Discovering Novel Actions in an Open World with Object-Grounded Visual Commonsense Reasoning

Sathyanarayanan N. Aakur[*], Sanjoy Kundu, Shubham Trehan
Department of Computer Science
Oklahoma State University
Stillwater, OK 74078
{saakurn, sanjoy.kundu, shubham.trehan}@okstate.edu

Abstract

Learning to infer labels in an open world, i.e., in an environment where the target "labels" are unknown, is an important characteristic for achieving autonomy. Foundation models pre-trained on enormous amounts of data have shown remarkable generalization skills through prompting, particularly in zero-shot inference. However, their performance is restricted to the correctness of the target label's search space. In an open world where these labels are unknown, the search space can be exceptionally large. It can require reasoning over several combinations of elementary concepts to arrive at an inference, which severely restricts the performance of such models. To tackle this challenging problem, we propose a neuro-symbolic framework called ALGO - novel Action Learning with Grounded Object recognition that can use symbolic knowledge stored in large-scale knowledge bases to infer activities (verb-noun combinations) in egocentric videos with limited supervision using two steps. First, we propose a novel neuro-symbolic prompting approach that uses *object-centric* vision-language foundation models as a noisy oracle to ground objects in the video through evidence-based reasoning. Second, driven by prior commonsense knowledge, we discover plausible activities through an energy-based symbolic pattern theory framework and learn to ground knowledge-based action (verb) concepts in the video. Extensive experiments on two publicly available datasets (GTEA Gaze and GTEA Gaze Plus) demonstrate its performance on open-world activity inference and its generalization to unseen actions in an unknown search space. We show that ALGO can be extended to zero-shot settings and demonstrate its competitive performance to multimodal foundation models.

1 Introduction

Humans display a remarkable ability to recognize unseen concepts (actions, objects, etc.) by associating known concepts gained through prior experience and reasoning over their attributes. Key to this ability is the notion of "grounded" reasoning, where abstract concepts can be mapped to the perceived sensory signals to provide evidence to confirm or reject hypotheses. In this work, we aim to create a computational framework that tackles open-world egocentric activity understanding. We define an activity as a complex structure whose semantics are expressed by a combination of actions (verbs) and objects (nouns). To recognize an activity, one must be cognizant of the object label, action label, and the possibility of any combination since not all actions are plausible for an object. Supervised learning approaches [55, 49, 37, 17] have been the dominant approach to activity understanding but are trained in a "closed" world, where there is an implicit assumption about the target labels. The videos during inference will always belong to the label space seen during training.

[*] Corresponding Author.

Preprint. Under review.

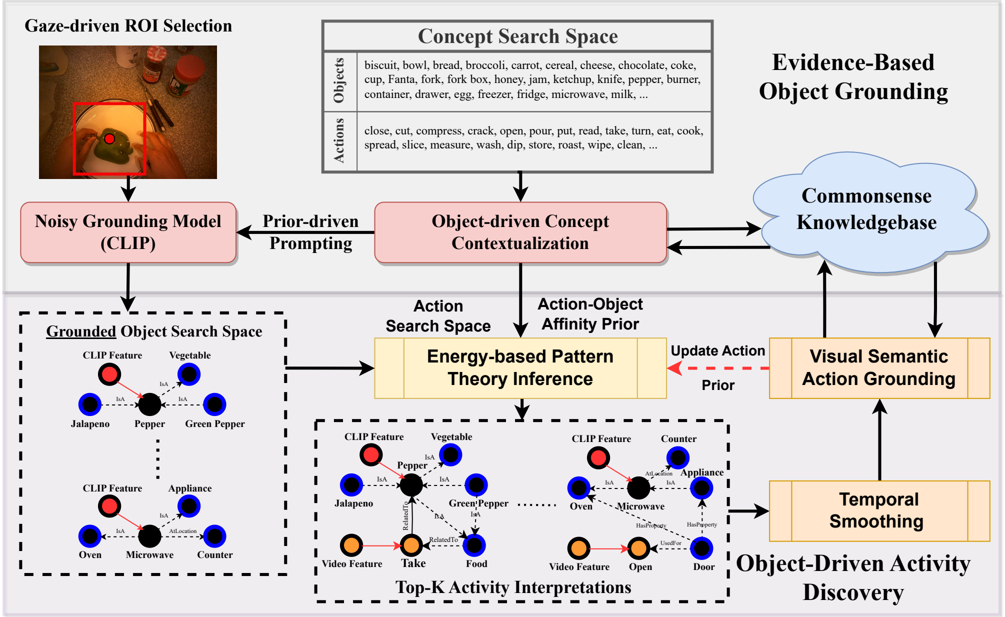

Gaze-driven ROI Selection
Concept Search Space
Objects
biscuit, bowl, bread, broccoli, carrot, cereal, cheese, chocolate, coke, cup, Fanta, fork, fork box, honey, jam, ketchup, knife, pepper, burner, container, drawer, egg, freezer, fridge, microwave, milk, ...
Actions
close, cut, compress, crack, open, pour, put, read, take, turn, eat, cook, spread, slice, measure, wash, dip, store, roast, wipe, clean, ...
Evidence-Based Object Grounding
Noisy Grounding Model (CLIP)
Prior-driven Prompting
Object-driven Concept Contextualization
Commonsense Knowledgebase
Grounded Object Search Space
CLIP Feature
Vegetable
IsA
Jalapeno
Pepper
Green Pepper
Appliance
AtLocation
Oven
Microwave
Counter
Action Search Space
Action-Object Affinity Prior
Energy-based Pattern Theory Inference
Update Action Prior
Visual Semantic Action Grounding
Temporal Smoothing
RelatedTo
Video Feature
Take
Food
HasProperty
UsedFor
Open
Door
Top-K Activity Interpretations
Object-Driven Activity Discovery

Figure 1: **Overall architecture** of the proposed approach (ALGO) is illustrated here. Using a two-step process, we first *ground* the objects within a gaze-driven ROI using CLIP [47] as a noisy oracle before reasoning over the plausible activities performed in the video. The inferred activity and action (verb) are grounded in prior knowledge and visual features to refine the activity interpretations.

Zero-shot learning approaches [61, 33, 62, 6] relax this assumption by considering disjoint "seen" and "unseen" label spaces where all labels are not necessarily represented in the training data. This setup is a *known* world, where the target labels are pre-defined and aware during training. In this work, we define an *open* world to be one where the target labels are unknown during both training and inference. The goal is to recognize elementary concepts and infer the activity.

Foundation models [8], pre-trained on large amounts of data, have shown tremendous performance on different problems such as question answering [15], zero-shot object recognition [47], and action recognition [33]. However, their ability to perform open-world inference is constrained by two factors. First, the search space (i.e., target label candidates) must be well-defined since their output is constrained to what is presented to them (or "prompted"), which requires prior knowledge about the environment. Second, their performance is dependent on the span of their pre-training data. Models trained on third-person views may not generalize to egocentric videos due to the limited capability to *ground* semantics in visual data and *reason* over object affordances. Learning these associations during pre-training is challenging since it requires data encompassing every possible combination of concepts. Yet, it restricts the model's functionality to a domain with a specific set of rules.

In this work, we propose to tackle this problem using a neuro-symbolic framework that leverages advances in multimodal foundation models to ground concepts from symbolic knowledge bases, such as ConceptNet [52], in visual data. The overall approach is shown in Figure 1. Using the energy-based pattern theory formalism [5, 2, 22] to represent symbolic knowledge, we ground objects (nouns) using CLIP [47] as a noisy oracle. Driven by prior knowledge, novel activities (verb+noun) are inferred, and the associated action (verb) is grounded in the video to learn visual-semantic associations for novel, unseen actions. The contributions of this work are three-fold: (i) We present a neuro-symbolic framework to leverage compositional properties of objects to prompt CLIP for evidence-based grounding. (ii) We propose object-driven activity discovery as a mechanism to reason over prior knowledge and provide action-object affinities to constrain the search space. (iii) We demonstrate that the inferred activities can be used to ground unseen actions (verbs) from symbolic knowledge in egocentric videos, which can generalize to unseen and unknown action spaces.

2 Related Works

Egocentric video analysis has been extensively explored in computer vision literature, having applications in virtual reality [25] and human-machine interaction. Varioys tasks have been proposed, such as question-answering [19], summarization [35], gaze prediction [31, 20, 3], and action recognition [30], among others. Success has been driven by the development of large-scale datasets such as Ego-4D [21], Charade-Ego [49], GTEA Gaze [20], GTEA Gaze Plus [31], and EPIC-Kitchens [13]. In the context of egocentric activity recognition, which is the focus of this work, supervised learning has been the predominant approach. Researchers have explored various techniques, such as modeling spatial-temporal dynamics [53], using appearance and motion cues for recognition [36], hand-object interaction [63, 56], and time series modeling of motion information [48], to name a few. Some studies have addressed the data-intensive nature by exploring zero-shot learning [61, 49]. KGL [5] is one of the first works to address the problem of **open-world understanding**. They represent knowledge elements derived from ConceptNet [52], using pattern theory [2, 14, 22]. However, their method relies on an object detector to ground objects in a source domain before mapping concepts to the target space using ConceptNet-based semantic correspondences. This approach has limitations: (i) false alarms may occur when the initial object detector fails to detect the object-of-interest, leading to the use of the *closest* object to the gaze, and (ii) reliance on ConceptNet for correspondences from the source domain to the target domain, resulting in objects being disregarded if corresponding probabilities are zero. Other efforts in **open-world learning** have primarily focused on *object-centric* tasks, such as open-world object detection [24, 18, 16], which do not address the combinatorial problems inherent in open-world *activity* recognition.

Vision-language modeling has gained significant attention in the community, driven by the success of transformer models [54] in natural language processing, such as BERT [15], RoBERTa [34], OpenAI's GPT series [45, 46, 9, 10], and ELECTRA [12]. The development of object-centric foundation models has enabled impressive capabilities in zero-shot object recognition in images, as demonstrated by CLIP [47], DeCLIP [32], and ALIGN [26]. These approaches rely on large amounts of image-text pairs, often in the order of *billions*, to learn visual-semantic representations trained various forms of contrastive learning [29, 11]. Recent works, such as EGO-VLP [33], Hier-VL [6], LAVILLA [62], and CoCa [59] have expanded the scope of multimodal foundation models to include egocentric videos and have achieved impressive performance in zero-shot generalization. However, these approaches require substantial amounts of curated pre-training data to learn semantic associations among concepts for egocentric activity recognition. **Neuro-symbolic models** [43, 27, 57, 5] show promise in reducing the increasing dependency on data. Our approach extends the idea of neuro-symbolic reasoning to address egocentric, open-world activity recognition.

3 Proposed Framework: ALGO

Problem Formulation. We address the task of recognizing unknown activities in egocentric videos within an open world setting. Our objective is to develop a system that can learn to identify elementary concepts, establish semantic associations, and systematically explore, evaluate, and reject combinations to arrive at an interpretation that best describes the observed activity class. In this context, we define the target classes as activities, which are composed of elementary concepts such as actions (verbs) and objects (nouns). These activities are formed by combining concepts from two distinct sets: an object search space (G_{obj}) and an action search space (G_{act}). These sets define the pool of available elementary concepts (objects and actions) that can be used to form an activity (referred to as the "target label"). The main challenge lies in effectively navigating through clutter and discovering unknown activities by leveraging visual cues from the observed video V_i and semantic cues based on knowledge.

To this end, we propose ALGO (Action Learning with Grounded Object recognition), illustrated in Figure 1, to tackle the problem of discovering novel actions in an open world. Given a search space (both known and unknown) of elementary concepts, we first explore the presence of plausible objects through evidence-based object grounding (Section 3.1) by exploring prior knowledge from a symbolic knowledge base. A noisy grounding model provides visual grounding to generate a grounded object search space. We then use an energy-based inference mechanism (Section 3.2) to discover the plausible actions that can be performed on the ground object space, driven by prior knowledge from symbolic knowledge bases, to recognize unseen and unknown activities (action-object combinations) without supervision. A visual-semantic action grounding mechanism (Sections 3.3) then provides feedback to ground semantic concepts with video-based evidence for discovering composite activities without explicit supervision. Although our framework is flexible to work with any noisy grounding model and knowledge base, we use CLIP [47] and ConceptNet [52], respectively.

Knowledge Representation. We use Grenander's pattern theory formalism [22] to represent the knowledge elements and build a contextualized activity interpretation that integrates neural and symbolic elements in a unified, energy-based representation. Pattern theory provides a flexible framework to help reason over variables with varying underlying dependency structures by representing them as compositions of simpler patterns. These structures, called configurations, are composed of atomic elements called *generators* ($\{g_1, g_2, \ldots g_i\} \in G_s$), which combine through local connections called *bonds* ($\{\beta_1, \beta_2, \ldots \beta_i\} \in g_i$). The collection of all generators is called the *generator space* (G_s), with each generator possessing an arbitrary set of bonds, defined by its *arity*. Bonds between generators are constrained through local and global *regularities*, as defined by an overarching graph structure. A probability structure over the representations captures the diversity of patterns. We refer the reader to Aakur *et al.* [2] and de Souza *et al.* [14] for a deeper exploration of the pattern theory formalism and Chapter 6 of [23] for its relation to other graphical models.

3.1 Evidence-based Object Grounding

The first step in our framework is to assess the plausibility of each object concept (represented as generators $\{g_1^o, g_2^o, \ldots g_i^o\} \in G_{obj}$) by *grounding* them in the input video V_i. In this work, we define *grounding* as gathering evidence from the input data to support the presence (or absence) of a concept in the final interpretation. While object-centric vision-language foundation models such as CLIP [47] have shown impressive abilities in zero-shot object recognition in images, egocentric videos provide additional challenges such as camera motion, lens distortion, and out-of-distribution object labels. Follow-up work [39] has focused on addressing them to a certain extent by probing CLIP for explainable object classification. However, they do not consider *compositional* properties of objects and alternative labels for verifying their presence in the video. To address this issue, we propose a neuro-symbolic *evidence-based* object grounding mechanism to compute the likelihood of an object in a given frame. For each object generator (g_i^o) in the search space (G_{obj}), we first compute a set of compositional *ungrounded* generators by constructing an ego-graph of each object label ($E_{g_i^o}$) from ConceptNet [52] and limiting edges to those that express *compositional* properties such as `IsA, UsedFor, HasProperty` and `SynonymOf`. Using ego-graph helps preserve the contextual information within the semantic locality of the object to filter high-order noise induced by regular k-hop neighborhoods. Given this set of *ungrounded* generators ($\{\bar{g}_i^o\} \forall g_i^o \in G_{obj}$), we then prompt CLIP to provide likelihoods for each ungrounded generator $p(\bar{g}_i^o|I_t)$ to compute the *evidence-based* likelihood for each *grounded* object generator $\underline{g}_i^o$ as defined by the probability function in Equation 1.

$$p(\underline{g}_i^o|\bar{g}_i^o, I_t, K_{CS}) = p(g_i^o|I_t) * \left\| \sum_{\forall \bar{g}_i^o} p(g_i^o, \bar{g}_i^o|E_{g_i^o}) * p(\bar{g}_i^o)|I_t) \right\| \qquad (1)$$

where $p(g_i^o, \bar{g}_i^o|E_{g_i^o})$ is the edge weight from the edge graph $E_{g_i^o}$ (sampled from a knowledge graph K_{CS}) that acts as a prior for each ungrounded evidence generator $\bar{g}_i^o$ and $p(\bar{g}_i^o)|I_t)$ is the likelihood from CLIP for its presence in each frame I_t. Hence the probability of the presence of a *grounded* object generator is determined by (i) its image-based likelihood, (ii) the image-based likelihood of its evidence generators, and (iii) support from prior knowledge for the presence of each evidence generator. Hence, we ground the object generators in each video frame by constructing and evaluating the evidence to support each grounding assertion and provide an interpretable interface to video object grounding. To navigate clutter and focus only on the object involved in the activity (i.e., the packaging problem [38]), we use a gaze-driven ROI selection process. Specifically, we take a fixed 200×200 region centered around the gaze position (from the human user if available, else we use the center bias [31] to approximate it) and use it as input to CLIP for object grounding.

3.2 Object-driven Activity Discovery

The second step in our approach is to discover plausible activities performed in the given video. Our approach is inspired by philosophical theories of knowledge [50], which hypothesize that each object is defined as such because of its affordance (actions permitted on it), which is constrained based on its "essence" or functionality. We take an object affordance-based approach to activity inference, constraining the activity label (verb+noun) to those that conform to affordances defined in prior knowledge. We first construct an "*action-object affinity*" function that provides a *prior* probability for the validity of an activity. Using ConceptNet as the source of knowledge, all possible paths between the action-object pair, which can include direct connections (if it exists) and indirect connections, are generated. We compute the prior probability of the action-object combination by taking a weighted sum of the edge weights along each path connecting them. Each term is weighted by an exponential decay function that reduces its contribution to the prior probability to avoid generating excessively long paths that can introduce noise into the reasoning process. Finally, we filter out paths that do *not* contain compositional assertions (`UsedFor, HasProperty, IsA`) since generic assertions such as (`RelatedTo`) may not capture the "essence" of the object to compute affordances. The probability of an activity (defined by an action generator g_i^a and a grounded object generator $\underline{g}_j^o$) is given by

$$p(g_i^a, \underline{g}_j^o | K_{CS}) = \underset{\forall E \in K_{CS}}{\arg\max} \sum_{(\bar{g}_m, \bar{g}_n) \in E} w_k * K_{CS}(\bar{g}_m, \bar{g}_n) \tag{2}$$

where E is the collection of all paths between g_i^a and $\underline{g}_j^o$ in a commonsense knowledge graph K_{CS}, w_i is a weight drawn from an exponential decay function based on the distance of the node $\bar{g}_n$ from g_i^a. After filtering for compositional properties, the path with the maximum weight is chosen with the optimal prior probability representative of the action-object affinity.

Energy-based Activity Inference. To reason over the different activity combinations, we assign an energy term to each activity label, represented as a *configuration*, a complex structure composed of individual generators that combine through bonds dictated by their affinity functions. In our case, each activity interpretation is a configuration composed of a grounded object generator ($\underline{g}_i^o$), its associated ungrounded evidence generators ($\bar{g}_j^o$), an action generator (g_k^a) and ungrounded generators from their affinity function, connected via an underlying graph structure. This graph structure will vary for each configuration depending on the presence of affinity-based bonds derived from ConceptNet. Hence, the *energy* of a configuration c_i is given by

$$E(c) = \phi(p(\underline{g}_i^o | \bar{g}_j^o, I_t, K_{CS})) + \phi(p(g_k^a, \underline{g}_i^o | K_{CS})) + \phi(p(g_k^a | I_t)) \tag{3}$$

where the first term provides the energy of grounded object generators (from Equation 1), the second term provides the energy from the affordance-based affinity between the action and object generators (from Equation 2), and the third term is the likelihood of an action generator. We initially set $\phi(p(g_k^a | I_t)) = 1$ to reason over all possible actions for each object and later update this using a posterior refinement process (Section 3.3). Hence, activity inference becomes an optimization over Equation 3 to find the configuration (or activity interpretation) with the least energy. For tractable computation, we use the MCMC-based simulated annealing mechanism proposed in KGL [5].

3.3 Visual-Semantic Action Grounding

The third step in our framework is the idea of visual-semantic action grounding, where we aim to learn to ground the inferred actions (verbs) from the overall activity interpretation. While CLIP provides a general purpose, if noisy, object grounding method, a comparable approach for actions does not exist. Hence, we learn an action grounding model by bootstrapping a simple function ($\psi(g_i^a, f_V)$) to map clip-level visual features to the semantic embedding space associated with ConceptNet, called ConceptNet Numberbatch [52]. The mapping function is a simple linear projection to go from the symbolic generator space ($g_i^a \in G_{act}$) to the semantic space (f_i^a), which is a 300-dimension ($\mathbb{R}^{1 \times 300}$) vector representation explicitly trained to capture concept-level attributes captured in ConceptNet. While there can be many sophisticated mechanisms [33, 6], including contrastive loss-based training, we use the mean squared error (MSE) loss as the objective function to train the mapping function since our goal is to provide a mechanism to ground abstract concepts from the knowledge-base in the video data. We leave the exploration of more sophisticated grounding mechanisms to future work.

Temporal Smoothing Since we predict frame-level activity interpretations to account for gaze transitions, we first perform temporal smoothing to label the entire video clip before training the mapping function $\psi(g_i^a, f_V)$ to reduce noise in the learning process. For each frame in the video clip, we take the five most common actions predicted at the *activity* level (considering the top-10 predictions) and sum their energies. This allows us to consolidate activity predictions and provides some leeway for erroneous predictions at the top-1 level. We then repeat the process for the entire clip,

Approach	Search Space	GTEA Gaze			GTEA GazePlus		
		Object	Action	Activity	Object	Action	Activity
Two-Stream CNN	Closed	38.05	59.54	53.08	61.87	58.65	44.89
IDT	Closed	45.07	75.55	40.41	53.45	66.74	51.26
Action Decomposition	Closed	**60.01**	**79.39**	**55.67**	**65.62**	**75.07**	**57.79**
Random	Known	3.22	7.69	2.50	3.70	4.55	2.28
Action Decomposition ZSL	Known	40.65	**85.28**	**39.63**	43.44	27.68	15.98
ALGO ZSL (Ours)	Known	**49.47**	74.74	27.34	**47.67**	**29.31**	**16.68**
KGL	Open	5.12	8.04	4.91	14.78	6.73	10.87
KGL+CLIP	Open	10.36	8.15	9.21	20.49	9.23	14.86
ALGO (Ours)	Open	**13.07**	**17.05**	**15.05**	**26.23**	**11.44**	**18.84**

Table 1: **Open-world activity recognition** performance on the GTEA Gaze and GTEA Gaze Plus datasets. We compare approaches with a closed search space, those with a known search space, and those with a partially open one. Accuracy is reported for predicted objects, actions, and activities.

i.e., get the top-5 actions based on their frequency of occurrence at the frame level and consolidated energies across frames. These five actions provide targets for the mapping function $\psi(g_i^a, f_V)$, which is then trained with the MSE function. We use the top-5 action labels as targets to restrict the influence of frequency bias from the commonsense knowledge base.

Posterior-based Activity Refinement. The final step in our framework is an iterative refinement process that updates the action concept priors (the third term in Equation 3) based on the predictions of the visual-semantic grounding mechanism described in Section 3.3. Since our predictions are made on a per-frame basis, it does not consider the overall temporal coherence and visual dynamics of the clip. Hence, there can be contradicting predictions for the actions done over time. Similarly, when setting the action priors to 1, we consider all actions equally plausible and do not restrict the action labels through grounding, as done for objects in Section 3.1. Hence, we iteratively update the action priors for the energy computation to re-rank the interpretations based on the clip-level visual dynamics. This prior could be updated to consider predictions from other models, such as EGO-VLP [33] through prompting mechanisms similar to our neuro-symbolic object grounding. However, we aim to iteratively refine the activity labels and update the visual-semantic action grounding modules simultaneously. We alternate between posterior update and action grounding until the generalization error (i.e., the performance on unseen actions) saturates, which indicates overfitting.

Implementation Details. We use an S3D-G network pre-trained by Miech *et al.* [40, 41] on Howto100M [40] as our visual feature extraction for visual-semantic action grounding. We use a CLIP model with the ViT-B/32 [17] as its backbone network. ConceptNet was used as our source of commonsense knowledge for neuro-symbolic reasoning, and ConceptNet Numberbatch [52] was used as the semantic representation for action grounding. The MCMC-based inference from KGL [5] was used as our reasoning mechanism. The mapping function, defined in Section 3.3, was a 1-layer feedforward network trained with the MSE loss for 100 epochs with a batch size of 256 and learning rate of 10^{-3}. Generalization errors on unseen actions were used to pick the best model. Experiments were conducted on a desktop with a 32-core AMD ThreadRipper and an NVIDIA Titan RTX.

4 Experimental Evaluation

We evaluate the proposed ALGO framework under two settings. In Section 4.1, we evaluate it on open-world inference, where the goal is to identify the activity in egocentric videos given only a known search space for each elementary concept, i.e., the ground truth activity is unknown. In Section 4.1.1, we map the inferred activity interpretation to the closest ground truth label and compare its performance against vision-language models. Finally, in Section 4.1.2, we evaluate the generalization capability of the learned action recognition model to unknown and unseen actions.

Data. To evaluate the open-world inference capabilities, we evaluate the approach on GTEA Gaze [20] and GTEA GazePlus [31] datasets, which contain egocentric, multi-subject videos of meal preparation activities. Since they have frame-level gaze information and activity labels, they provide an ideal test bed for our setup. The GTEA Gaze dataset consists of 14 subjects performing activities composed of 10 verbs and 38 nouns across 17 videos. The Gaze Plus dataset has 27 nouns

Approach	Visual Backbone	Pre-Training?	Pre-Training Data			mAP
			Ego?	Source	Size	
EGO-VLP w/o EgoNCE	TimeSformer [7]	VisLang	✗	Howto100M [40]	136M	9.2
EGO-VLP w/o EgoNCE	TimeSformer	VisLang	✗	CC3M+WebVid-2m	5.5M	20.9
EGO-VLP + EgoNCE	TimeSformer	VisLang	✓	EgoClip [33]	3.8M	23.6
HierVL	FrozenInTime	VisLang	✓	EgoClip	3.8M	26.0
LAVILA	TimeSformer	VisLang	✓	Ego4D	4M	**26.8**
ALGO (Ours)	S3D-G [40]	Vision Only	✗	Howto100M	136M	**17.3**
ALGO (Ours)	S3D [58]	Vision Only	✗	Kinetics-400 [28]	240K	16.8

Table 2: Evaluation of ALGO under **zero-shot** learning settings on Charades-Ego where the search space is constrained to ground truth activity semantics. VisLang: Vision Language Pre-Training.

and 15 verbs from 6 subjects performing 7 meal preparation activities across 37 videos. The gaze information is collected at 30 frames per second for both datasets. We also evaluate on Charades-Ego, a larger egocentric video dataset focused on activities of daily living, to evaluate on the zero-shot setting. It contains 7,860 videos containing 157 activities. Following prior work [33], we use the 785 egocentric clips in the test set for evaluation.

Evaluation Metrics. Following prior work in open-world activity recognition [5, 2], we use accuracy to evaluate action and object recognition and use *word-level* accuracy for evaluating the activity (verb+noun) recognition performance. It provides a less-constrained measurement to measure the quality of predictions beyond accuracy by considering all units without distinguishing between insertions, deletions, or misclassifications. This allows us to quantify the performance while not penalizing semantically similar interpretations. To evaluate the zero-shot learning setup, we use the official class-wise mAP metric as defined in the benchmark [49]. Finally, to measure the generalization capability of the approach to unknown actions, we use the word similarity score (denoted as NB-WS) to measure the semantic similarity between the predicted and ground truth actions. NB-WS has demonstrated the ability to capture attribute-based representations when computing similarity [51].

Baselines. We compare against various egocentric action recognition approaches, including those with a closed-world learning setup. For open-world inference, we compare it against Knowledge Guided Learning (KGL) [5], which introduced the notion of open-world egocentric action recognition. We also create a baseline called "KGL+CLIP" by augmenting KGL with CLIP-based grounding by including CLIP's similarity score for establishing semantic correspondences. We also compare with supervised learning models such as Action Decomposition [61], IDT [55], and Two-Stream CNN [37], with a strong closed-world assumption and a dependency on labeled training data. We also compare against the zero-shot version of Action Decomposition, which can work under a known world where the final activity labels are known. For zero-shot inference, we compare against large vision-language models, such as EGO-CLP [33], HierVL [6], and LAVILA [62].

4.1 Open World Activity Recognition

Table 1 summarizes the evaluation results under the open-world inference setting. Top-1 prediction results are reported for all approaches. As can be seen, CLIP-based grounding significantly improves the performance of object recognition for KGL, as opposed to the originally proposed, prior-only correspondence function. However, our neuro-symbolic grounding mechanism (Section 3.1) improves it further, achieving an object recognition performance of 13.0% on Gaze and 26.33% on Gaze Plus. It is interesting to note that naïvely adding CLIP as a mechanism for grounding objects, while effective, does not provide significant gains in the overall action recognition performance (an average of 2% across Gaze and Gaze Plus). We attribute it to the fact that the camera motion inherent in egocentric videos introduces occlusions and visual variations that make it hard for consistent recognition of actions. Evidence-based grounding, as proposed in ALGO, makes it more robust to such changes and improves the performance of both object and action recognition. Similarly, the posterior-based action refinement module (Section 3.3) helps achieve a top-1 action recognition performance of 17.05% on Gaze and 11.44% on Gaze Plus, outperforming KGL (8.04% and 6.73%). Note that the predictions are not separate for verbs and nouns, but computed from the final predicted activity. These are remarkable results, considering that the search space is open, i.e., the verb+noun combination is unknown and can be large (380 combinations for Gaze and 405 for Gaze Plus).

Training Data		Evaluation Data		Unknown Verbs?	Search Space	Accuracy	NB-WS
Dataset	**# Verbs**	**Dataset**	**# Verbs**				
Gaze	10	Gaze	10	✗	K	14.11	27.24
Gaze Plus	15	Gaze Plus	15	✗	K	11.44	24.45
Charades-Ego	33	Charades-Ego	33	✗	K	11.92	36.02
Gaze	10	Charades-Ego	33	✓	K	13.55	34.83
Gaze Plus	15	Charades-Ego	33	✓	K	10.24	31.11
Gaze Plus	15	Gaze	10	✓	K	5.27	29.68
Charades-Ego	33	Gaze	10	✓	K	10.17	32.65
Gaze	10	Gaze Plus	15	✓	K	10.37	23.55
Charades-Ego	33	Gaze Plus	15	✓	K	11.22	24.25
Gaze	10	Gaze	10	✓	U	9.87	14.51
Gaze Plus	15	Gaze Plus	15	✓	U	8.45	11.78

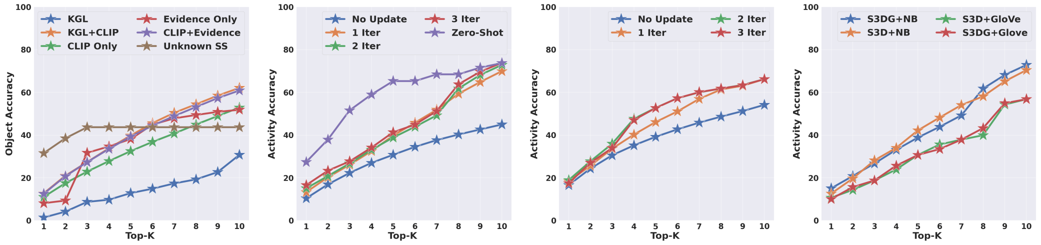
Table 3: **Generalization studies** to analyze the performance of the action (verb) recognition models learned in an open-world setting. The models are trained in one domain and evaluated in another, containing possible unknown and unseen actions. NB-WS: ConceptNet Numberbatch Word Similarity

4.1.1 Extension to Zero-Shot Egocentric Activity Recognition

Open-world learning involves the combinatorial search over the different, plausible compositions of elementary concepts. In activity recognition, this involves discovering the action-object (verb-noun) combinations that make up an activity. However, in many applications such as zero-shot recognition, the search space is known, and there is a need to predict pre-specified labels. To compare our approach with such foundation models, we evaluate ALGO on the Charades-Ego dataset and summarize the results in Table 2. We consider the top-10 interpretations made for each clip and perform a nearest neighbor search using ConceptNet Numberbatch embedding to the set of ground-truth labels and pick the one with the least distance. It provides a simple yet effective mechanism to extend our approach to zero-shot settings. We achieve an mAP score of 16.8% using an S3D [58] model pre-trained on Kinetics-400 [28] and an S3D-G [41] model pre-trained on Howto100M [40]. This significantly outperforms a comparable TimeSFormer [7] model pre-trained with a vision-language alignment objective function and provides competitive performance to state-of-the-art vision-language models, with significantly lower training requirements. We observe a similar performance in the Gaze and GazePlus datasets as shown in Table 1. We obtain 27.34% on Gaze and 16.69% on Gaze Plus, performing competitively with the zero-shot approaches. These results are obtained without large amounts of paired text-video pairs and a simple visual-semantic grounding approach. Diverse datasets and better alignment methods will help reduce the need for large-scale pre-training.

4.1.2 Generalization of Learned Actions to Unknown Vocabulary

We evaluate ALGO's ability to recognize actions from out of its training distribution by presenting videos from datasets with unseen actions and an unknown search space. Specifically, we refer to actions not in the original training domains as "unseen" actions, following convention from zero-shot learning. Similarly, in an unknown search space, i.e., *completely open world inference*, the search space is not pre-specified but inferred from general-purpose knowledge sources. For these experiments, we prompted GPT-4 [10] using the ChatGPT interface to provide 100 everyday actions that can be performed in the kitchen to construct our search space. The results are summarized in Table 3, where we present the verb accuracy and the ConceptNet Numberbatch Word Similarity (NB-WS) score. ALGO generalizes consistently across datasets. Of particular interest is the generalization from Gaze and Gaze Plus to Charades-Ego, where there is a significantly higher number of unseen and unknown actions. Models trained on GTEA Gaze, which has more variation in camera quality and actions, generalize better than those from Gaze Plus. With unseen actions and unknown search space, the performance was competitive, achieving an accuracy of 9.87% on Gaze and 8.45% on Gaze Plus. NB-WS was higher, indicating better agreement with the ground truth.



(a) (b) (c) (d)

Figure 2: **Ablation studies** showing (a) the quality of different object grounding techniques, (b) the impact of posterior-based action refinement, (c) the impact of iterative action refinement on generalization capabilities, and (d) the choice of visual and semantic representations.

4.2 Ablation Studies

We systematically examine the impact of the individual components on the overall performance of the approach. We experiment on the GTEA Gaze dataset and discuss the results below.

Quality and Impact of Object Grounding. First, we evaluate the object recognition performance of different object grounding techniques and present results in Figure 2(a). We consider 5 different techniques: the prior-based approach proposed in KGL, updating the prior with CLIP-based likelihood (KGL+CLIP), näively using CLIP to recognize the object in the gaze-based ROI (CLIP Only), the proposed evidence-based object grounding (CLIP+Evidence), and using evidence only without checking object-level likelihood (Evidence Only). As can be seen, using CLIP improves performance significantly across the different approaches while using evidence provides gains over the näive CLIP Only method. KGL+CLIP and the proposed CLIP+Evidence approaches perform similarly, with KGL+CLIP being slightly better when considering more than the top-5 recognized objects. We also evaluated the performance of CLIP+Evidence on an unknown search space by prompting GPT-4 to provide a list of 100 objects commonly found in the kitchen. The Top-3 performance is excellent, reaching 45% before saturating, which is remarkable considering that the *unknown* search space.

Impact of Posterior-based Action Refinement. One of the major contributions of ALGO is the use of continuous posterior-based action refinement, where the energy of the action generator is refined based on an updated prior from the visual-semantic action grounding to improve the activity recognition performance. Figure 2(b) visualizes the activity recognition performance with different levels of iteration, along with the results of a constrained search space (zero-shot) approach. As can be seen, the first two iterations significantly improved the performance, while the third iteration provided very negligible improvement, which provided indications of overfitting. Constraining the search space in the zero-shot setting significantly improves the performance.

Generalization of Visual-Semantic Action Grounding. To evaluate the impact of the posterior-based refinement on the generalization capabilities, we evaluated the trained models, at different iterations, on the GTEA Gaze Plus dataset. As can be seen from Figure 2, each iteration improves the performance of the model before the performance starts to stagnate (at the third iteration). These results indicate that while iterative refinement is useful, it can lead to overfitting to the domain-specific semantics and can hurt the generalization capabilities of the approach. To this end, we keep the termination criteria for the iterative posterior-based refinement based on the generalization performance of the action grounding model on unseen actions.

Impact of Visual-Semantic Features. Finally, we evaluate ALGO with different visual and semantic features and visualize the results in Figure 2 (d). We see that the use of ConceptNet Numberbatch (NB) considerably improves the performance of the approach as opposed to using GloVe embeddings [44]. The choice of visual features (S3DG vs. S3D) does not impact the performance much. We hypothesize that the NB's ability to capture semantic attributes [51] allows it to generalize better than GloVe.

5 Discussion, Limitations, and Future Work

In this work, we proposed ALGO, a neuro-symbolic framework for open-world egocentric activity recognition that aims to learn novel action and activity classes without explicit supervision. By grounding objects and using an object-centered, knowledge-based approach to activity inference, we reduce the need for labeled data to learn semantic associations among elementary concepts. We demonstrate that the open-world learning paradigm is an effective inference mechanism to distill commonsense knowledge from symbolic knowledge bases for grounded action understanding. While showing competitive performance, there are two key limitations: (i) it is restricted to ego-centric videos due to the need to navigate clutter by using human attention as a contextual cue for object grounding, and (ii) it requires a defined search space to arrive at an interpretation. While we demonstrated its performance on an unknown search space, much work remains to effectively build a search space (both action and object) to move towards a truly open-world learning paradigm. We aim to explore the use of attention-based mechanisms [42, 1] to extend the framework to third-person videos and using abductive reasoning [4, 60] to integrate visual commonsense into the reasoning.

6 Acknowledgements

Acknowledgements. This research was supported in part by the US National Science Foundation grants IIS 2143150, and IIS 1955230. We thank Dr. Anuj Srivastava (FSU) and Dr. Sudeep Sarkar (USF) for their thoughtful feedback during the discussion about the problem formulation and experimental analysis phase of the project.

References

[1] S. Aakur and S. Sarkar. Actor-centered representations for action localization in streaming videos. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, pages 70–87. Springer, 2022.

[2] S. Aakur, F. de Souza, and S. Sarkar. Generating open world descriptions of video using common sense knowledge in a pattern theory framework. *Quarterly of Applied Mathematics*, 77(2):323–356, 2019.

[3] S. N. Aakur and A. Bagavathi. Unsupervised gaze prediction in egocentric videos by energy-based surprise modeling. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2021.

[4] S. N. Aakur and S. Sarkar. Abductive reasoning as self-supervision for common sense question answering. *arXiv preprint arXiv:1909.03099*, 2019.

[5] S. N. Aakur, S. Kundu, and N. Gunti. Knowledge guided learning: Open world egocentric action recognition with zero supervision. *Pattern Recognition Letters*, 156:38–45, 2022.

[6] K. Ashutosh, R. Girdhar, L. Torresani, and K. Grauman. Hiervl: Learning hierarchical video-language embeddings, 2023.

[7] G. Bertasius, H. Wang, and L. Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning*, pages 813–824. PMLR, 2021.

[8] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[9] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[10] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[11] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.

[12] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

[13] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141, 2020.

[14] F. D. de Souza, S. Sarkar, A. Srivastava, and J. Su. Pattern theory for representation and inference of semantic structures in videos. *Pattern Recognition Letters*, 72:41–51, 2016.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[16] N. Dong, Y. Zhang, M. Ding, and G. H. Lee. Open world detr: Transformer based open world object detection, 2022.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

[18] Y. Du, F. Wei, Z. Zhang, M. Shi, Y. Gao, and G. Li. Learning to prompt for open-vocabulary object detection with vision-language model, 2022.

[19] C. Fan. Egovqa-an egocentric video question answering benchmark dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[20] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision*, pages 314–327. Springer, 2012.

[21] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[22] U. Grenander. *Elements of pattern theory*. JHU Press, 1996.

[23] U. Grenander, M. I. Miller, et al. *Pattern theory: from representation to inference*. Oxford University Press, 2007.

[24] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui. Open-vocabulary object detection via vision and language knowledge distillation, 2022.

[25] S. Han, B. Liu, R. Cabezas, C. D. Twigg, P. Zhang, J. Petkau, T.-H. Yu, C.-J. Tai, M. Akbay, Z. Wang, et al. Megatrack: monochrome egocentric articulated hand-tracking for virtual reality. *ACM Transactions on Graphics (ToG)*, 39(4):87–1, 2020.

[26] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021.

[27] J. Jiang and S. Ahn. Generative neurosymbolic machines. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12572–12582. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/94c28dcfc97557df0df6d1f7222fc384-Paper.pdf`.

[28] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[29] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

[30] H. Li, Y. Cai, and W.-S. Zheng. Deep dual relation modeling for egocentric interaction recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7932–7941, 2019.

[31] Y. Li, A. Fathi, and J. M. Rehg. Learning to predict gaze in egocentric video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3216–3223, 2013.

[32] Y. Li, F. Liang, L. Zhao, Y. Cui, W. Ouyang, J. Shao, F. Yu, and J. Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm, 2022.

[33] K. Q. Lin, A. J. Wang, M. Soldan, M. Wray, R. Yan, E. Z. Xu, D. Gao, R. Tu, W. Zhao, W. Kong, C. Cai, H. Wang, D. Damen, B. Ghanem, W. Liu, and M. Z. Shou. Egocentric video-language pretraining, 2022.

[34] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[35] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2714–2721, 2013.

[36] M. Ma, H. Fan, and K. M. Kitani. Going deeper into first-person activity recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[37] M. Ma, H. Fan, and K. M. Kitani. Going deeper into first-person activity recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1894–1903, 2016.

[38] M. J. Maguire and G. O. Dove. Speaking of events: event word learning and event representation. *Understanding Events: How Humans See, Represent, and Act on Events*, pages 193–218, 2008.

[39] S. Menon and C. Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022.

[40] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019.

[41] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020.

[42] R. Mounir, A. Shahabaz, R. Gula, J. Theuerkauf, and S. Sarkar. Towards automated ethogramming: Cognitively-inspired event segmentation for streaming wildlife video monitoring. *International Journal of Computer Vision*, pages 1–31, 2023.

[43] M. Nye, M. Tessler, J. Tenenbaum, and B. M. Lake. Improving coherence and consistency in neural sequence models with dual-system, neuro-symbolic reasoning. *Advances in Neural Information Processing Systems*, 34:25192–25204, 2021.

[44] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[45] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. 2018.

[46] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[47] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[48] M. S. Ryoo, B. Rothrock, and L. Matthies. Pooled motion features for first-person videos. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[49] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari. Actor and observer: Joint modeling of first and third-person videos. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7396–7404, 2018.

[50] A. Silverman. Plato's middle period metaphysics and epistemology. 2003.

[51] R. Speer and J. Lowry-Duda. Luminoso at SemEval-2018 task 10: Distinguishing attributes using text corpora and relational knowledge. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 985–989, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-1162. URL `https://aclanthology.org/S18-1162`.

[52] R. Speer, J. Chin, and C. Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[53] S. Sudhakaran, S. Escalera, and O. Lanz. Lsta: Long short-term attention for egocentric action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[55] H. Wang and C. Schmid. Action recognition with improved trajectories. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3551–3558, 2013.

[56] X. Wang, L. Zhu, H. Wang, and Y. Yang. Interactive prototype learning for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8168–8177, October 2021.

[57] T. Wu, M. Tjandrasuwita, Z. Wu, X. Yang, K. Liu, R. Sosic, and J. Leskovec. Zeroc: A neuro-symbolic model for zero-shot concept recognition and acquisition at inference time. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 9828–9840. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/3ff48dde82306fe8f26f3e51dd1054d7-Paper-Conference.pdf`.

[58] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision*, pages 305–321, 2018.

[59] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu. Coca: Contrastive captioners are image-text foundation models, 2022.

[60] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[61] Y. C. Zhang, Y. Li, and J. M. Rehg. First-person action decomposition and zero-shot learning. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 121–129, 2017.

[62] Y. Zhao, I. Misra, P. Krähenbühl, and R. Girdhar. Learning video representations from large language models, 2022.

[63] Y. Zhou, B. Ni, R. Hong, X. Yang, and Q. Tian. Cascaded interactional targeting network for egocentric video analysis. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.