benz will move right until it is above twitter"

Figure 2: Bisk et al. [10] blocks data sample.

The dataset is comprised of a sequence of state representations paired with natural language instructions on how to transition between states. While the blocks dataset is designed for robot manipulation, several assumptions were made to simplify annotation by crowdworkers. Crucially, physics in the environment were turned off while the structures were being constructed and annotated, resulting in some physically infeasible structures. To facilitate research in the blocks domain, and make the dataset more useful for training and evaluating embodied agents, we create a filtered version of the dataset, where physically infeasible examples are removed. Our filtering process removes $49.34\%, 43.93\%$ and 38.23% of train, dev, and test, respectively. Appendix C.1 contains more details on the translation process and data preprocessing. To model the dataset's world state sequences in the VGP framework, we translate the 3D Cartesian coordinates of each block given by the dataset's state representation into a top-down grid view, where each pixel corresponds to the ID of the block located at that position. We refer to this image as the scene's "state representation."

While the blocks dataset contains all the state information required to produce image-like state representations, this cannot generally be assumed to be the case. A more likely scenario mimicks Section 4.1, where the model's input is a combined color and depth image. After translating the dataset into the simulated environment used in Section 4.1, we collect color and depth images for the whole blocks dataset. In the original dataset, each block type is identified by a company logo. Since the logos were downsampled by the simulator, we additionally assign a color to each type. Experiments and results for this naturalistic dataset are given in Section 5.2

5 Experiments

5.1 Simulation Experiments with Synthetic Commands

An operator guiding a robot through a manipulation task may have a particular outcome in mind: for example, a stack with a particular order of blocks. Language can naturally express such constraints;

i.e. "stack the red block on the yellow block." Following the template generation and data annotation in Section 4.1, we train baseline and Transformer VGP models to produce output masks $\hat{M}^{\text{grasp}}$ and $\hat{M}^{\text{place}}$ and combine them with a SPOT-Q model to manipulate blocks in a simulated environment. These masks are filtered and thresholded using 0-1 block masks, following Hundt et al. [9]. To evaluate the VGP system in isolation, we report the following metrics:

- *Block Accuracy* is the percentage predictions for which the grasp mask overlaps with the correct block, i.e. the model chooses the correct block to pick up. Higher is better.
- *Idealized Error Score (IES)* captures the performance of an idealized robot on the manipulation task, choosing a block to move using the grasp mask, and a location using the place mask. After the chosen block is teleported to the predicted location, we measure the Euclidean distance between the block's new location and its reference location in block lengths. Lower is better.

We compare our model against baseline UNet-based VGP model; specifically, we extend the LINGUNET model presented by Misra et al. [4] with an attentional fusion mechanism. While the original LINGUNET used the final state of an LSTM for the language representation, we use a weighted average of BiLSTM states produced by an attention mechanism.

Further details on the masking, as well as the simulation environment, can be found in Appendix B.1. The filtered/thresholded masks $\hat{M}^{\text{grasp}}$ and $\hat{M}^{\text{place}}$ are stacked into a single mask $\hat{M} \in \mathbb{R}^{n \times n \times 2}$. A final matrix of Q' of Q values constrained by the language is given by $Q' = \hat{M} \odot Q$, where $\odot$ is the Hadamard product; Q' is then used to choose an action and a location at which to grasp/place in the simulation.

Table 1: Visual goal prediction (VGP) metrics for models trained on one or both tasks. Higher accuracy and lower Idealized Error Score (IES, measured in blocks), is better. TFMR is Transformer.

Model	Train	Test	Block Acc.	IES
UNet	Rows	Rows	57.1%	2.4
TFMR	Rows	Rows	84.8%	0.9
UNet	Both	Rows	66.5%	1.2
TFMR	Both	Rows	**91.3%**	**0.6**
UNet	Stacks	Stacks	27.9%	1.0
TFMR	Stacks	Stacks	81.5%	**0.4**
UNet	Both	Stacks	75.2%	0.6
TFMR	Both	Stacks	**90.4%**	**0.4**

We measure success by whether the model builds the row or stack while respecting the original order constraints. In addition, we measure the percentage of times the gripper picked up the correct block, and the number of actions taken to reach a successful row or stack. We contrast our model with the SPOT-Q framework alone, which has no access to the order constraints, and with our UNet baseline. Each model was evaluated with the same SPOT-Q model and tested on 100 trials. Hundt et al. [9] set a 30 action limit for failed trials; we added several evaluation heuristics to terminate failed trials when an irrecoverable state was reached, measuring color ordering errors. Stacking tasks use 8 block colors and row-making tasks have 4.

Table 1 shows the visual goal prediction metrics for all model variants trained on one or both tasks. We see that the Transformer-based models (TFMR) uniformly out-perform the UNet baseline across all metrics, often by large margins. For both UNet and Transformer architectures, training on two tasks simultaneously covers a broader range of semantically different examples with a corresponding increase in performance on all metrics. Notably, these examples differ by task: firstly, the row-making tasks contain a subset of the colors available in stacking. Secondly, while the grasp data across both tasks looks roughly identical, the data for place masks looks quite different. For example, both "stack the blue block on the red block" and "move the blue block next to the red block" result in grasp masks over the blue block, but whereas the place mask for stacking picks out the block on which to stack (the red block), the row mask specifies a location *near* the target block. Thus, it is intriguing that combining tasks benefits not only the block accuracy, but also the IES.

Table 2: Simulation Metrics. Higher task and color grasping success percentages are better. A lower average number of actions per trial is better.

Task	Model	Task %	Color %	Actions
Rows	RL only	10%	31.2%	15.3
Rows	UNet	10%	32.5%	19.2
Rows	TFMR	**53%**	**76.7%**	**12.3**
Stacks	RL only	0%	13.0%	NaN
Stacks	UNet	0%	15.2%	NaN
Stacks	TFMR	**29%**	**78.6%**	**11.6**

Table 2 shows the task completion percentage for the best UNet and Transformer models (trained on both tasks) in the simulated environment, as well as the percentage of successful grasps that picked up the correct color, and the average number of actions taken to complete a successful trial. In the aggregate, we see that high performance on VGP metrics does not translate to equivalent

performance in the simulation environment. For example, both for stacking and row-making, the percentage of correct color grasps is substantially lower that the VGP equivalent metric, block accuracy. For the Transformer model, this may be due to the imprecision of the patch-wise tiling of the image: when heuristically combined with a common-sense mask, a patch may cover multiple adjacent blocks. This is an inherent limitation of our approach. When compared to the RL-only and UNet baselines, our Transformer-based model succeeds in significantly more trials. Additionally, the mean number of actions across successful trials is lower, indicating that the Transformer-based masks result in fewer actions per successful trial. Note that for stacking, both baselines succeeded on 0 out of the 100 trials; due to toppling and occlusion, stacking poses a significantly harder task than row-making, and it is harder to recover from a misplaced block, even if the trial is not ended. We see that even a fairly high block accuracy and low IES are insufficient for high task performance. This highlights the one of the main challenges of multistep tasks: small error rates compound across multiple actions, yielding low task performance. A full error analysis is given in Appendix B.3

Transfer to Real Images – In Figure 3, we plot the IES of row-making and stacking models when trained on increasing subsets of real data. We see that for both rows and stacks, performance improves slightly as we increase the amount of training data, with diminishing returns after 30% of the available data (1794 examples) are used. Given that Transformers typically require several orders of magnitude more data to reach good performance [35, 31, 36], the fact that our architecture can be trained with mere thousands of simulated training examples, and successfully finetuned with even fewer real examples, extracted from only a few hundred real images, represents a significant achievement. For a proof-of-concept demonstration of our model running on a real robot, see Appendix B.4 and the supplementary materials.

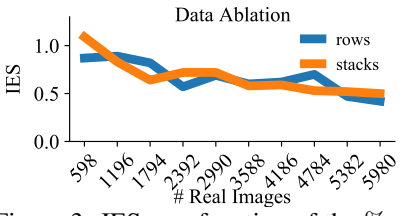


Figure 3: IES as a function of the % of real fine-tuning data used. IES is error measured in blocks, lower is better.

Compositionality – By generating held-out data splits as described in Section 4.3, we can test the ability of our model to generalize compositionally, i.e. to recombine known symbols in unseen contexts. A model that has learned compositional color representations will be able to combine color representations across contexts; for example, if it has seen red blocks in the context of green blocks, it should generalize to a blue block context. We would therefore expect the model to perform roughly on par with a model trained on a random data split, where all contexts are seen at train time. Conversely, if the model has learned to interpret each utterance as an atomic statement, we should expect much lower performance on the held-out subset.

Our results are reported in Table 3. Both rows and stacks see a slight drop in block accuracy from the baseline. Nevertheless, the lack of a large disparity between the baseline and the held-out condition suggests that the model has largely learned to generalize compositionally to unseen block combinations. This finding suggests that, although the model certainly has the capacity to memorize 56 combinations, and although it is not in any way trained to perform compositional reasoning, it naturally learns to bind symbols to percepts in a way that generalizes across contexts (i.e. "red block" refers to one particular block, regardless of its context). These result are promising but limited to simulated language; accordingly, in Section 5.2 we evaluate our results with real language data.

Table 3: VGP metrics on a held-out subset of unseen color pairs. Baseline models trained on a randomly-shuffled split.

Task	Data Split	Block Acc.	IES
Stacks	Random	85.7%	0.4
Rows	Random	89.1%	1.7
Stacks	Unseen	81.6%	0.6
Rows	Unseen	84.3%	1.3

5.2 Naturalistic Language Commands with Complex Scenes

In Section 5.1, we demonstrated the performance of our model on data in a simulated environment, with language commands generated via templates. Here, we model the naturalistic data described in Section 4.4, where state representations are coupled instead with diverse crowdsourced commands. We train models on state representations (cf. Section 4.4) as well as birds-eye color and depth images. This highlights the flexibility of our model: with no modifications to the architecture, we are able to train from images and from state representations.

Large pre-trained encoders have been shown to encode useful linguistic features at a number of levels of analysis [37, 38, 39]; accordingly, we make use of `bert-base-cased` [40] as an input

feature, contrasting it with GloVe [41] word embeddings. We conduct similar hyperparameter tuning as in Section 5.1, and report the same metrics.

Table 4 shows the VGP metrics for models trained from image-like representations of the state (described in Section 4.4) as well as from images. For brevity, only the best model for each embedding type is reported, with full results in Appendix C.4. On state representations, the Transformer-based models show a clear advantage over the baseline, with IES being substantially lower for both Transformer models. When using GloVe embeddings with state representation input, 2×2 patches resulted in the best performance, while a patch size of 4 yielded the best results for BERT embeddings. Overall, while the BERT-based model has slightly higher block accuracy, the IES are equal between the two models when trained on state representations.

Table 4: Visual Goal Prediction metrics for models trained on state information. Best Transformer models out-perform UNet, both on state and image inputs.

Model	Input	k	Block Acc.	IES
UNet + GloVe	State	1	56.5%	3.2
UNet + GloVe	Image	1	63.4%	3.1
TFMR + GloVe	State	2	88.4%	**2.1**
TFMR + GloVe	Image	2	*88.7%*	*2.8*
TFMR + BERT	State	4	**90.5%**	**2.1**
TFMR + BERT	Image	2	83.6	3.4

In Table 4, we also see that the performance shifts substantially when changing the input from state to images. While the baseline improves slightly in IES, the Transformers' performance decreases dramatically. These results stand in contrast to those in Section 5.1, where a Transformer trained on images unequivocally outperformed the baseline. Nevertheless, the best Transformer model still outperforms the baseline. We speculate that the baseline models may perform better in this condition because the architecture was developed for image segmentation problems. Interestingly, while a BERT-based Transformer performed well when trained from state, the same models' performance here is greatly diminished. Improving the Transformer's real-world performance is a direction for future work.

Data Translation and Filtering In Section 5.1, we observed that high performance in a static evaluation against a dataset did not entail equally high performance when using a realistic robot for multi-step tasks in a simulated environment. In its current form, the blocks dataset can only be evaluated using static metrics. Furthermore, the images provided in the dataset are taken from a fixed 45 degree projection, and our method assumes access to birds-eye images. To obtain birds-eye images of the environment and to facilitate further research into combining rich, naturalistic language with action in manipulation tasks, we translate the dataset into the same environment used in Section 5.1, and filter the dataset to only physically feasible examples; this process is described in Section 4.4.

Table 5 reports the VGP metrics on the the physically feasible subset of the data for the Transformer models. Firstly, we see that across most metrics, the performance decreases. However, for the UNet baseline, the IES improves on the physically feasible subset. In contrast, the Transformer models decrease in performance across both metrics. Nevertheless, the Transformer model still performs on par with the baseline on IES, and substantially better on block accuracy. These results suggest that the physically feasible subset is more challenging to model, even when evaluating the models statically. In other words, even when the evaluation and models are physically unconstrained, modeling physically feasible scenes is more difficult than modeling scenes in general. This result highlights the difficulty of developing embodied agents as well as the importance of testing them in realistic environments.

Table 5: VGP metrics for models trained on images of the physically feasible subset of the blocks dataset.

Model	Input	k	Block Acc.	IES
UNet + GloVe	Image	1	56.4	**3.0**
TFMR + GloVe	Image	2	**84.4**	**3.0**
TFMR + BERT	Image	2	78.9	3.6

6 Conclusion

We have presented a general approach to multi-step rearrangement and manipulation tasks that factorizes each step into learning how to act and learning where to act. To decide where to act, we have presented a Transformer-based model for converting state representations and natural language instructions to masks over an action space that learns with remarkably few examples. We have shown the ability of our model to act in a realistic simulated environment for two multi-step block rearrangement tasks, and naturally learn a compositional mapping from text to actions in a perceptual space. We have demonstrated our model performs well under both these realistic physical assumptions and with real and complex natural language via a dataset of crowdsourced instructions.

References

[1] S. Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3), 1990.

[2] C. Paxton, A. Hundt, F. Jonathan, K. Guerin, and G. D. Hager. CoSTAR: Instructing collaborative robots with behavior trees and vision. *ICRA*, 2017. URL https://arxiv.org/abs/1611.06145.

[3] A. Hundt, V. Jain, C.-H. Lin, C. Paxton, and G. D. Hager. The costar block stacking dataset: Learning with workspace constraints. *IROS, 2019*, 2019.

[4] D. Misra, J. Langford, and Y. Artzi. Mapping instructions and visual observations to actions with reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1004–1015, 2017.

[5] V. Blukis, N. Brukhim, A. Bennett, R. A. Knepper, and Y. Artzi. Following high-level navigation instructions on a simulated quadcopter with imitation learning. *arXiv preprint arXiv:1806.00047*, 2018.

[6] V. Blukis, Y. Terme, E. Niklasson, R. A. Knepper, and Y. Artzi. Learning to map natural language instructions to physical quadcopter control using simulated flight. In *Conference on Robot Learning (CoRL)*, 2019.

[7] V. Blukis, R. A. Knepper, and Y. Artzi. Few-shot object grounding and mapping for natural language robot instruction following. *arXiv preprint arXiv:2011.07384*, 2020.

[8] V. Blukis, Y. Terme, E. Niklasson, R. A. Knepper, and Y. Artzi. Learning to map natural language instructions to physical quadcopter control using simulated flight. In *CoRL*, 2020.

[9] A. Hundt, B. Killeen, N. Greene, H. Wu, H. Kwon, C. Paxton, and G. D. Hager. "Good Robot!": Efficient reinforcement learning for multi-step visual tasks with sim to real transfer. *IEEE RA-L*, 5(4), 2020.

[10] Y. Bisk, K. J. Shih, Y. Choi, and D. Marcu. Learning interpretable spatial operations in a rich 3d blocks world. In *AAAI*, 2018.

[11] Y. Bisk, D. Yuret, and D. Marcu. Natural language communication with robots. In *NAACL HLT 2016*, 2016.

[12] S. Gubbi Venkatesh, A. Biswas, R. Upadrashta, V. Srinivasan, P. Talukdar, and B. Amrutur. Spatial reasoning from natural language instructions for robot manipulation. *arXiv:2012.13693*, 2020.

[13] T. Nguyen, N. Gopalan, R. Patel, M. Corsaro, E. Pavlick, and S. Tellex. Robot object retrieval with contextual natural language queries. *arXiv preprint arXiv:2006.13253*, 2020.

[14] S. Stepputtis, J. Campbell, M. Phielipp, S. Lee, C. Baral, and H. B. Amor. Language-conditioned imitation learning for robot manipulation tasks. *Neurips 2020*, 2020.

[15] I. Perera, J. Allen, C. M. Teng, and L. Galescu. Building and learning structures in a situated blocks world through deep language understanding. In *SPLU 2018*, 2018.

[16] A. Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, V. Sindhwani, and J. Lee. Transporter networks: Rearranging the visual world for robotic manipulation. *CoRL*, 2020.

[17] L. Shao, T. Migimatsu, Q. Zhang, K. Yang, and J. Bohg. Concept2robot: Learning manipulation concepts from instructions and human demonstrations. In *Proceedings of RSS*, 2020.

[18] N. Gopalan, E. Rosen, G. Konidaris, and S. Tellex. Simultaneously learning transferable symbols and language groundings from perceptual data for instruction following. 2020.

[19] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*. Springer, 2015.

[20] P. Goyal, S. Niekum, and R. Mooney. Using natural language for reward shaping in reinforcement learning. In *IJCAI 2019*, 2019.

[21] P. Goyal, S. Niekum, and R. J. Mooney. Pixl2r: Guiding reinforcement learning using natural language by mapping pixels to rewards. *arXiv preprint arXiv:2007.15543*, 2020.

[22] T.-J. Fu and W. Y. Wang. Semi-supervised policy initialization for playing games with language hints. In *NAACL 2021*. ACL, June 2021.

[23] A. Zeng, S. Song, S. Welker, J. Lee, A. Rodriguez, and T. Funkhouser. Learning synergies between pushing and grasping with self-supervised deep reinforcement learning. In *IROS*. IEEE, 2018.

[24] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, et al. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *CoRL*, 2018.

[25] I. Akinola, J. Varley, and D. Kalashnikov. Learning precise 3d manipulation from multiple uncalibrated cameras. In *ICRA*, 2020.

[26] S. Song, A. Zeng, J. Lee, and T. Funkhouser. Grasping in the wild: Learning 6dof closed-loop grasping from low-cost demonstrations. In *IEEE RA-L*, volume 5, 2020.

[27] M. Q. Mohammed, K. L. Chung, and C. S. Chyi. Review of deep reinforcement learning-based object grasping: Techniques, open challenges, and recommendations. *IEEE Access*, 8, 2020.

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[29] R. Ranftl, A. Bochkovskiy, and V. Koltun. Vision transformers for dense prediction. *arXiv preprint arXiv:2103.13413*, 2021.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Neurips*, 2017.

[31] T. Q. Nguyen and J. Salazar. Transformers without tears: Improving the normalization of self-attention. *arXiv preprint arXiv:1910.05895*, 2019.

[32] B. Lake and M. Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *ICML*. PMLR, 2018.

[33] B. M. Lake, T. Linzen, and M. Baroni. Human few-shot learning of compositional instructions. *arXiv preprint arXiv:1901.04587*, 2019.

[34] L. Ruis, J. Andreas, M. Baroni, D. Bouchacourt, and B. M. Lake. A benchmark for systematic generalization in grounded language understanding. *arXiv preprint arXiv:2003.05161*, 2020.

[35] M. Popel and O. Bojar. Training tips for the transformer model. *arXiv preprint arXiv:1804.00247*, 2018.

[36] A. Araabi and C. Monz. Optimizing transformer for low-resource neural machine translation. *arXiv preprint arXiv:2011.02266*, 2020.

[37] J. Hewitt and P. Liang. Designing and interpreting probes with control tasks. In *EMNLP-IJCNLP*, 2019.

[38] I. Tenney, D. Das, and E. Pavlick. Bert rediscovers the classical nlp pipeline. In *ACL*, 2019.

[39] G. Jawahar, B. Sagot, and D. Seddah. What does bert learn about the structure of language? In *ACL*, 2019.

[40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT*. ACL, 2019.

[41] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[42] E. Rohmer, S. P. N. Singh, and M. Freese. Coppeliasim (formerly v-rep): a versatile and scalable robot simulation framework. In *IROS 2013*, 2013. www.coppeliarobotics.com.