EUROPEAN CONFERENCE ON COMPUTER VISION

MILANO
2024

Discovering Novel Actions from Open World Egocentric Videos with Object-Grounded Visual Commonsense Reasoning

Sanjoy Kundu, Shubham Trehan, Sathyanarayanan Aakur*

Auburn University

*Corresponding author

Motivation

- 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 like CLIP 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, this target search space can be unknown or exceptionally large and can restrict their performance.

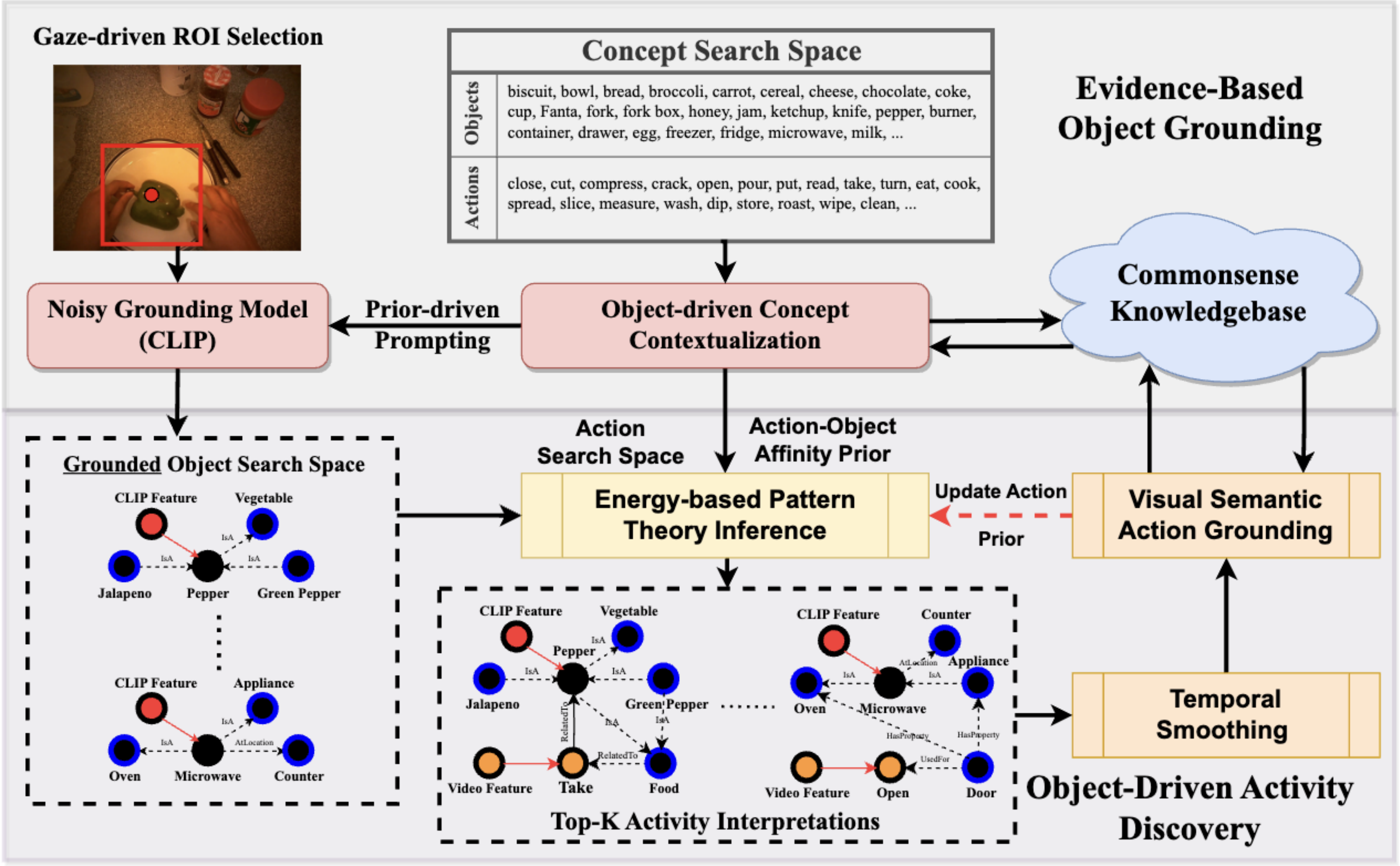
What is an open world?

- We define an activity as a complex structure whose semantics are expressed by a combination of actions (verbs) and objects (nouns).
- Semantics can be learned in different settings:
  - Supervised learning: closed world. There is a 1:1 mapping between training-time and test-time semantics.
    - Labels of samples in the test set are always seen in the training set.
  - Zero-shot learning: known world. There is an overlapping mapping between training and test set.
    - Labels of samples in test set are known but not always seen during training.
  - Open-world learning: open world. Semantics of labels unknown during both train- and test-times.
    - Elementary concept labels (objects, verbs/actions) are known without any corresponding examples. Goal is to learn to associate them during training and inference.

Intuition

- Locating and recognizing *objects* in egocentric videos can provide context for recognizing *actions* through affordance-based reasoning.
  - Knowing the object restricts the actions that can be performed on them.
- Our approach is inspired by philosophical theories of knowledge, 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.
- For example, a *chair* is part of a larger concept space called *furniture* whose purpose is constrained to a set of actions such as *sitting* and *standing*.
- 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.

Framework



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

Action Search Space

Action-Object Affinity Prior

Energy-based Pattern Theory Inference

Update Action Prior

Visual Semantic Action Grounding

Top-K Activity Interpretations

Temporal Smoothing

Object-Driven Activity Discovery

Step 1: Object Grounding

- The first step in our framework is to assess the plausibility of each object concept by grounding them in the input video.
- 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 in the search space, we first compute a set of compositional *ungrounded* generators by constructing an ego-graph of each object label from ConceptNet.

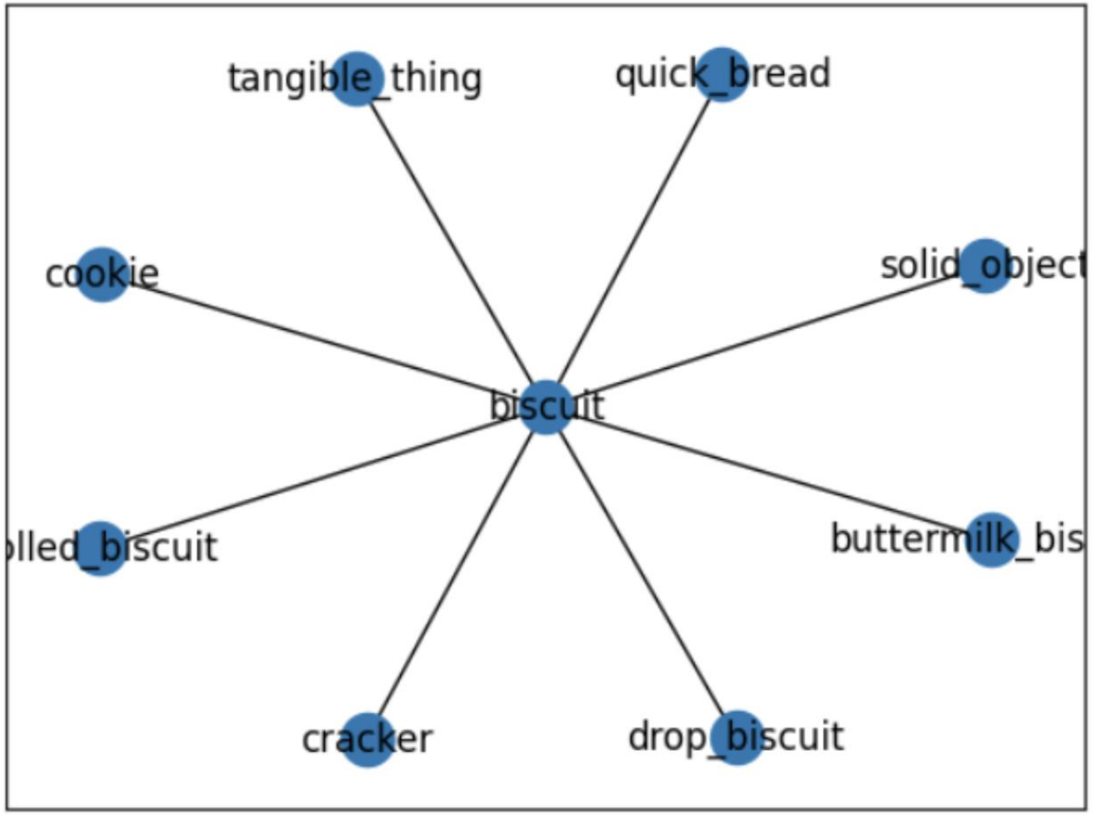
EUROPEAN CONFERENCE ON COMPUTER VISION
MILANO
2024

Step 1: Object Grounding

- For each object generator in the search space, we first compute a set of compositional *ungrounded* generators to assess its probability.

$$p(\underline{g}_i^o | \bar{g}_i^o, I_t, K_{CS}) = p(\underline{g}_i^o | I_t) * p(\bar{g}_i^o | I_t, K_{CS})$$

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



Step 2: Object-driven Activity Discovery

- We first construct an action-object affinity function that provides a *prior* probability for the validity of an activity.
- The probability of each action-object combination is computed by taking a weighted sum of the edge weights along each path (direct and indirect) that connects them in ConceptNet.
- An exponential decay function is applied to each term to avoid generating excessively long paths that can introduce noise into the reasoning process.

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

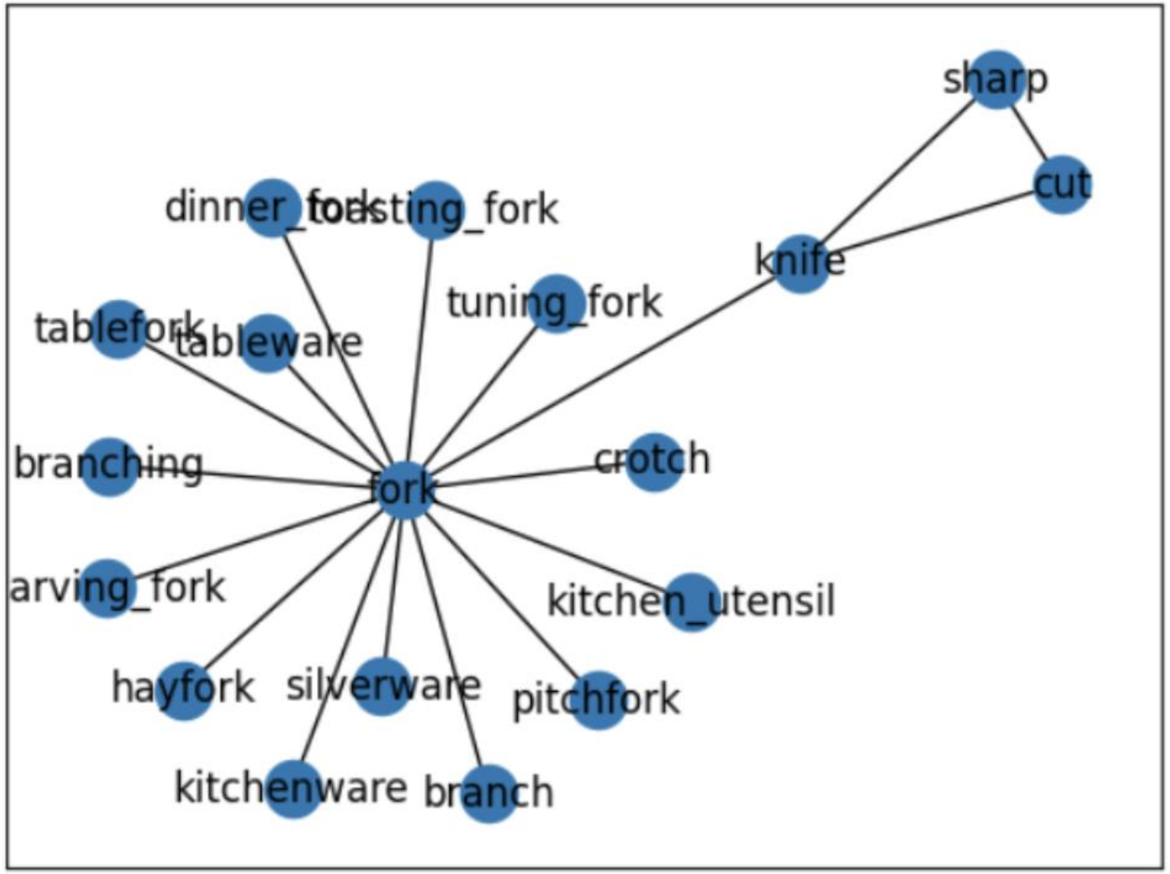
Step 2: Object-driven Activity Discovery

- To reason over the different activity combinations, we assign an energy term to each activity label, represented as a *configuration*.
- Each activity interpretation is a configuration composed of a grounded object generator (g_i^o), its associated ungrounded evidence generators ($\overline{g_j^o}$), an action generator (g_k^a) and ungrounded generators from their affinity function, connected via a graph structure.

$$E(c) = \overbrace{\phi(p(\underline{g}_i^o|\bar{g}_j^o, I_t, K_{CS}))}^{\text{Energy of grounded object generators}} + \overbrace{\phi(p(g_k^a, \underline{g}_i^o|K_{CS}))}^{\text{Energy of action generators given the object of interest}}$$

$$+ \underbrace{\phi(p(g_k^a|I_t))}_{\text{Prior probability of action generators}}$$

Example of an interpretation



Step 3: Visual-semantic Action Grounding

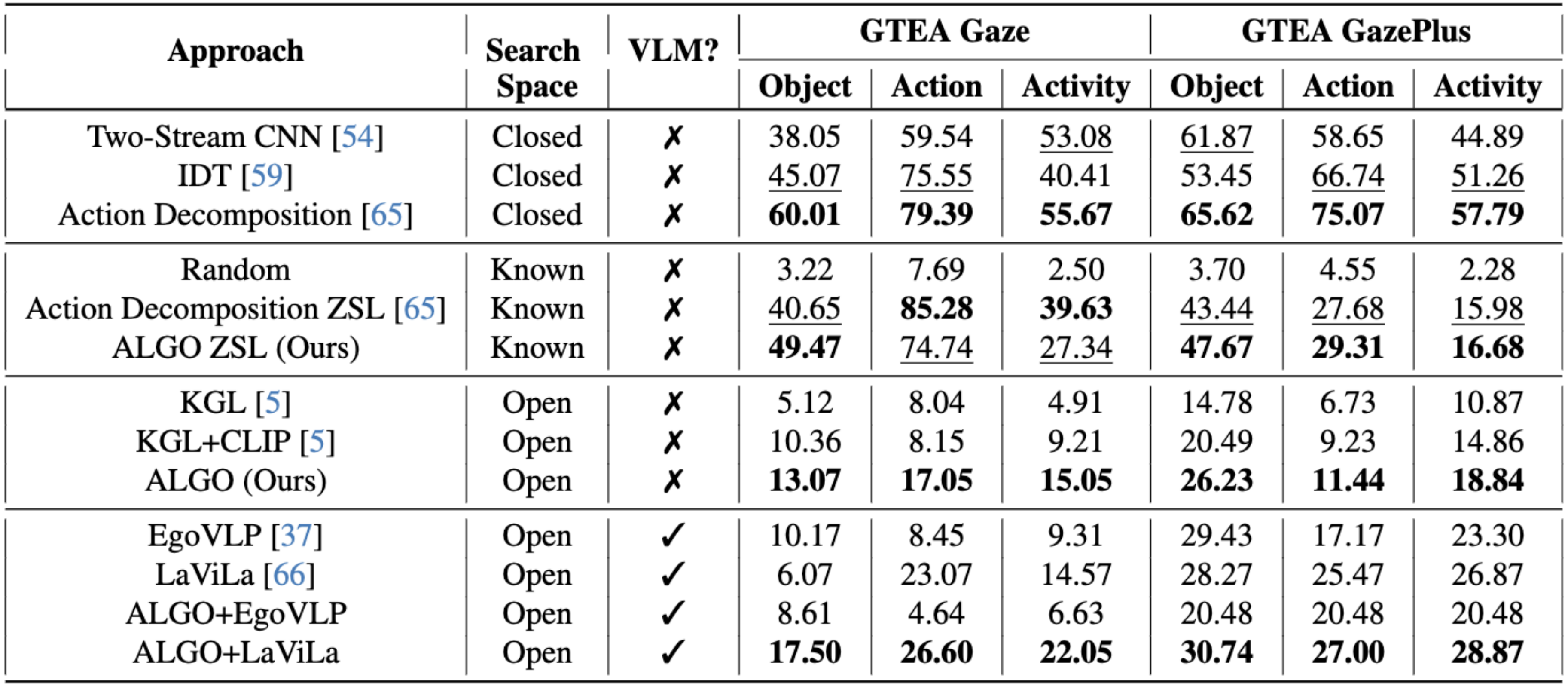
- The third step in our framework is the idea of visual-semantic action grounding, where we aim to learn to ground the inferred actions (verbs) from the overall activity interpretation.
- We learn an action grounding model by bootstrapping a simple function ($\psi(g_i^a, f_V)$) to map clip-level visual features to the semantic embedding space associated with ConceptNet, called ConceptNet Numberbatch. 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 300d vector representation trained on ConceptNet.
- **Temporal Smoothing:** we 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 and sum their energies to consolidate activity predictions and account for erroneous predictions.
  - We use the top-5 action labels as targets to limit the effect of frequency bias.
  - temporal smoothing acts as a regularizer to reduce overfitting by forcing the model to predict the embedding for the top five actions for each video clip.

Step 3: Visual-semantic Action Grounding

- **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 4) based on the predictions of the visual-semantic grounding mechanism.
    - 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.
    - 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.
    - Hence, we iteratively update the action priors for the energy computation to re-rank the interpretations based on the clip-level visual dynamics.

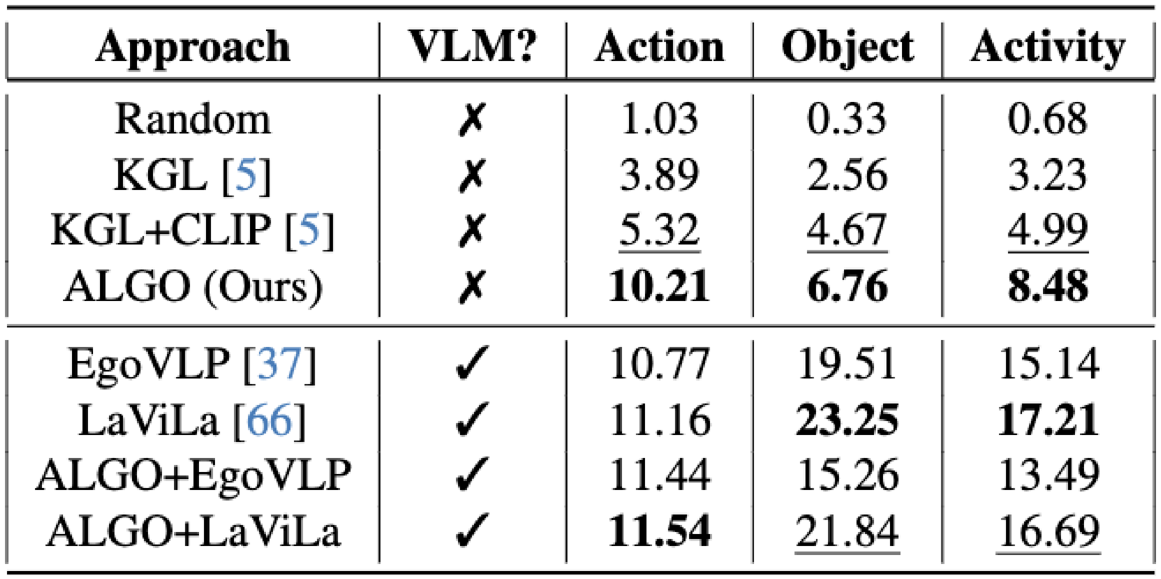
EUROPEAN CONFERENCE ON COMPUTER VISION
MILANO
2024

Performance on GTEA Gaze and GTEA Gaze+



Approach	**Search Space**	**VLM?**	**GTEA Gaze**			**GTEA GazePlus**		
			Object	**Action**	**Activity**	**Object**	**Action**	**Activity**
Two-Stream CNN [54]	Closed	✗	38.05	59.54	53.08	61.87	58.65	44.89
IDT [59]	Closed	✗	45.07	75.55	40.41	53.45	66.74	51.26
Action Decomposition [65]	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 [65]	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 [5]	Open	✗	5.12	8.04	4.91	14.78	6.73	10.87
KGL+CLIP [5]	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**
EgoVLP [37]	Open	✓	10.17	8.45	9.31	29.43	17.17	23.30
LaViLa [66]	Open	✓	6.07	23.07	14.57	28.27	25.47	26.87
ALGO+EgoVLP	Open	✓	8.61	4.64	6.63	20.48	20.48	20.48
ALGO+LaViLa	Open	✓	**17.50**	**26.60**	**22.05**	**30.74**	**27.00**	**28.87**

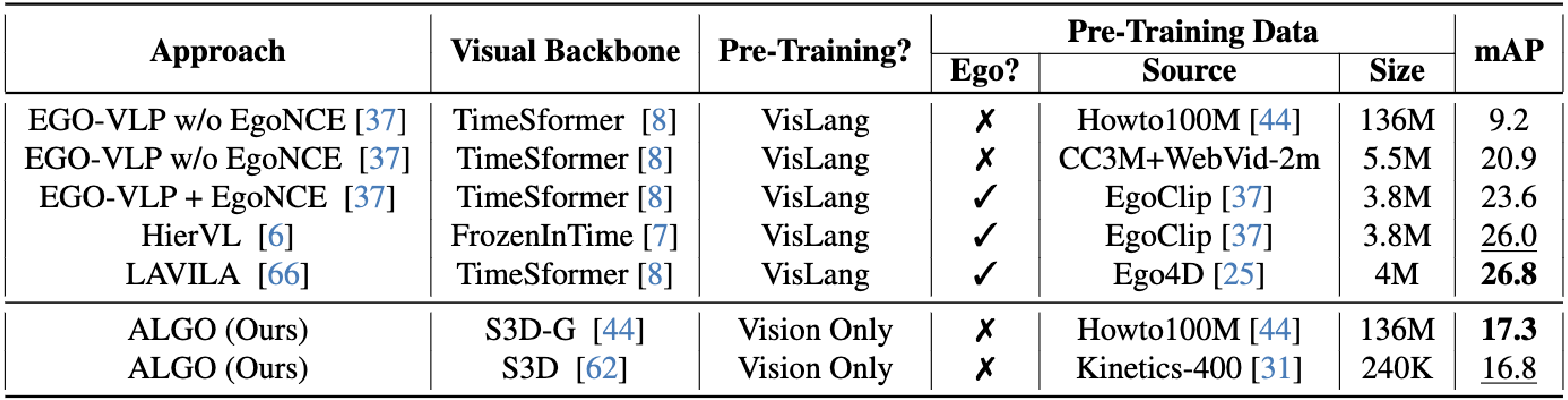
Performance on EPIC-Kitchens-100



Approach	VLM?	Action	Object	Activity
Random	✗	1.03	0.33	0.68
KGL [5]	✗	3.89	2.56	3.23
KGL+CLIP [5]	✗	5.32	4.67	4.99
ALGO (Ours)	✗	**10.21**	**6.76**	**8.48**
EgoVLP [37]	✓	10.77	19.51	15.14
LaViLa [66]	✓	11.16	**23.25**	**17.21**
ALGO+EgoVLP	✓	11.44	15.26	13.49
ALGO+LaViLa	✓	**11.54**	21.84	16.69

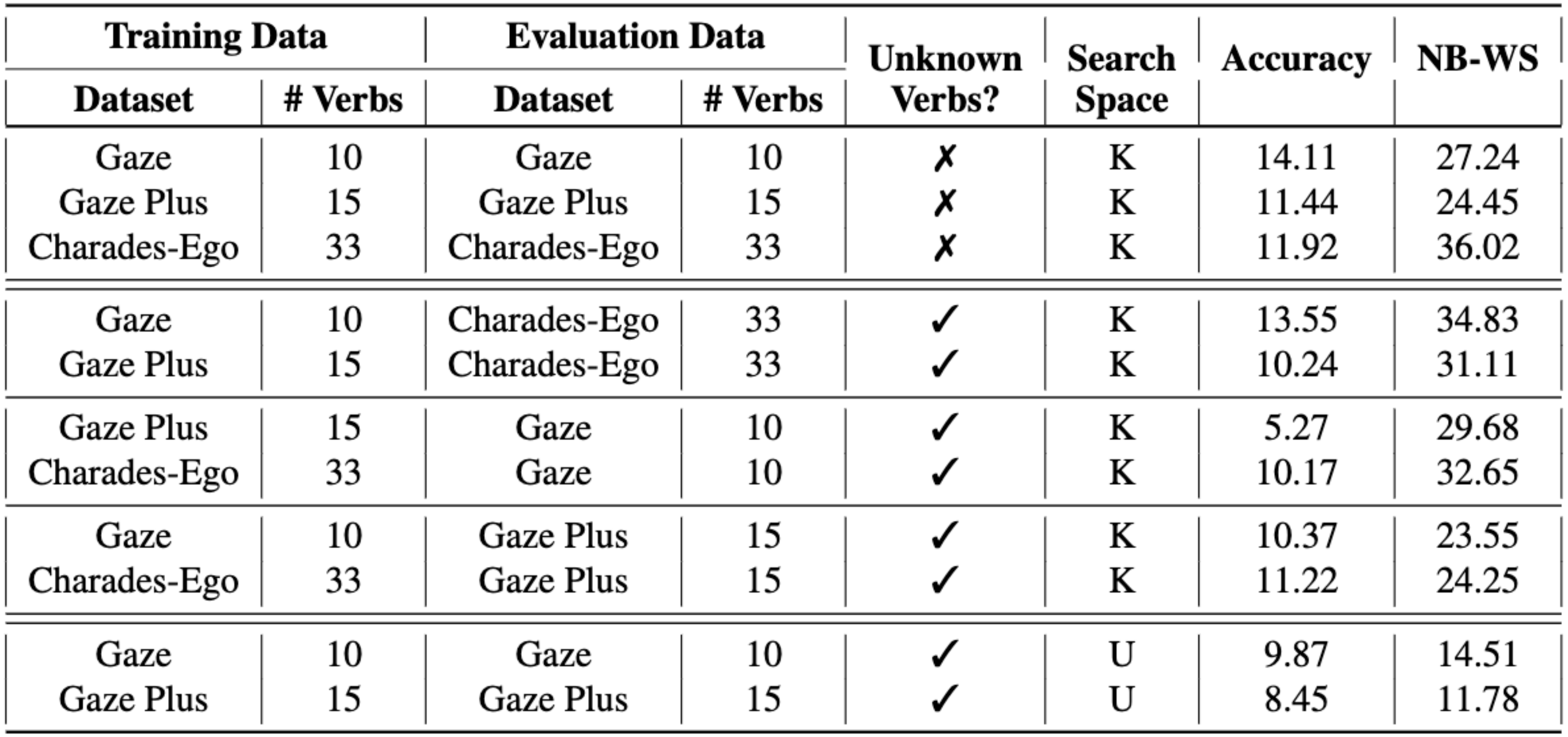
Table 2. Evaluation on the EPIC-Kitchens-100 dataset. VLM: Vision-Language pre-training on egocentric data. Accuracy for actions, objects, and activity are reported.

Performance on Charades-Ego under Zero-Shot settings



Approach	Visual Backbone	Pre-Training?	Pre-Training Data			mAP
			Ego?	Source	Size	
EGO-VLP w/o EgoNCE [37]	TimeSformer [8]	VisLang	✗	Howto100M [44]	136M	9.2
EGO-VLP w/o EgoNCE [37]	TimeSformer [8]	VisLang	✗	CC3M+WebVid-2m	5.5M	20.9
EGO-VLP + EgoNCE [37]	TimeSformer [8]	VisLang	✓	EgoClip [37]	3.8M	23.6
HierVL [6]	FrozenInTime [7]	VisLang	✓	EgoClip [37]	3.8M	26.0
LAVILA [66]	TimeSformer [8]	VisLang	✓	Ego4D [25]	4M	**26.8**
ALGO (Ours)	S3D-G [44]	Vision Only	✗	Howto100M [44]	136M	**17.3**
ALGO (Ours)	S3D [62]	Vision Only	✗	Kinetics-400 [31]	240K	16.8

Generalization Studies



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

Acknowledgements

- This research was supported in part by the US National Science Foundation grants IIS 2348689 and IIS 2348690. We thank Dr. Anuj Srivastava (FSU) and Dr. Sudeep Sarkar (USF) for their thoughtful feedback during the discussion about the project's problem formulation and experimental analysis phase.