

Do Trajectories Encode Verb Meaning?

Dylan Ebert

Brown University

dylan_ebert@brown.edu

Chen Sun

Brown University

chensun@brown.edu

Ellie Pavlick

Brown University

ellie_pavlick@brown.edu

Abstract

Distributional models learn representations of words from text, but are criticized for their lack of grounding, or the linking of text to the non-linguistic world. Grounded language models have had success in learning to connect concrete categories like nouns and adjectives to the world via images and videos, but can struggle to isolate the meaning of the verbs themselves from the context in which they typically occur. In this paper, we investigate the extent to which *trajectories* (i.e. the position and rotation of objects over time) naturally encode verb semantics. We build a procedurally generated agent-object-interaction dataset, obtain human annotations for the verbs that occur in this data, and compare several methods for representation learning given the trajectories. We find that trajectories correlate as-is with some verbs (e.g., *fall*), and that additional abstraction via self-supervised pretraining can further capture nuanced differences in verb meaning (e.g., *roll* vs. *slide*).

1 Introduction

While large distributional language models such as BERT (Devlin et al., 2019) and GPT (Radford, 2020; Brown et al., 2020) have had empirical success in deriving representations of words and sentences from large text corpora, most of these models lack *grounding*, or a connection between the words and their real-world referents. Grounding, in addition to being necessary for multimodal tasks like video recognition, has been argued to lie at the core of language *understanding* (Bender and Koller, 2020). Work on grounded language learning associates language with the non-linguistic world, typically by learning from large-scale image (Bruni et al., 2011) or video (Sun et al., 2019) datasets.

Much prior work on language grounding has focused on concrete nouns (objects) and adjectives (attributes), which are captured well by patterns of pixels. Verbs, however, have received less attention,

despite being essential for building models that can interact in realistic 3D environments (Shridhar et al., 2020a; Bisk et al., 2020). Verbs are especially challenging to model, given that they take place over time. Image and video data alone is insufficient to fully capture verb semantics, as demonstrated by prior work (Yatskar et al., 2016), in many cases failing to isolate the meaning of the verb from context in which it typically occurs. For example, Chao et al. 2018 show that an image of a person laying in the snow next to a snowboard is labeled “standing on a snowboard”. Moreover, recent work has introduced datasets and benchmarks based on situated 3D environments (Gan et al., 2020; Deitke et al., 2020; Ebert and Pavlick, 2020; Shridhar et al., 2020a) that demonstrate the challenges of learning task-oriented behavior, which demands a combination of object and verb grounding.

In this paper, we test the hypothesis that the semantics of (concrete) verbs are grounded in the 3D trajectories of objects: i.e., the absolute and relative paths objects take through 3D space. We investigate if and when verb meanings appear to be a product of raw perception of objects in 3D space, and when differentiating verb meanings requires additional abstraction and representation beyond what is available via direct perception. To study this, we collect a clean dataset of 3D object trajectories in simulation. We collect human descriptions of these perceived world dynamics, i.e., to determine whether or not a given event constitutes a *fall* or a *tumble*. We then propose a self-supervised pretraining approach, whereby we train a time-series prediction model to obtain representations of trajectories in a 3D environment without any linguistic input. We evaluate the learned representations on how well they encode verb semantics for specific verbs. We show that the pretrained model learns to represent events in a way that aligns well with the meaning of English verbs, e.g. differentiating *slide* from *roll*. In summary, our primary contributions

are:

1. We introduce a new, clean dataset of 3D object trajectories paired with human judgments about whether or not each trajectory falls within the extension of each of 24 different verbs. To the best of our knowledge, this is the first dataset of its kind, and provides a valuable resource for empirical studies of lexical semantics. Our data is available at <https://github.com/dylanebert/simulated>.
2. We compare several representation learning methods in terms of their ability to capture verb semantics *without any linguistic signal during training*. In particular, we investigate the roll of abstraction (via self-supervised pre-training) compared to raw perception in capture verb meanings. To our knowledge, this is the first work to apply neural networks and (pre-linguistic) concept learning to the study of verb semantics.

2 Related Work

Grounded Language with Deep Learning.

Our contributions add to a large body of work on grounded representation learning. Much of this work augments language modeling objectives with images (Silberer and Lapata, 2012; Lazaridou et al., 2015; Kiela et al., 2017) and videos (Sun et al., 2019). In this work, we focus on representations that encode verb semantics. Prior work on verb learning has been conducted in the computer vision community, typically described as “human-object interactions” (Regneri et al., 2013; Chao et al., 2018; Sun et al., 2018; Ji et al., 2019). Most closely related to our approach, which focuses on trajectory data, is work on learning affordances for human-robot communication. For example, Kalkan et al. (2014); Ugur et al. (2009) learn affordance representations based on the state changes of objects, but do not encode the full trajectory between states. Also related is work in grounded language in text-only models which investigates models ability to reason about objects through space and time (Aroca-Ouellette et al., 2021).

Outside of NLP, models have been trained on trajectory data for applications like human motion path forecasting (Giuliani et al., 2021) and human activity recognition (Wang et al., 2018). Our work lies at the intersection of grounded language learning and spatiotemporal machine learning, using

representations of trajectory data to study verb semantics.

Grounding and Lexical Semantics. Prior work in formal semantics attempts to build feature-based representations of verb meaning in terms of the 3D trajectories and state transitions entailed by those verbs (Pustejovsky and Krishnaswamy, 2014; Siskind, 2001; Steedman, 2002). Such work is related more generally to the idea of mental simulation as a means for representing and reasoning about linguistic concepts (Feldman, 2008; Bergen et al., 2007; Bergen, 2012). We view our contribution as consistent with and complementary to this formal semantics program. While the prior work has sought to codify the precise truth conditions of motion verbs, we investigate whether such representations could emerge organically from data-driven processes.

While we focus on concrete verbs in this paper, other work has argued that motor processing and mental simulation plays a more general role in language processing. For example, Gärdenfors (2019) makes a case for grounded distributional “conceptual spaces” as the foundation for modeling linguistic concepts. Dorr and Olsen (2018) discusses the role of metaphor in modeling abstract uses of words like *push*. Borghi and Riggio (2009) argues for the notion of a “motor prototype” as a key component of recognizing and processing objects, and Mazzuca et al. (2021) presents evidence that the sensorimotor system (in particular the interactive aspects) drive acquisition of abstract concepts.

3 Dataset

3.1 Overview

To carry out the proposed study, we require a dataset that contains continuous 3D recordings of an agent interacting with an object. While our representation learning methods will not use linguistic supervision, we require verb labels in order to evaluate our models. Thus, in our data, we require that each recording is annotated with verbs describing the motion of the object. For example, if the agent throws a bouncy ball across a room, we’d expect the recording to be annotated with a verb sequence such as *be thrown, fall, bounce, bounce, bounce, roll, stop*.

To produce such data, we build a simple Markovian agent which interacts with a variety of objects in a 3D virtual environment. We record the result-

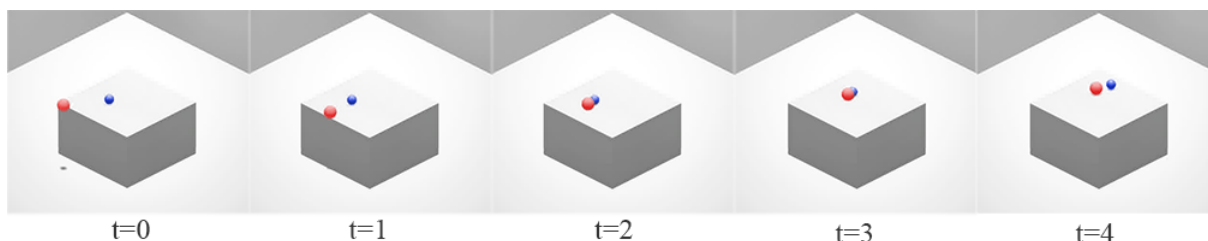


Figure 1: The *Simulated Spatial Dataset* consists of procedurally generated motion data of a virtual agent interacting with an object. In this sequence the agent (red sphere) pushes the object (blue sphere). At $t=0$ and $t=1$, the agent approaches the ball. Then, in $t=2$ and $t=3$, the agent pushes to ball. Finally, at $t=4$, the ball is rolling away from the agent.

ing trajectory of the object and then, using crowd-sourcing, ask humans to determine which verbs could accurately describe which portions of the object’s movement. An example sequence from the dataset is shown in Figure 1.

3.2 Data Generation and Terminology

In this section we provide details on how we generate the data, and introduce terminology that will be used throughout the rest of the paper.

Environment. The dataset is generated in Unity, a game engine seeing increased use by researchers (Deitke et al., 2020; Gan et al., 2020) for its accessible rendering and physics simulation via the underlying Nvidia PhysX physics engine. The dataset and simulation source code are publicly available.¹

Trajectory data. We define trajectory data as the position and rotation of entities in space, represented with three-dimensional XYZ coordinates and four-dimensional $XYZW$ quaternions respectively. We choose to focus on only these features, ignoring other possibilities like object shape or identity, in order to focus on learning generalizable aspects of verb semantics that are independent of the object.

Sessions. The dataset is generated in 3-minute continuous segments we refer to as **sessions**. Within each session, several parameters are randomized, including object shape, mass, drag, friction, and bounciness.

Action Primitives. The data generation is driven by a Markov Chain with a set of randomly parameterized **action primitives**. In this Markov Chain, the States are whether the object is *Held*, *OnCounter* and *OnGround*. The transitions between

¹<https://github.com/dylanebert/simulated>

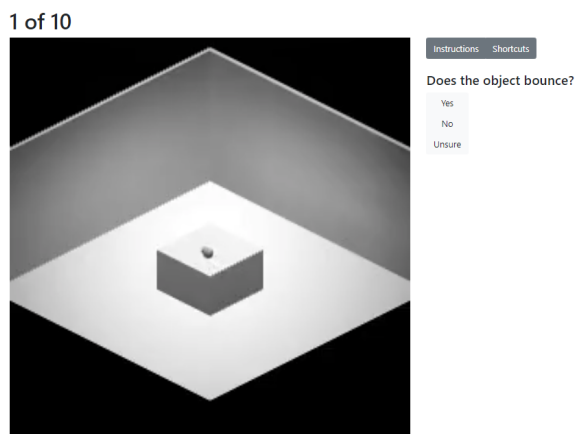


Figure 2: Crowd annotation task. Crowdworkers make binary judgments on whether the verb applies to the clip. In this example, the worker is asked *Does the object bounce?* about the 1.5s video clip on the left.

these states are action primitives like *PickUp*, *Put-Down*, or *Throw*. For example, when the object is in the state *OnCounter*, the agent may execute a *PickUp*, after which the object is *Held*. These action primitives, combined with the physics of the objects (e.g., their shape, mass, friction, bounciness, etc) are intended to produce a wide range of object motions corresponding to a range of verbs, and we do not expect that the primitives will map directly to the verbs that one would use to describe the resulting object behavior. For example, when we simulate a *Throw* primitive, the result might be that the object flies across the room, hits the wall, falls to the floor, and bounces until it comes to a rest. We parameterize the execution of each action with action-specific parameters, e.g. the force of a throw. The combination of session- and action-level parameters can result in a wide variety of object motion from each primitive action. A full description of the parameters for each action can be found in Appendix A.

Verbs. We highlight a distinction between action primitives and the *high-level actions* or **verbs** that emerge from them. For example, if the object is *pushed*, it may then *slide*, *bounce*, *roll*, *tumble*, or any combination thereof. We refer to all of these as *verbs*, though only *push* is an action primitive. We highlight this distinction because we are most interested in studying the nuanced verbs that emerge from the simulation, rather than the action primitives that drive it explicitly.

Frames. Our atomic unit is **frames**, also referred to as **timesteps**, which represent a single point in time. Our dataset is collected at 60 fps, or 10,800 frames per session. For each frame, we record the position and rotation of the object, as well as the position of the agent. This is sufficient to reconstruct and render the scene from an arbitrary perspective as needed. We choose this high framerate because it’s relatively fast and inexpensive to rapidly produce trajectory data, which can be subsampled as needed for rendering or modeling.

3.3 Crowdsourced Annotation

We collect labels for which *verbs* occur in the data, and when they occur. To do this, we extract short clips from the dataset, and ask crowdworkers to provide binary judgments on whether the clip falls in the extension of the verb.

Clips. We extract short clips from the dataset using Hierarchical Dynamic Clustering with Motion energy-based pooling (Zhang et al., 2018), a self-supervised action segmentation framework that can be summarized as follows:

1. The 3D space is divided into clusters using the provided trajectory data. The framework uses Hierarchical Dynamic Clustering, which is similar to k-means but shown to outperform it on human motion parsing tasks.
2. A sliding window is applied to the cluster labels for a given positional sequence. The number of transitions between clusters in a window are defined as its *motion energy*.
3. The subsequent motion energy curve is smoothed using a Gaussian kernel with a tuned smoothing factor.
4. The peaks of the motion energy curve are considered motion *segments*, with lengths varying with respect to the width of the peak.

This algorithm is shown to perform well on human motion parsing, which we find transfers well

to our dataset when applied to *object position*. This yields easily identifiable patterns of motion, e.g. from the time the object is thrown to when it slows to a stop. We find that, in contrast to a random sliding window, this approach avoids cutting clips in the middle of salient patterns of motion.

In our case, a disadvantage of this approach is that the extracted segments are variable-length. To simplify our pipeline, we filter to only segments of length 72 to 96, then crop the segment to length 90, or 1.5 seconds. We call each 1.5s segment a **clip**. We choose this length to make the clip as short as possible to avoid crowdworker fatigue, but give sufficient time for a human observer to recognize what’s happening.

Verbs. We produce 24 queries, each corresponding to a verb, e.g. *Does the object bounce?* To do this, the authors curate a list of 24 verbs² of interest which are likely to occur in the simulated data and range from general descriptions (e.g., *fall*) to more subtle descriptions of object motion (e.g., *tumble*). When asking annotators whether a verb applies to a clip, we always frame the question with the object as the subject. That is, when a carry event occurs, annotators are asked “is the object carried”.

We then consider every possible (clip, query) pair a potential crowdsourcing *task*. We apply conservative heuristics to filter out (clip, query) pairs that are guaranteed to have a negative label. For example, if the *Held* state was never present in a clip, we don’t ask if the object is *carried*. This results in approximately 110k tasks, from which we sample 100 tasks per query, for a total 2400 crowdsourcing tasks, such as the one shown in Figure 2.

Labels. For each crowdsourcing task, we obtain responses from five workers, then take the majority response as the *label* for that clip. The same clip is shown for all applicable queries, resulting in a supervised dataset of 24-dimensional vectors, representing binary verb labels for each clip.³ The dataset and all unaggregated annotations are available for download.⁴

²fall, carry, fall off, fall over, bounce, drop, pick up, push, topple, bump, tumble, roll, put down, hit, throw, flip, toss, tip, stop, spin, slap, slide, start, turn

³Labels are only yes or no. *Unsure* was not the majority label for any task. Tasks that were filtered out during crowdsourcing are assigned a mask value that is ignored during training and validation.

⁴<https://lunar.cs.brown.edu/simulated/>

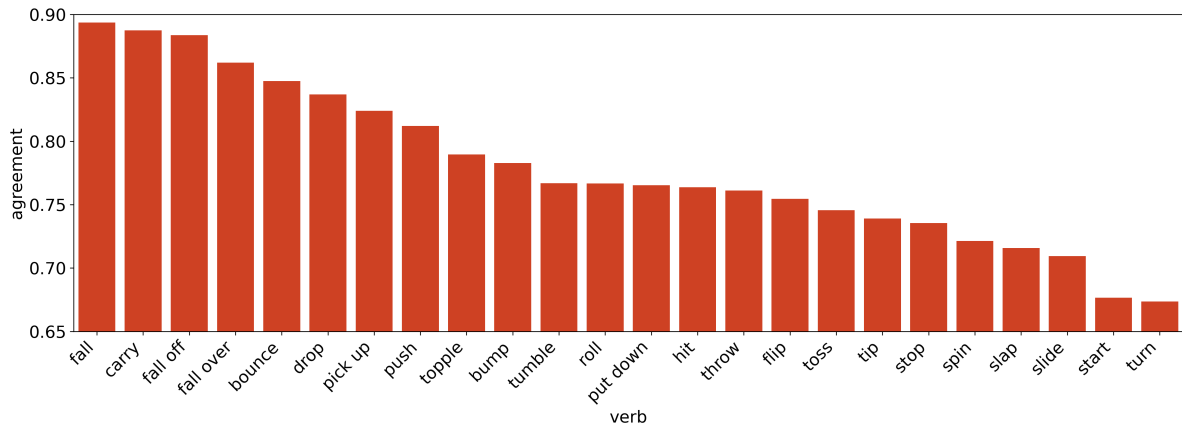


Figure 3: Crowd annotation agreement by verb. Workers agree most on when verbs of gravity occur, such as *fall*, *drop*, *bounce*, and least on when verbs of rotation occur, i.e. *turn*, *spin*, *tip*.

4 Dataset Analysis

In this section, we analyze trends in the dataset annotations, including worker agreement, and comparisons between semantically related verbs.

4.1 Agreement

Annotation agreement on a clip is the proportion of responses that match the majority label for that clip. Figure 3 shows annotation agreement by verb. A noticeable trend is that agreement is higher for particular semantic categories. Specifically, verbs that involve gravity, i.e. *fall*, *fall off*, *drop*, and *bounce* have higher agreement. On the other hand, verbs of rotation, i.e. *turn*, *spin*, *tip*, *flip* have lower agreement, alongside abstract verbs *start* and *stop*. For *start* in particular, we even received feedback from crowdworkers that they weren't sure whether the object *started* moving during the clip or not.

4.2 Co-occurrence

Figure 4 shows co-occurrence: specifically, given that a clip is labeled by at least one worker as verb v_1 , how often is it labeled by other workers as verb v_2 ? Co-occurrence allows us to answer questions like *how often is a toss considered a throw?* and vice-versa. We highlight some interesting verb relationships.

General co-occurrence. Verb co-occurrence is high in general. The average number of verbs used to describe a given clip is 4 (where a verb is considered “used” if at least three workers use it). This highlights the challenge of verb learning, as opposed to more concrete nouns and adjectives. Verbs

are applicable to a wide variety of behavior, even if it isn't a prototypical instance of that verb.

Lexical entailments. All dogs are animals but not all animals are dogs. These types of semantic containment relationships are also ascribed to verbs. Analyzing our collected data, in some cases, we observe the opposite of what's expected. For example, according to WordNet (Fellbaum, 2010), *toss* is a type of *throw*. However, using the majority labels, we find *throws* to be annotated as *tosses* more often *tosses* than are annotated as *throws*. That is, $p(\text{toss}|\text{throw}) = .67 < p(\text{throw}|\text{toss}) = .75$.

Frequent co-occurrences. *Hit*, *push*, and *bump* stand out as the most frequently co-occurring verbs, having over 90% co-occurrence with each other. These likewise occur when many other verbs do, but not reciprocally. For example, most *slaps* are *hits*, but only 41% of *hits* are *slaps*. In many cases, this can be explained by other verbs being immediately preceded by the agent making contact with the object, which gets labeled *hit*, *push*, and *bump*.

Fine-grained distinctions. Workers distinguish **roll** from **slide** - only 50% of *rolls* are also considered *slides*, and vice-versa. This validates that verbs with similar trajectories, which may be challenging for models, are indeed differentiated by humans. Additionally, verbs with similar but nuanced meanings are differentiated. For example, *tip*, *tumble*, *fall over*, and *topple* tend to co-occur around 70-80% of the time. These also fall into “verbs of rotation” category, which have the lowest annotator agreement. It isn't clear the extent to which these are nuanced distinctions, or annotation

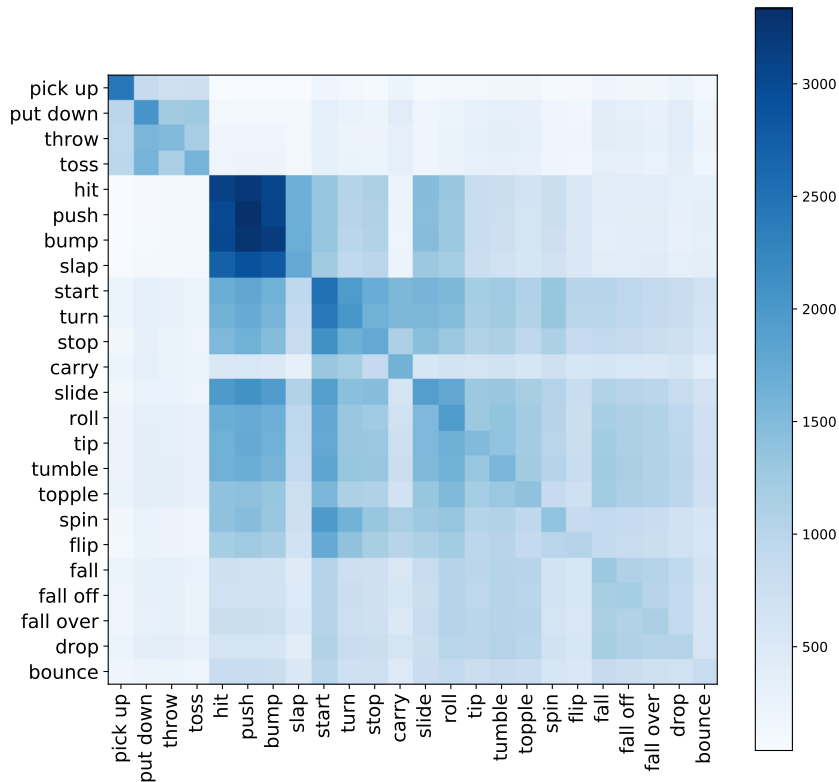


Figure 4: Co-occurrence of different verbs for the same clip. Specifically, given that at least one worker labeled a clip as v_1 (row), how many times did other workers label the clip as v_2 (column).

noise.

5 Experiments

Our hypothesis is that representation learning in the 3D visuospatial world (without language supervision) can yield concept representations that align to English verb semantics—i.e. can the representations capture nuanced distinctions like *throw* vs. *toss* or *slide* vs. *roll*? To test this, we pretrain a self-supervised model on a time-series prediction task, and then use a perceptron classifier to evaluate its learned representations.

We evaluate four approaches, described in detail below. First, we train a simple perceptron to evaluate the representational capacity of the trajectory data as-is, as a comparative baseline. Second, we train a fully supervised model to determine a soft upper bound on the task without pretraining. Third, we evaluate our self-supervised model. And finally, we fine-tune the self-supervised model to determine an upper bound with pretraining.

5.1 Experimental Setup

For all approaches, we evaluate representation quality with a multi-way verb classification task. Specif-

ically, we predict the verb labels for the 1.5s clips gathered through the crowdsourcing task described in Section 3.3.

Each input sample $X_{t_{1..90}}$ is a 90x10 matrix of position and rotation data, corresponding to 90 frames per clip and 10 spatial features⁵ per frame. The output Y is a 24-dimensional multi-hot vector indicating the whether each of our 24 verb classes apply to the clip.

5.2 Approaches

Perceptron. We wish to evaluate the representational capacity of the raw trajectory data itself. To do so, we train a single 24-dimensional dense layer with sigmoid activation, equivalent to a perceptron for each class. While very simple, this approach gives an idea of how well trajectory data represents verbs as-is, and provides a naive comparative baseline against which to evaluate our more complex pretraining techniques.

Fully Supervised. The fully supervised approach is similar to the perceptron, but adds a dense layer and LSTM layer in-between. This is equiv-

⁵Object Position XYZ, Hand Position XYZ, and Object Rotation XYZW.

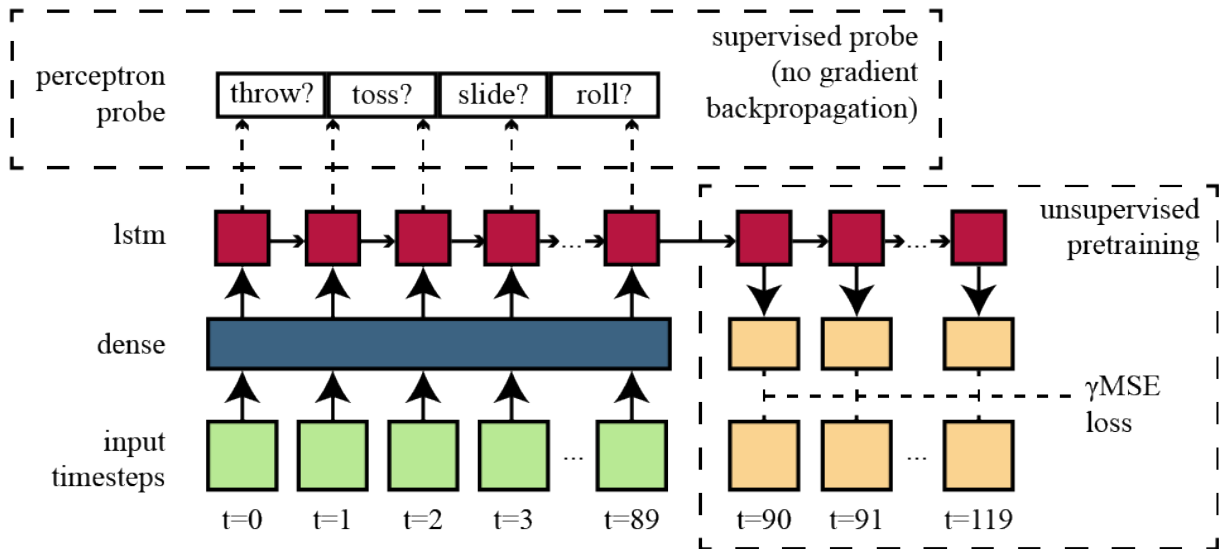


Figure 5: Our pretraining setup. During pretraining, the model learns to encode and represent input timesteps for time-series prediction. To evaluate these learned representations, a perceptron probe is trained on the LSTM outputs, without propagating gradients to the pretrained model.

alent to the model shown in Figure 5, but trained end-to-end without pretraining. The purpose of this approach is to provide an upper bound to the experimental setup *without* pretraining.

Self-supervised Pretraining. To evaluate the capacity of self-supervised models to represent trajectory data, we pretrain a time-series prediction model on a large unlabeled dataset of 400k sessions. That is, given n input frames $X_{t_{1..n}}$, the model is trained to predict k output frames $Y_{t_{n+1..n+k}}$. The model consists of a dense layer followed by an LSTM layer unrolled k timesteps, as shown in Figure 5. We use a discounted mean squared error loss as shown in Equation 1, which discounts loss by how far it is into the future by factor γ .

$$\gamma\text{MSE} = \sum_{t=n}^{n+k} \gamma^t (y_t - \hat{y}_t)^2 \quad (1)$$

We tune discount factor γ , output length k , model width, and batch size using a grid search on validation performance, resulting in values of 0.85, 60, 128, and 1024, respectively. Input length n is fixed at 90 to match the length of clips.

We consider the concatenated LSTM outputs as the *representation* of a clip. To evaluate this representation compared to raw trajectory data, we freeze the weights of the pretrained model and, as when evaluating the raw trajectory data, train a perceptron for each class.

Approach	mAP (%)
Random Stratified	41.4
Perceptron	65.3
Fully Supervised	72.2
Pretraining + Probe	76.3
Pretraining + Finetuning	77.4

Table 1: Mean Average Precision (mAP) for each approach. The pretrained approaches outperform others on verb classification.

Fine-tuning. To provide an upper bound for our experimental setup *with* pretraining, we fine-tune the self-supervised model. This is the same as the previous approach, but allows the gradients in the perceptron step to pass through the entire model.

6 Results

We report Mean Average Precision on unseen test data for each approach in Table 1. We compare these to random stratified predictions that are based on the class distribution of the training data.

Perceptron. The perceptron approach evaluates the representational capacity of raw trajectory data as-is, with a lower bound of random stratified and soft upper bound of fully supervised. The perceptron performs relatively well for its simplicity, being only 7 points below the fully supervised upper bound. This suggests that the trajectory data itself encodes a significant amount of verb meaning, but leaves plenty of room for improvement.

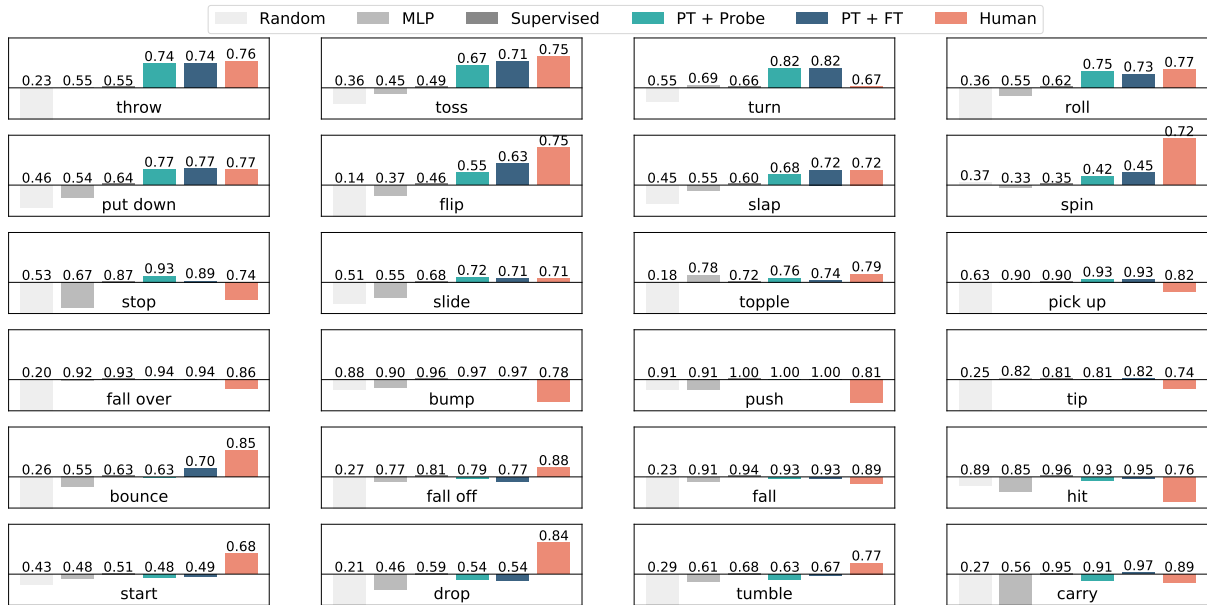


Figure 6: Comparison of each approach by verb. As we are primarily interested in the effect of pretraining, all bars are shown relative to the performance of the fully supervised model. I.e., the fully supervised model (which represents a soft upper bound on how well one can do given raw state information without any pretraining) is visualized as 0, and all other models are visualized relative to that. Human performance is inner-annotator agreement (% of workers who agree on the majority label).

Self-supervised pretraining. The pretraining + probe approach evaluates the ability of self-supervised models to encode verb meaning from trajectory data. This is equivalent to the perceptron approach, but with learned hidden representations as input rather than raw trajectory data. The pre-trained model does outperform the perceptron, as well as the fully supervised approach. Fine-tuning only improves on this slightly, highlighting that self-supervised pretraining can yield representations that successfully encode verb meaning.

Breakdown by verb. Figure 6 shows a comparison of average precision for each verb. There are some patterns worth highlighting. In particular, we can categorize verbs into three main groups: *trivial*, *tractable*, and *hard*.

Trivial verbs are verbs that can be well-represented by trajectory data as-is, i.e. those with high performance with the perceptron approach. These include *fall*, *fall off*, *fall over* and *pick up*.⁶ Many of these have high agreement, and may be explained by the object’s change in height.

Tractable verbs are those that see significant benefit from pretraining, including *slide*, *roll*, *throw*,

toss, *put down*, *turn*, *flip*, and *stop*. An intuition behind this is that these verbs involve manner distinctions, and in particular, rotations of the object relative to itself. Such information doesn’t fall directly out of raw state descriptions, but is likely to be well modeled by a pretraining objective that tries to predict the object’s future position.

Hard verbs are those with low performance that don’t benefit much from pretraining. These include *bounce*, *drop*, *tip*, *topple*, and *spin*. Many of these are verbs which have lower agreement. *Bounce*, *slap* and *spin* appear to benefit a bit from both pretraining and fine-tuning, suggesting that they may be tractable with similar but more robust pretraining. *Tip* and *topple* have fairly high performance, and may almost be categorized as trivial, perhaps being explained by the object’s change in rotation. However, they are noticeably lower than other trivial verbs, despite seeing no benefit from pretraining, suggesting that there is nuance to their meaning in the dataset, which isn’t captured by any approach. Finally, *drop* is a great example of a hard verb, as it is similar to trivial verbs like *fall*. However, *drop* involves interaction between the agent and object that is highly agreed upon by annotators, but doesn’t appear to be captured by our approaches, despite the model receiving both object and agent data. More challenging examples may be able to

⁶We exclude *hit*, *push*, and *bump* from trivial verbs, as these have high average precision with random stratified, showing that they are very positively skewed, not necessarily well-represented.

unveil a similar story for other verbs of interaction like *pick up* and *put down*.

7 Discussion and Conclusion

We test the hypothesis that verb meanings can be grounded in 3D trajectories, i.e., the position of objects over time. Specifically, we investigate the extent to which representations of object trajectories, learned without any linguistic supervision, naturally encode concepts that align to English verb semantics. Our primary contributions are twofold. First, we build a procedurally generated agent-object-interaction dataset for which we collect crowdsourced annotations. This is the first dataset of its kind, and provides a rich inventory of human judgments about the extensions of 24 verbs of motion. Second, we compare a variety of representation learning approaches, specifically contrasting approaches which operate directly on perceptual inputs to approaches which learn abstractions over the raw perception (via pretraining). We find that some verbs meanings (e.g., *fall* and *push*) are captured easily by the raw state information, while others (e.g., *roll* and *turn*) require additional processing to be represented well.

This work is a first step toward exploring ways to capture fine-grained distinctions in grounded verb semantics that are trivial for humans, but challenging for models. Recent benchmarks at the intersection of NLP, vision and robotics (Deitke et al., 2020; Shridhar et al., 2020b) illuminate unsolved challenges in AI that demand a more robust understanding of verb semantics and spatial reasoning. As these benchmarks continue to be developed, and rich multimodal datasets from technologies like virtual reality become increasingly abundant, we envision that future work in this vein will be especially relevant.

In the future, we plan to explore more sophisticated models for self-supervised pretraining, and evaluate how well these models transfer to more naturalistic language learning settings (Ebert and Pavlick, 2020). Beyond this, there is a large body of related research questions to be explored. For example, can representations of trajectory data be fused with visually-grounded representations to yield better encodings of verb semantics? Collaborative efforts will be key to addressing these next milestones in natural language understanding.

Acknowledgments

This work was supported by the DARPA GAILA program. Thank you to George Konidakis, Carsten Eickhoff, Roman Feiman, Jack Merullo, Charles Lovering, Gabor Brody, and the members of the Brown LUNAR lab for helpful discussion and feedback.

References

- Stéphane Aroca-Ouellette, Cory Paik, Alessandro Roncone, and Katharina Kann. 2021. Prost: Physical reasoning of objects through space and time. *arXiv preprint arXiv:2106.03634*.
- Emily M. Bender and Alexander Koller. 2020. *Climbing towards NLU: On meaning, form, and understanding in the age of data*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.
- Benjamin K Bergen. 2012. *Louder than words: The new science of how the mind makes meaning*. Basic Books.
- Benjamin K Bergen, Shane Lindsay, Teenie Matlock, and Srini Narayanan. 2007. Spatial and linguistic aspects of visual imagery in sentence comprehension. *Cognitive science*, 31(5):733–764.
- Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. *Experience grounds language*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.
- Anna M Borghi and Lucia Riggio. 2009. Sentence comprehension and simulation of object temporary, canonical and stable affordances. *Brain Research*, 1253:117–128.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. *Language models are few-shot learners*. *CoRR*, abs/2005.14165.
- Elia Bruni, Giang Binh Tran, and Marco Baroni. 2011. *Distributional semantics from text and images*. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*,

- pages 22–32, Edinburgh, UK. Association for Computational Linguistics.
- Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. 2018. [Learning to detect human-object interactions](#).
- Matt Deitke, Winson Han, Alvaro Herrasti, Anirudha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, et al. 2020. Robothor: An open simulation-to-real embodied ai platform. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3164–3174.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Bonnie Dorr and Mari Broman Olsen. 2018. Lexical conceptual structure of literal and metaphorical spatial language: A case study of “push”. In *Proceedings of the First International Workshop on Spatial Language Understanding*, pages 31–40.
- Dylan Ebert and Ellie Pavlick. 2020. [A visuospatial dataset for naturalistic verb learning](#). In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 143–153, Barcelona, Spain (Online). Association for Computational Linguistics.
- Jerome Feldman. 2008. *From molecule to metaphor: A neural theory of language*. MIT press.
- Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.
- Chuang Gan, Jeremy Schwartz, Seth Alter, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, Megumi Sano, et al. 2020. Threedworld: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954*.
- Peter Gärdenfors. 2019. Using event representations to generate robot semantics. *ACM Transactions on Human-Robot Interaction (THRI)*, 8(4):1–21.
- Francesco Giuliani, Irtiza Hasan, Marco Cristani, and Fabio Galasso. 2021. Transformer networks for trajectory forecasting. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10335–10342. IEEE.
- Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. 2019. [Action genome: Actions as composition of spatio-temporal scene graphs](#).
- Sinan Kalkan, Nilgün Dag, Onur Yürüten, Anna M Borghi, and Erol Şahin. 2014. Verb concepts from affordances. *Interaction Studies*, 15(1):1–37.
- Douwe Kiela, Alexis Conneau, Allan Jabri, and Maximilian Nickel. 2017. Learning visually grounded sentence representations. *arXiv preprint arXiv:1707.06320*.
- Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. [Combining language and vision with a multimodal skip-gram model](#). pages 153–163.
- Claudia Mazzuca, Chiara Fini, Arthur Henri Michal-land, Ilenia Falcinelli, Federico Da Rold, Luca Tum-olini, and Anna M Borghi. 2021. From affordances to abstract words: The flexibility of sensorimotor grounding. *Brain Sciences*, 11(10):1304.
- James Pustejovsky and Nikhil Krishnaswamy. 2014. [Generating simulations of motion events from verbal descriptions](#). In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 99–109, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.
- Benjamin Radford. 2020. [Seeing the forest and the trees: Detection and cross-document coreference resolution of militarized interstate disputes](#). In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 35–41, Marseille, France. European Language Resources Association (ELRA).
- Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36.
- Kumar Shridhar, Harshil Jain, Akshat Agarwal, and Denis Kleyko. 2020a. [End to end binarized neural networks for text classification](#). In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 29–34, Online. Association for Computational Linguistics.
- Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020b. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749.
- Carina Silberer and Mirella Lapata. 2012. [Grounded models of semantic representation](#). In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1423–1433, Jeju Island, Korea. Association for Computational Linguistics.

- Jeffrey Mark Siskind. 2001. Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Journal of artificial intelligence research*, 15:31–90.
- Mark Steedman. 2002. Plans, affordances, and combinatory grammar. *Linguistics and Philosophy*, 25(5):723–753.
- Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. *arXiv preprint arXiv:1904.01766*.
- Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. 2018. [Actor-centric relation network](#).
- Emre Ugur, Erol Sahin, and Erhan Oztop. 2009. Predicting future object states using learned affordances. In *2009 24th International Symposium on Computer and Information Sciences*, pages 415–419. IEEE.
- Lei Wang, Yangyang Xu, Jun Cheng, Haiying Xia, Jianqin Yin, and Jiaji Wu. 2018. Human action recognition by learning spatio-temporal features with deep neural networks. *IEEE access*, 6:17913–17922.
- Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yan Zhang, Siyu Tang, He Sun, and Heiko Neumann. 2018. Human motion parsing by hierarchical dynamic clustering. In *BMVC*, page 269.

A Dataset parameters

The following tables describe the session-level and action-level parameters for our procedural data generation protocol described in Section 3.2.

Parameter	Description	Random Values
Start Location	Initial agent location	<i>random from waypoint set</i>
Start Rotation	Initial agent rotation in degrees	(0, 360)
Target Mesh	Shape of the object	<i>cube/sphere/capsule/cylinder</i>
Target Position	Initial object location	<i>random location on counter</i>
Target Rotation	Initial object rotation in degrees	(0, 360)
Target Mass	Mass in kg of the object	(0.1, 10)
Target Drag	Hinders object motion	(0, 2)
Target Angular Drag	Hinders object angular motion	(0.1, 1)
Dynamic Friction	Friction when object is moving	(0, 1)
Static Friction ⁷	Friction when object is not moving	(0, 1)
Bounciness	Energy retained on bounce	(0, 1)

Table 2: Session-level parameters, which add variety between 3-minute sessions.

Parameter	Description	Values
Pick Speed	Hand velocity for Pick motion	(1, 3)
Put Speed	Hand velocity for Put motion	(1, 3)
Push Speed	Hand velocity for Push motion	(1, 3)
Throw Force	Object force for Throw motion	(25, 125)
Hit Force	Object force for Hit motion	(25, 125)

Table 3: Action-level parameters, which add variety in the execution of each action primitive in the dataset.