This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# ProbRes: Probabilistic Jump Diffusion for Open-World Egocentric Activity Recognition

Sanjoy Kundu*, Shanmukha Vellamcheti*, Sathyanarayanan N. Aakur
CSSE Department, Auburn University
Auburn, Alabama, USA 36849
{szk0266, szv0080, san0028}@auburn.edu

## Abstract

*Open-world egocentric activity recognition poses a fundamental challenge due to its unconstrained nature, requiring models to infer unseen activities from an expansive, partially observed search space. We introduce ProbRes, a Probabilistic Residual search framework based on jump-diffusion that efficiently navigates this space by balancing prior-guided exploration with likelihood-driven exploitation. Our approach integrates structured commonsense priors to construct a semantically coherent search space, adaptively refines predictions using Vision-Language Models (VLMs) and employs a stochastic search mechanism to locate high-likelihood activity labels while minimizing exhaustive enumeration efficiently. We systematically evaluate ProbRes across multiple openness levels (L0–L3), demonstrating its adaptability to increasing search space complexity. In addition to achieving state-of-the-art performance on benchmark datasets (GTEA Gaze, GTEA Gaze+, EPIC-Kitchens, and Charades-Ego), we establish a clear taxonomy for open-world recognition, delineating the challenges and methodological advancements necessary for egocentric activity understanding. Our results highlight the importance of structured search strategies, paving the way for scalable and efficient open-world activity recognition.*

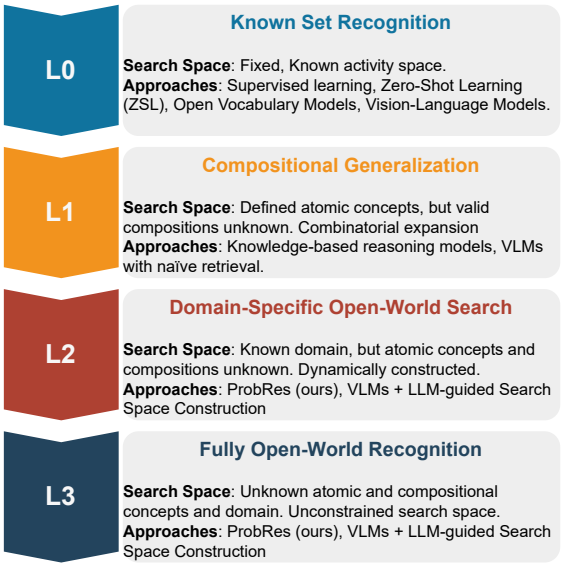


Figure 1. *Taxonomy of Openness in Egocentric Activity Recognition.* We define four levels of openness based on search space constraints: L0 (fixed activity set), L1 (known atomic concepts, unknown compositions), L2 (known domain, inferred activities), and L3 (fully unconstrained search).

## 1. Introduction

Egocentric activity recognition in open-world settings presents a fundamental challenge in perception-driven reasoning, where an intelligent system must infer ongoing activities from a vast, unconstrained space of possibilities. Unlike closed-set classification, where predefined categories limit ambiguity, open-world recognition requires models to adaptively infer plausible activities, even when faced with unknown or unseen combinations of actions and objects. This challenge is particularly relevant for assistive robotics, wearable AI, and autonomous agents, where real-time understanding of human actions is crucial for interaction and decision-making. While multimodal foundation models such as Vision-Language Models (VLMs) have demonstrated remarkable zero-shot generalization, their reliance on exhaustive enumeration for inference makes them inefficient for large-scale open-world reasoning. To bridge this gap, structured knowledge is essential to move beyond naive brute-force enumeration, enabling scalable and efficient activity understanding in real-world applications.

The increasing focus on open-world learning [2, 13, 20, 26] has led to ambiguity in defining openness, as different

*Equal Contribution

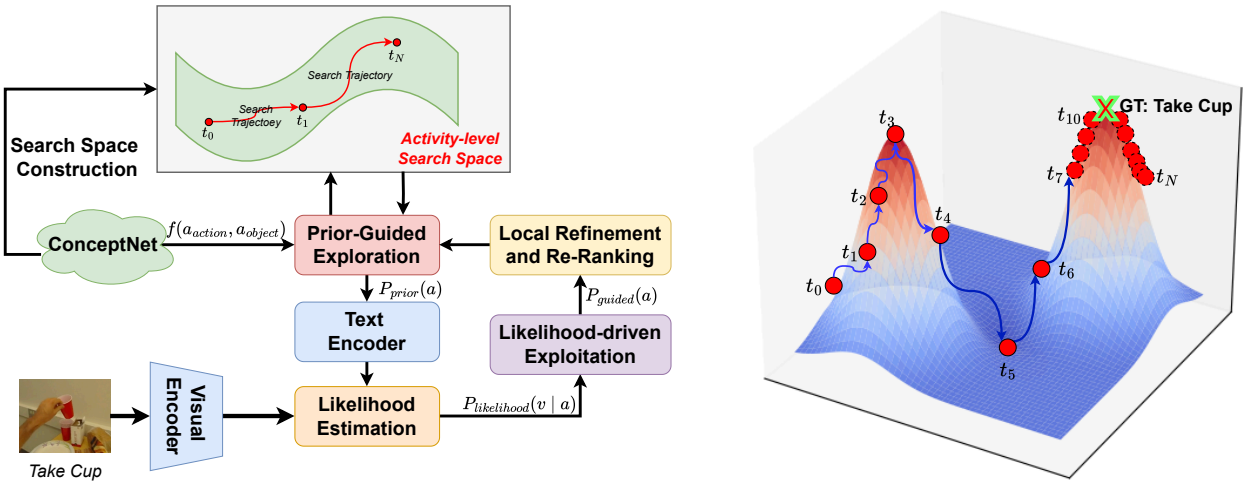


Figure 2. **(a) ProbRes framework** for open-world egocentric activity recognition. The search space is structured using ConceptNet priors, enabling guided exploration. The model iteratively refines candidates via likelihood estimation, balancing exploration and exploitation, followed by local refinement and re-ranking. **(b) Search trajectory visualization**, showing how ProbRes navigates likelihood regions to reach high-confidence predictions near the ground truth.

factors, objects, actions, and domains can independently contribute to it. Without a clear distinction, comparing methods and assessing generalization remain challenging. To address this, we define a hierarchy of openness levels that systematically capture increasing uncertainty in the search space (Figure 1). At **L0**, as in traditional zero-shot learning, all possible activity categories (action-object combinations) are predefined, but some may lack training examples. This setting benefits directly from Vision-Language Models (VLMs), which can generalize to unseen instances but operate within a bounded conceptual space. At **L1**, the atomic concept space — objects and actions — is known, but their composition (activities) is undefined, requiring models to reason about plausible pairings. At **L2**, the domain is known (e.g., cooking), but the underlying search space is unconstrained. Finally, at **L3**, no prior knowledge about the domain, actions, or objects is available, requiring models to construct the entire search space on the fly. As openness increases, exhaustive enumeration becomes intractable and requires structured priors for adaptive search.

To address these, we introduce Probabilistic Residual Search (ProbRes), a structured search framework designed to navigate unconstrained activity spaces efficiently. Instead of relying solely on VLM-based likelihood estimation, ProbRes integrates structured priors from external knowledge sources to guide exploration, ensuring semantically meaningful search trajectories without supervision. As illustrated in Figure 2, ProbRes operates in three key phases: (1) exploration, where prior-driven sampling identifies plausible activity candidates; (2) exploitation, where the search refines predictions based on VLM likelihoods; and (3) residual refinement, which further optimizes results by re-ranking activities based on their decomposed action-object components. By balancing structured priors with data-driven reasoning, ProbRes efficiently locates high-likelihood activities while drastically reducing VLM queries.

Our **contributions** are four-fold: (i) we propose *ProbRes*, a probabilistic residual search strategy that integrates structured priors with likelihood-driven inference, effectively balancing exploration and exploitation for open-world activity recognition, (ii) we introduce a clear hierarchical taxonomy (L0–L3) to systematically define and evaluate open-world recognition challenges, providing a principled foundation for future research, (iii) we show that ProbRes drastically reduces VLM queries while maintaining or improving recognition accuracy, demonstrating the importance of structured priors in large-scale search, and (iv) we highlight the limitations of current VLM-based text embeddings for search efficiency with qualitative analysis and highlight the need for improved semantic structuring.

## 2. Related Work

**Egocentric video understanding** has gained significant traction in the computer vision community in recent years due to its critical applications in areas such as imitation learning for robotic skills [45], virtual reality [21], and autonomous skill acquisition [4, 22, 27]. A wide range of tasks have been proposed to analyze egocentric video understanding, including procedure understanding, Ego-

Pose [18, 19], proficiency estimation, question answering [15], gaze prediction [1, 16, 29], activity recognition [28], and summarization [33], driven by the development of large-scale datasets such as EgoExo4D [19], Ego4D [18], EPIC-Kitchens [10], Charades-Ego [46], GTEA Gaze [29], GTEA Gaze Plus [16], and Assembly101 [44].

Despite these advancements, action recognition in egocentric videos remains challenging, particularly under conditions of rapid camera movement and significant occlusion, leading to issues with accuracy and generalization across different environments. While early approaches to egocentric action recognition were dominated by supervised learning methods, such as spatial-temporal modeling [49], motion-based representations [34], hand-object interactions [53, 56, 62], and time-series modeling [43], recent research has shifted towards self-supervised learning [57, 59], leveraging large-scale pretraining followed by fine-tuning for downstream tasks. Both paradigms remain data-driven and require extensive training to achieve optimal performance. Multi-modal LLM-based models such as MM-Ego [58] and GPT4Ego [9] have recently been explored for egocentric action recognition. Zero-shot learning [46, 60] offers a potential avenue to reduce dependence on labeled datasets, but its effectiveness is constrained by the inherent limitations of pre-trained models in reasoning about open-world activity compositions.

**Open-world understanding**, particularly in egocentric activity recognition, remains an underexplored area. While significant progress has been made in open-world object detection, such as Open World DETR [13] and open-vocabulary object detection [14, 20], which leverage VLMs through prompting and knowledge distillation, challenges remain in handling novel class biases, as highlighted by UMB [55]. One of the first works addressing open-world egocentric activity recognition, KGL [2], employed a neuro-symbolic framework utilizing ConceptNet [47] for knowledge-driven inference. However, its reliance on object detectors for grounding concepts and the semantic gap between knowledge bases and video data limits its applicability. Chatterjee et al. [7] proposed a verb encoder and prompt-based object encoder for open-vocabulary action recognition in egocentric videos. More recently, ALGO [26] extended neuro-symbolic prompting to refine action-object affinities iteratively, combining object grounding with video foundation models. Other works explore related directions, such as incremental open-world action recognition [39] and skeleton-based methods for open-set action recognition [36].

**Vision-Language Models** have gained prominence following the success of transformer-based LLMs such as BERT [12], RoBERTa [32], GPT series [5, 6, 40, 41], and open-source models like LLaMA [17, 51, 52] and DeepSeek [11]. Image-language models, including CLIP [42], DeCLIP [30], and ALIGN [23], leverage large-scale image-text pairs with contrastive learning [8, 24] to achieve strong zero-shot classification. Video-language models extend these capabilities to egocentric video understanding. EGO-VLP [31] employs separate video and text encoders, Hier-VL [3] captures multi-scale video information hierarchically and LAVILA [61] enhances video-text embeddings with LLM-generated narrations. EgoVLPv2 [38] integrates cross-modal fusion, while ALANAVLM [50] advances large-scale multimodal pretraining for embodied AI. However, these models are inherently constrained by the fixed search space defined during inference, limiting their generalization ability in open-world scenarios.

## 3. ProbRes: Jump Diffusion-based Reasoning

**Overview.** We aim to infer an activity label $a \in \mathcal{S}$ from a given video feature $v$. Unlike conventional closed-set recognition [49], where all activities are known during training, open-world settings require reasoning over a vast, partially observed search space, making exhaustive enumeration computationally prohibitive. Given a prior probability distribution $P_{\text{prior}}(a)$ capturing activity occurrence, a Vision-Language Model (VLM) likelihood estimator $P_{\text{likelihood}}(v|a)$, and a search space $\mathcal{S}$, the goal is to estimate the most probable activity efficiently:

$$a^* = \arg\max_{a \in \mathcal{S}} P_{\text{likelihood}}(v \mid a). \tag{1}$$

However, directly evaluating all candidates in $\mathcal{S}$ is computationally inefficient and often infeasible as the search space grows. Instead, we propose *Probabilistic Residual Search (ProbRes)*, an adaptive search framework that efficiently navigates the text embedding space of activities while being guided by symbolic priors derived from a structured knowledge graph. The overall procedure is summarized in Algorithm 1. ProbRes operates in three key phases: (1) *Exploration*, where search hypotheses are sampled from a prior-driven distribution; (2) *Exploitation*, where search is refined using VLM-based likelihoods; and (3) *Residual Refinement*, where candidates are refined using fine-grained re-ranking mechanism.

### 3.1. Constructing and Structuring the Search Space

We combine knowledge-driven priors with embedding-based organization to efficiently navigate the open-world activity space. The prior space estimates the plausibility of action-object pairs using structured commonsense knowledge, guiding search toward semantically valid activities. In parallel, the search space structuring reorders VLM embeddings to ensure local jumps during search remain meaningful. This structured search space allows ProbRes to make jumps that respect semantic locality while maintaining computational efficiency.

**Algorithm 1** Probabilistic Residual Search (ProbRes) for Open-World Activity Recognition

**Input**: Prior probabilities $P_{\text{prior}}(a)$, VLM likelihoods $P_{\text{likelihood}}(v \mid a)$, search space $\mathcal{S}$, exploration weight $\lambda$ and duration $T_{\text{explore}}$, max iterations $T$.
**Output**: Predicted activity $a^*$

**Initialize:** $t \leftarrow 0$, $a_0 \sim P_{\text{prior}}(a)$
**while** $t < T$ **do** ▷ *Iterative Search Process*
  **if** $t < T_{\text{explore}}$ **then** ▷ *Exploration Phase*
    $P_{\text{explore}}(a) = \frac{\lambda P_{\text{prior}}(a) + (1-\lambda)\frac{1}{|\mathcal{S}|}}{\sum_{a'} \lambda P_{\text{prior}}(a') + (1-\lambda)}$
    $a_{t+1} \sim P_{\text{explore}}(a)$ ▷ *Sample with exploration*
  **else** ▷ *Exploitation Phase*
    $P_{\text{guided}}(a) \propto P_{\text{prior}}(a) \cdot P_{\text{likelihood}}(v \mid a)$
    $a_{t+1} \sim P_{\text{guided}}(a)$ ▷ *Sample with exploitation*
  **end if**
  $t \leftarrow t + 1$
**end while**
$\mathcal{A}_{\text{refine}} = \text{topk}_{a \in \mathcal{S}} P_{\text{likelihood}}(v \mid a)$ ▷ *Top $k$ candidates for refinement*
**for** each $a \in \mathcal{A}_{\text{refine}}$ **do**
  $a \rightarrow (a_{\text{action}}, a_{\text{object}})$ ▷ *Decompose activity*
  $S_a = v^T \phi(a_{\text{action}})$ ▷ *Action Likelihood*
  $S_o = v^T \phi(a_{\text{object}})$ ▷ *Object Likelihood*
  $S_{\text{final}}(a) = P_{\text{likelihood}}(v \mid a) + \lambda_a S_a + \lambda_o S_o$
**end for**
**Return:** $a^* = \arg\max_{a \in \mathcal{A}_{\text{refine}}} S_{\text{final}}(a)$

#### 3.1.1. Constructing a prior space.

To guide search in open-world activity recognition, we construct a prior probability distribution over action-object pairs using external semantic knowledge. Specifically, we leverage ConceptNet [47, 48] to estimate the likelihood of an action-object pair occurring based on structured commonsense relationships. Given a finite set of actions $\mathcal{A}$ and objects $\mathcal{O}$, we define the prior probability of an activity $a = (a_{\text{action}}, a_{\text{object}})$ as $P_{\text{prior}}(a) \propto f(a_{\text{action}}, a_{\text{object}})$, where $f(a_{\text{action}}, a_{\text{object}})$ is a semantic score computed as the average weight of all shortest paths in ConceptNet between each object and action pair. While this is a good semantic similarity measure, it does not capture the compositional nature of affordance-based affinity between the concepts.

**Affinity Scoring.** We model ConceptNet as a directed graph $G$ with nodes as concepts (e.g., *wash*, *dish*) and weighted edges representing semantic relations. The score $f(a_{\text{action}}, a_{\text{object}})$ is derived from the shortest path between $a_{\text{action}}$ and $a_{\text{object}}$, computed as the sum of edge weights with an exponential decay factor $\lambda^i$ at step $i$. To refine the prior, we apply a relation-based adjustment that penalizes negative relations (e.g., `NotCapableOf`, `NotUsedFor`) and boosts positive ones (e.g., `UsedFor`, `CapableOf`). Specifically, we compute the final affinity score as

$$f(a_{\text{action}}, a_{\text{object}}) \leftarrow f(a_{\text{action}}, a_{\text{object}}) \cdot R(a_{\text{action}}, a_{\text{object}}) \quad (2)$$

where $R(a_{\text{action}}, a_{\text{object}})$ is a multiplicative weight derived from ConceptNet relations. The prior is then normalized across all candidate activities to form a probability distribution, serving as the exploration distribution in ProbRes to guide search toward semantically plausible activities before likelihood-based refinement. Empirically, in Section 5, we see that this prior space guides the search process much more efficiently than a random search.

#### 3.1.2. Search Space Structuring

The search is performed in the text embedding space of activities, where each candidate is represented by a Vision-Language Model (VLM) embedding $\phi(a)$. However, raw VLM embeddings do not always encode the fine-grained affordance relationships necessary for effective search. We construct an ordered search space to ensure that the search moves in semantically meaningful directions. First, we extract action-object phrase embeddings from the VLM to form an initial unordered set $\mathcal{S} = \{a_i = (a_{\text{action}}, a_{\text{object}})\}$. We define a similarity metric based on the Euclidean distance between embeddings, $d(a_i, a_j) = ||\phi(a_i) - \phi(a_j)||$, where $\phi(a)$ represents the embedding of activity $a$. The search space is structured by selecting an anchor point $a_{\text{ref}}$ and sorting activities based on their distance to $a_{\text{ref}}$. This ensures that similar activities remain proximate, enabling ProbRes to perform jumps that respect semantic locality.

### 3.2. Adaptive Search via Jump Diffusion

Probabilistic Residual Search (ProbRes) efficiently estimates the most probable activity from a large open-world search space without exhaustive evaluation. It iteratively refines predictions by optimizing for

$$a^* = \arg\max_{a \in \mathcal{S}} [P_{\text{search}}(a) + \lambda_a S_a + \lambda_o S_o] \quad (3)$$

where $P_{\text{search}}(a) = P_{\text{explore}}(a)$ for $t < T_{\text{explore}}$ (exploration phase, Eqn 4) and transitions to $P_{\text{search}}(a) = P_{\text{guided}}(a)$ for $t \geq T_{\text{explore}}$ (exploitation phase, Eqn. 5). $P_{\text{explore}}(a)$ leverages prior knowledge to guide broad search, while $P_{\text{guided}}(a)$ refines the search trajectory using likelihood estimates. $S_a$ and $S_o$ capture the fine-grained likelihood of the activity's action and object components. A local refinement step does a fine-grained search in high-likelihood areas to distill activity labels. We detail this process next.

#### 3.2.1. Sampling-based Search

A key challenge in search-based optimization is efficiently identifying high-likelihood regions in the large search space while avoiding excessive VLM consultations, which

can be expensive. We address this by balancing exploration (prior-guided sampling to discover diverse candidates) and exploitation (likelihood-guided refinement to focus on promising hypotheses). This two-phase process ensures the search remains computationally efficient while converging to high-likelihood solutions.

**Exploration as Prior-Guided Search.** In the initial phase, the search process favors broad exploration to efficiently locate high-likelihood regions in $\mathcal{S}$. Instead of randomly sampling from the entire space, ProbRes leverages prior knowledge to guide exploration, ensuring that early search iterations prioritize semantically plausible activities:

$$P_{\text{explore}}(a) = \frac{\lambda P_{\text{prior}}(a) + (1-\lambda)\frac{1}{|\mathcal{S}|}}{\sum_{a'} P_{\text{prior}}(a') + (1-\lambda)} \tag{4}$$

Here, $\lambda \in [0,1]$ controls the tradeoff between prior-driven and uniform random sampling. When $\lambda \approx 1$, the search is strongly guided by priors, favoring activities that align with commonsense affordances. When $\lambda \approx 0$, the search behaves like a random walk over $\mathcal{S}$. This mechanism allows the search to avoid implausible candidates.

**Exploitation as Likelihood-Guided Refinement**. Once high-likelihood regions begin to emerge, the search shifts focus to exploitation. The goal is to refine the search trajectory and concentrate computational resources on the most promising candidates. Instead of sampling from the entire prior space, the search becomes likelihood-driven $P_{\text{guided}}(a) \propto P_{\text{likelihood}}(v \mid a)$. Hypotheses are then selected based on their estimated likelihood given the observed video, prioritizing candidates that the VLM deems most probable. This ensures the search process refines the results rather than continuing broad exploration.

**Balancing Exploration and Exploitation.** The transition from exploration to exploitation is governed by integrating prior probabilities and likelihood estimates. Instead of relying on an explicit temperature parameter $\tau$, the search gradually shifts towards exploitation by weighting priors and likelihoods dynamically:

$$P_{\text{guided}}(a) = \frac{P_{\text{prior}}(a) P_{\text{likelihood}}(v \mid a)}{\sum_{a'} P_{\text{prior}}(a') P_{\text{likelihood}}(v \mid a')} \tag{5}$$

Early in the search, priors provide a structured means of exploration, ensuring that plausible activities are prioritized. As the search progresses and likelihood estimates become more informative, the search increasingly favors high-likelihood candidates while retaining prior knowledge for regularization. This gradual transition enables an efficient balance between discovering new hypotheses, refining the most probable ones, and reducing computational overhead.

#### 3.2.2. Localized Refinement in High-Likelihood Areas

While the exploitation phase prioritizes high-likelihood candidates, it remains probabilistic and may fail to leverage strong hypotheses fully. ProbRes performs localized refinement to improve robustness, a deterministic re-evaluation of the highest-confidence candidates. Instead of continuing broad sampling, search is restricted to a subset of top-ranked hypotheses ($|\mathcal{A}_{\text{refine}}| = K$) defined as $\mathcal{A}_{\text{refine}} = \arg\max_{a \in \mathcal{S}} P_{\text{likelihood}}(v \mid a)$. Unlike exploitation, which still samples from a distribution, this refinement stage deterministically focuses on fine-tuning a small set of high-confidence candidates. This is crucial because early likelihood estimates can be noisy, and a single-pass sampling strategy may not be sufficient to distinguish the most accurate activity labels. By revisiting the strongest hypotheses, ProbRes prevents premature convergence to suboptimal solutions. This localized refinement stabilizes the ranking of predictions, ensuring that minor variations in likelihood do not disrupt the final decision. Instead of additional broad search iterations, the final stage is spent refining existing strong candidates to enhance efficiency and robustness.

#### 3.2.3. Concept Decomposition and Re-Ranking

After obtaining a refined candidate set, ProbRes applies a final re-ranking step, decomposing each activity into its action and object components. Instead of treating activities as atomic units, this step evaluates their semantic consistency separately by computing individual similarity scores for the action and object. Concept decomposition mitigates errors from Vision-Language Models (VLMs) by separately evaluating the action and object components, reducing the impact of incorrect phrase-level embeddings. Using finer-grained representations can help compensate for VLM ambiguities and ensure that activities are ranked based on semantically consistent elements rather than potentially noisy joint embeddings. Given a video representation $v$ and VLM embeddings $\phi(a_{\text{action}})$ and $\phi(a_{\text{object}})$, we compute their respective alignment scores as $S_a = v^T \phi(a_{\text{action}})$ and $S_o = v^T \phi(a_{\text{object}})$. These scores adjust the likelihood-based ranking by incorporating finer-grained semantic relevance. The final score for re-ranking is defined as $S_{\text{final}} = P_{\text{likelihood}}(v \mid a) + \lambda_a S_a + \lambda_o S_o$, where $\lambda_a$ and $\lambda_o$ control the relative influence of the action and object alignment scores. The highest-ranked candidate from the refined set $\mathcal{A}_{\text{refine}}$ is then selected as $a^* = \arg\max_{a \in \mathcal{A}_{\text{refine}}} S_{\text{final}}$. This hierarchical re-ranking ensures that activities with semantically consistent components receive higher confidence while mitigating errors from overly generic or ambiguous predictions.

**Implementation Details.** We use EGOVLP [31] and LAVILA [61] as VLM backbones for likelihood estimation, guided by structured priors from ConceptNet. For L2 and L3 evaluations, search spaces were generated using Gemini 2.0 Flash [37] and refined to remove duplicates. The search process is regulated by the exploration-exploitation balance parameter ($\lambda$) and search iterations ($T$), tuned via grid search for efficiency and accuracy. We set $\lambda$=0.5 for prior-guided exploration before likelihood-driven exploita-

| Approach | Datasets | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | GTEA Gaze | | | | GTEA Gaze+ | | | | EK100 | | | | CharadesEgo | | | |
| | VLM Calls | Activity | Phrase | WUPS | VLM Calls | Activity | Phrase | WUPS | VLM Calls | Activity | Phrase | WUPS | VLM Calls | Activity | Phrase | WUPS |
| ALGO | N/A* | 15.06 | 1.80 | 46.89 | N/A* | 18.84 | 3.40 | 43.29 | N/A* | 8.49 | 1.72 | 38.53 | N/A* | 2.62 | 0.10 | 36.68 |
| KGL | N/A** | 6.58 | 0.30 | 35.83 | N/A** | 10.76 | 1.10 | 39.41 | N/A** | 3.07 | 0.94 | 36.85 | N/A** | 2.06 | 0.05 | 31.48 |
| ALGO+LAVILA | N/A* | 22.05 | 3.50 | 49.42 | N/A* | 28.87 | 8.30 | 53.38 | N/A* | 17.69 | 4.39 | 34.47 | N/A* | 2.94 | 0.21 | 37.21 |
| LAVILA-Decomp | 380 | 14.57 | 2.10 | 48.65 | 405 | 26.87 | 7.89 | 53.12 | 29100 | 17.21 | 4.28 | 43.37 | 1254 | 11.82 | 1.51 | 38.99 |
| EGOVLP-Decomp | 380 | 9.31 | 0.90 | 40.51 | 405 | 23.30 | 4.10 | 51.74 | 29100 | 15.14 | 2.05 | 39.94 | 1254 | 10.66 | 1.12 | 38.73 |
| LAVILA | 380 | 29.02 | 9.53 | 51.31 | 405 | 29.82 | 7.91 | 53.27 | 29100 | 18.81 | 4.43 | 43.52 | 1254 | 11.34 | 1.45 | 39.76 |
| EGOVLP | 380 | 17.07 | 1.80 | 46.77 | 405 | 27.00 | 3.95 | 51.48 | 29100 | 12.78 | 2.22 | 39.84 | 1254 | 10.55 | 1.21 | 38.43 |
| LAVILA + ProbRes (Ours) | 110 | **33.80** ↑ | **8.98** ↑ | **53.34** ↑ | 175 | **31.70** ↑ | **9.33** ↑ | **53.82** ↑ | 3000 | **18.89** ↑ | **4.61** ↑ | **43.55** ↑ | 300 | **13.71** ↑ | **2.12** ↑ | **40.91** ↑ |
| EGOVLP + ProbRes (Ours) | 110 | **19.12** ↑ | **1.97** ↑ | **49.25** ↑ | 178 | **29.84** ↑ | **6.32** ↑ | **53.98** ↑ | 3000 | **13.45** ↑ | **2.36** ↑ | **40.53** ↑ | 300 | **12.65** ↑ | **1.78** ↑ | **38.91** ↑ |

Table 1. **L1 evaluation results** on GTEA Gaze, GTEA Gaze+, EK100, and CharadesEgo. ProbRes consistently improves activity recognition across all datasets while significantly reducing VLM calls, demonstrating its efficiency in structured search. Green arrows indicate performance gains, while red arrows denote declines. *Frame-wise CLIP querying, **Frame-wise Faster R-CNN querying.

tion and cap iterations at $T$=3000 for smaller datasets and $T$=1000 for larger ones to control computational cost. Reranking weights $\lambda_a$ and $\lambda_o$ are optimized within $[0.3, 0.7]$ to balance action-object contributions. All experiments run on an NVIDIA RTX 3090 GPU, with an average inference time of 2 seconds per video.

## 4. Experimental Setup

We evaluate ProbRes on four egocentric activity recognition datasets - GTEA Gaze [29], GTEA Gaze+ [16], EPIC-Kitchens-100 (EK100) [10], and Charades-Ego [46] - covering structured kitchen tasks to unconstrained daily activities. GTEA Gaze and GTEA Gaze+ feature meal preparation with generic, often ambiguous action-object pairs, making disambiguation challenging due to frequent interaction overlap. EK100 provides a fine-grained, structured activity space with precise action-object interactions, serving as a benchmark for large-scale open-world inference. Charades-Ego, spanning daily activities, requires robust reasoning over diverse activity compositions, further testing generalization in open-world settings.

**Metrics.** To evaluate performance across different levels of openness, we adopt a suite of metrics tailored to open-world activity recognition. Traditional accuracy and mean Average Precision (mAP) are limited in open-world settings, assuming a closed, fully enumerated label space. Instead, we use WUPS (Wu-Palmer Similarity) [54] to quantify semantic similarity between predicted and ground-truth activity labels, capturing partial correctness even when exact matches are infeasible. WUPS is computed for objects, actions, and full activity labels. We also report exact phrase-level match accuracy to measure precise label retrieval.

**Baselines.** We compare ProbRes against both Vision-Language Models (VLMs) and neuro-symbolic approaches to comprehensively evaluate open-world activity recognition. EGOVLP [31] and LAVILA [61] are strong baselines, leveraging large-scale VLM training to match video embeddings with pre-defined activity labels. While effective in closed-set settings, their reliance on direct embedding similarity limits their adaptability in unconstrained search spaces. To assess the role of structured knowledge, we also evaluate against KGL [2] and ALGO [26], two neuro-symbolic models that integrate commonsense reasoning for activity inference. KGL utilizes ConceptNet [47, 48] relations to construct knowledge graphs for action-object reasoning, while ALGO refines activity representations through affordance-based priors.

## 5. Quantitative Results

**L1 evaluation** results are summarized in Table 1. ProbRes achieves higher accuracy with fewer VLM calls, highlighting the efficiency of a structured search over exhaustive querying. On GTEA Gaze, ProbRes reduces VLM queries from 380 to just 110, while on EK100, the reduction is even more pronounced—from $29,100$ to just $3,000$, while achieving higher WUPS and phrase-level accuracy across datasets. This efficiency-accuracy tradeoff is critical for real-world scalability, demonstrating that ProbRes can effectively navigate open-world spaces without brute-force enumeration. The performance gains hold even in larger, more structured datasets like EK100 and CharadesEgo, suggesting that ProbRes generalizes beyond small-scale benchmarks and is well-suited for real-world, large-scale, open-world activity recognition. Interestingly, neuro-symbolic approaches like KGL and ALGO struggle significantly, with KGL performing worse than even baseline VLMs (e.g., 35.83 WUPS on GTEA Gaze). This highlights a fundamental limitation of fixed knowledge graphs, which lack adaptability in dynamic, open-ended environments where the space of possible actions is fluid and context-dependent. Furthermore, ProbRes shows significant improvements at the phrase level, indicating that it

| Openness | Approach | # VLM Calls | GTEA Gaze | | | GTEA Gaze+ | | | EK100 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Object | Action | Activity | Object | Action | Activity | Object | Action | Activity |
| L2 | LAVILA | 37191 | 50.72 | 25.96 | 38.34 | 59.25 | **30.36** | 44.81 | 37.89 | **23.39** | 30.64 |
| | LAVILA + ProbRes (Ours) | 1500 | **53.49**↑ | **33.07**↑ | **43.28**↑ | **62.26**↑ | 28.03↓ | **45.15**↑ | **41.28**↑ | 22.56↓ | **31.92**↑ |
| | EGOVLP | 37191 | **56.77** | 18.34 | 37.56 | 60.77 | 22.27 | 41.52 | 40.87 | **21.32** | 31.10 |
| | EGOVLP + ProbRes (Ours) | 1500 | 56.27↓ | **19.98**↑ | **38.13**↑ | **62.89**↑ | **23.89**↑ | **43.39**↑ | **42.54**↑ | 21.07↓ | **31.81**↑ |
| L3 | LAVILA | 195714 | 59.85 | 32.27 | 46.06 | 60.43 | 28.73 | 44.58 | **41.60** | 20.51 | 31.06 |
| | LAVILA+ProbRes (ours) | 5000 | **59.89**↑ | **35.53**↑ | **47.71**↑ | **62.35**↑ | **29.04**↑ | **45.70**↑ | 41.20↓ | **23.96**↑ | **32.58**↑ |
| | EGOVLP | 195714 | **52.94** | 18.93 | 35.94 | 56.95 | 24.37 | 40.66 | **40.70** | 20.42 | 30.56 |
| | EGOVLP+ProbRes (Ours) | 5000 | 52.01↓ | **20.92**↑ | **36.47**↑ | **60.08**↑ | **24.75**↑ | **42.42**↑ | 39.91↓ | **24.13**↑ | **32.02**↑ |

Table 2. **L2 and L3 evaluation** results on GTEA Gaze, GTEA Gaze+, and EK100. ProbRes achieves significant accuracy gains while reducing VLM calls, demonstrating efficient search in unconstrained spaces. WUPS is reported for object, action, and activity levels.

predicts plausible activities and refines them into more semantically accurate action-object descriptions. These findings suggest that leveraging structured priors becomes even more critical for guiding inference as the search space complexity increases exponentially in L2 and L3 settings. The ability of ProbRes to reduce VLM reliance helps scale to unconstrained activity recognition tasks.

**L2 Evaluation** results are presented in Table 2. In this setting, the search space is constructed dynamically by querying an LLM for domain-specific (e.g., kitchen) actions and objects (see supplementary for details). This leads to a significantly expanded and fine-grained space, introducing high specificity (e.g., *julienne*, *emulsify*, *ladle*) and semantic overlap (e.g., *stir* vs. *whisk*), increasing the challenge of accurate retrieval. Despite this complexity, ProbRes demonstrates strong performance while drastically reducing VLM calls—from $37,191$ to just $1,500$, maintaining or improving WUPS across datasets. Notably, while ProbRes improves activity recognition scores overall, it particularly benefits action recognition, which is expected given the disambiguation challenges posed by the expanded action space. Interestingly, EGOVLP sees a small drop in object-level WUPS when paired with ProbRes, suggesting that its lower baseline object retrieval performance may limit downstream phrase and activity accuracy. These results reinforce our hypothesis that structured priors and guided search enable efficient inference in unconstrained spaces to balance specificity and generalization.

**L3 Evaluation** introduces a significantly more unconstrained setting where even the domain is unknown, necessitating the automatic construction of the search space. We use an LLM to generate a broad set of everyday actions and objects spanning multiple domains. (More details in the supplementary materials.) Unlike L2, which constrained search to kitchen-related activities, L3 contains highly generic activities such as "carry chair", "store container", or "hold book", alongside more specific ones like "boil pasta" or "slice onion". This results in a vastly expanded search space with increased ambiguity, making exhaustive enumeration impractical. Interestingly, in many cases, most models perform better in L3 than in L2. This is primarily due to the LLM-generated search space being more generic, allowing for better coverage of ground-truth activities. However, this also highlights a key insight: *the success of open-world recognition depends on the quality of the search space construction.* An uncurated or overly broad search space may introduce excessive noise, while an excessively restrictive one may fail to capture plausible activities. As shown in Table 2, ProbRes outperforms baselines despite the less semantically structured search space. The reliance on structured priors is even more pronounced here, as VLMs struggle with the broad combinatorial possibilities, and brute-force methods quickly become infeasible. Notably, the performance gap between ProbRes and naive VLM inference increases in L3, reinforcing the need for an intelligent search strategy in open-ended environments. These findings suggest that scalable open-world activity recognition requires methods capable of dynamically adapting the search space, making ProbRes a promising direction for real-world deployment.

**Ablation Studies.** We analyze the impact of key ProbRes components on open-world activity recognition using LAVILA as the VLM backbone on GTEA Gaze under L3 evaluation (Figure 3). Figure 3(a) examines search iterations. While WUPS scores steadily improve, exact match plateaus, highlighting diminishing returns from over-exploration. This tradeoff is more evident in larger datasets like EK100, where broader search improves semantic accuracy but reduces exact match likelihood. Figure 3(b) evaluates the exploration-exploitation tradeoff by varying $\lambda$. Pure likelihood-driven search ($\lambda = 0$) leads to premature convergence on high-confidence but incorrect activities, while excessive prior-driven exploration degrades performance. Optimal balance ($\lambda \approx 0.6$) yields the highest WUPS and exact

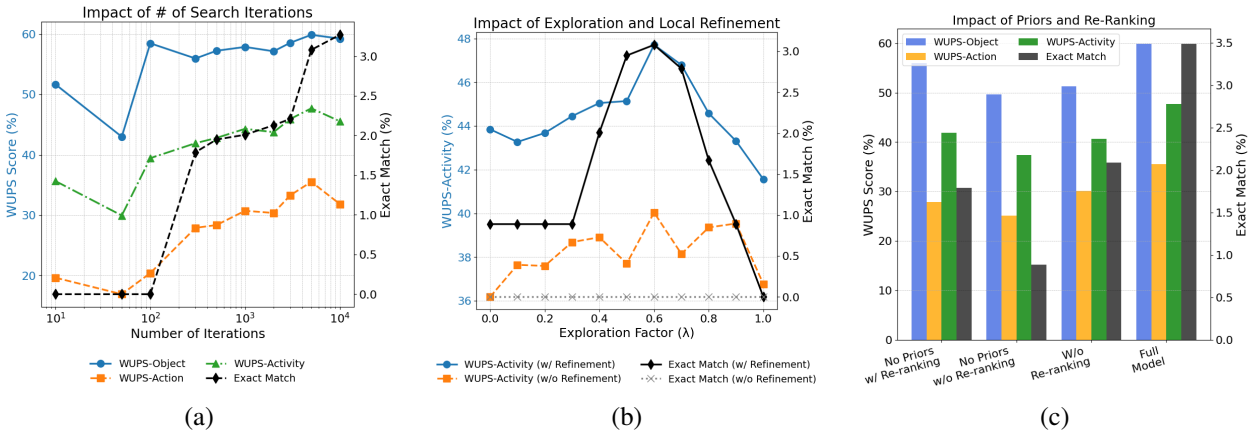


Figure 3. **Ablation Studies** that illustrate the impact of (a) the number of search iterations, (b) exploration-exploitation tradeoff and local refinement, and (c) knowledge-based priors and concept decomposition-based re-ranking, on the final performance.

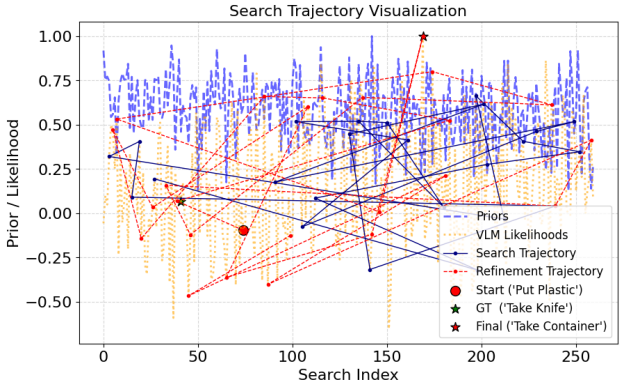


Figure 4. **Qualitative Visualization** of the search trajectory by ProbRes across different phases indicating exploration, exploitation, and the final refinement phase.

match, emphasizing the need for guided search. Figure 3(c) ablates knowledge priors and re-ranking. Removing ConceptNet priors significantly reduces performance, particularly in L3, where an unstructured search space increases uncertainty. Eliminating re-ranking lowers exact match, confirming its role in refining semantically similar but incorrect predictions.

**Qualitative Analysis.** Figure 4 visualizes the search trajectory of ProbRes in identifying the ground truth ("take fork") through exploration, exploitation, and refinement. In the exploration phase, the model samples diverse but loosely related actions (e.g., "put plastic" and "cut freezer"), demonstrating the influence of structured priors in steering broad yet relevant candidates. However, the unstructured nature of VLM embeddings results in abrupt transitions between unrelated activities (e.g., "pour burner" to "compress microwave"), disrupting semantic coherence. During exploitation, the search progressively narrows to more plausible candidates (e.g., "take fridge" and "put milk"), leveraging likelihood estimates to refine predictions. The refinement phase further structures the trajectory, filtering semantically similar but incorrect activities before converging near the ground truth. The inconsistencies in VLM embedding distances hinder an optimal search path, reinforcing the need for more structured representations [25], such as hyperbolic embeddings [35], to improve search efficiency.

## 6. Limitations, Discussion, and Future Work

While ProbRes significantly improves open-world egocentric activity recognition through structured search, challenges remain. Its reliance on Vision-Language Models (VLMs) inherits biases from pretraining data, occasionally leading to inefficient search trajectories. Structured priors help mitigate this, but the lack of semantic organization in text embeddings results in erratic jumps, suggesting that hyperbolic embeddings could enforce better locality. In higher openness levels (L2 and L3), LLM-generated search spaces enable scalable inference but require careful curation to avoid overly broad or noisy label distributions. Despite reducing VLM calls, further efficiency improvements—such as hierarchical search or adaptive refinement—are necessary for real-world deployment. ProbRes demonstrates that structured reasoning can replace brute-force enumeration in open-world settings. Integrating commonsense priors with probabilistic search lays the foundation for scalable and interpretable egocentric AI. Future work will explore additional cues, including human-object interactions and scene context, to enhance inference in unconstrained.

**Acknowledgement.** This research was supported in part by the US National Science Foundation grants IIS 2348689 and IIS 2348690. We thank Dr. Anuj Srivastava (FSU), Dr. Sudeep Sarkar (USF), and the anonymous reviewers for their thoughtful feedback to help present this work better.

## References


[1] Sathyanarayanan N. Aakur and Arunkumar Bagavathi. Unsupervised Gaze Prediction in Egocentric Videos by Energy-based Surprise Modeling, 2021. arXiv:2001.11580 [cs]. 3

[2] Sathyanarayanan N. Aakur, Sanjoy Kundu, and Nikhil Gunti. Knowledge guided learning: Open world egocentric action recognition with zero supervision. *Pattern Recognition Letters*, 156:38–45, 2022. Publisher: North-Holland. 1, 3, 6

[3] Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. HierVL: Learning Hierarchical Video-Language Embeddings. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23066–23078, Vancouver, BC, Canada, 2023. IEEE. 3

[4] Siddhant Bansal, Chetan Arora, and C. V. Jawahar. My view is the best view: Procedure learning from egocentric videos. In *Computer Vision – ECCV 2022*, pages 657–675, Cham, 2022. Springer Nature Switzerland. 2

[5] 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. Language Models are Few-Shot Learners, 2020. arXiv:2005.14165 [cs]. 3

[6] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of Artificial General Intelligence: Early experiments with GPT-4, 2023. arXiv:2303.12712 [cs]. 3

[7] Dibyadip Chatterjee, Fadime Sener, Shugao Ma, and Angela Yao. Opening the vocabulary of egocentric actions. *Advances in Neural Information Processing Systems*, 36: 33174–33187, 2023. 3

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations, 2020. arXiv:2002.05709 [cs]. 3

[9] Guangzhao Dai, Xiangbo Shu, Wenhao Wu, Rui Yan, and Jiachao Zhang. GPT4Ego: Unleashing the Potential of Pre-trained Models for Zero-Shot Egocentric Action Recognition, 2024. arXiv:2401.10039 [cs]. 3

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The EPIC-KITCHENS Dataset: Collection, Challenges and Baselines, 2020. arXiv:2005.00343 [cs]. 3, 6

[11] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, 2025. arXiv:2501.12948 [cs]. 3

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019. arXiv:1810.04805 [cs]. 3

[13] Na Dong, Yongqiang Zhang, Mingli Ding, and Gim Hee Lee. Open World DETR: Transformer based Open World Object Detection, 2022. arXiv:2212.02969 [cs]. 1, 3

[14] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to Prompt for Open-Vocabulary Object Detection with Vision-Language Model. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14064–14073, New Orleans, LA, USA, 2022. IEEE. 3

[15] Chenyou Fan. EgoVQA - An Egocentric Video Question Answering Benchmark Dataset. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4359–4366, Seoul, Korea (South), 2019. IEEE. 3

[16] Alireza Fathi, Yin Li, and James M. Rehg. Learning to Recognize Daily Actions Using Gaze. In *Computer Vision – ECCV 2012*, pages 314–327, Berlin, Heidelberg, 2012. Springer. 3, 6

[17] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie,

Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 Herd of Models, 2024. arXiv:2407.21783 [cs]. 3

[18] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4D: Around the World in 3,000 Hours of Egocentric Video, 2022. arXiv:2110.07058 [cs]. 3

[19] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 3

[20] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation, 2022. arXiv:2104.13921 [cs]. 1, 3

[21] Shangchen Han, Beibei Liu, Randi Cabezas, Christopher D Twigg, Peizhao Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Zheng Wang, et al. Megatrack: monochrome egocentric articulated hand-tracking for virtual reality. *ACM Transactions on Graphics (ToG)*, 39(4):87–1, 2020. 2

[22] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. In *Advances in Neural Information Processing Systems*, pages 3343–3360. Curran Associates, Inc., 2022. 2

[23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision, 2021. arXiv:2102.05918 [cs]. 3

[24] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*, pages 18661–18673. Curran Associates, Inc., 2020. 3

[25] Sanjoy Kundu and Sathyanarayanan N Aakur. Is-ggt: Iterative scene graph generation with generative transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6292–6301, 2023. 8

[26] Sanjoy Kundu, Shubham Trehan, and Sathyanarayanan N. Aakur. Discovering Novel Actions from Open World Egocentric Videos with Object-Grounded Visual Commonsense Reasoning, 2024. arXiv:2305.16602 [cs]. 1, 3, 6

[27] Shih-Po Lee, Zijia Lu, Zekun Zhang, Minh Hoai, and Ehsan Elhamifar. Error detection in egocentric procedural task videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18655–18666, 2024. 2

[28] Haoxin Li, Yijun Cai, and Wei-Shi Zheng. Deep Dual Relation Modeling for Egocentric Interaction Recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7924–7933, Long Beach, CA, USA, 2019. IEEE. 3

[29] Yin Li, Alireza Fathi, and James M. Rehg. Learning to Predict Gaze in Egocentric Video. In *2013 IEEE International Conference on Computer Vision*, pages 3216–3223, Sydney, Australia, 2013. IEEE. 3, 6

[30] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm, 2022. arXiv:2110.05208 [cs]. 3

[31] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, Chengfei Cai, Hongfa Wang, Dima Damen, Bernard Ghanem, Wei Liu, and Mike Zheng Shou. Egocentric Video-Language Pretraining, 2022. arXiv:2206.01670 [cs]. 3, 5, 6

[32] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019. arXiv:1907.11692 [cs]. 3

[33] Zheng Lu and Kristen Grauman. Story-Driven Summarization for Egocentric Video. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2714–2721, Portland, OR, USA, 2013. IEEE. 3

[34] Minghuang Ma, Haoqi Fan, and Kris M. Kitani. Going Deeper into First-Person Activity Recognition, 2016. arXiv:1605.03688 [cs]. 3

[35] Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. *arXiv preprint arXiv:2410.06912*, 2024. 8

[36] Kunyu Peng, Cheng Yin, Junwei Zheng, Ruiping Liu, David Schneider, Jiaming Zhang, Kailun Yang, M. Saquib Sarfraz, Rainer Stiefelhagen, and Alina Roitberg. Navigating Open Set Scenarios for Skeleton-based Action Recognition, 2023. arXiv:2312.06330 [cs]. 3

[37] Sundar Pichai, Demis Hassabis, and Koray Kavukcuoglu. Introducing gemini 2.0: Our new ai model for the agentic era, 2024. Accessed: [Insert Date Accessed Here]. 5

[38] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. EgoVLPv2: Egocentric Video-Language Pre-training with Fusion in the Backbone. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5262–5274, Paris, France, 2023. IEEE. 3

[39] Derek S. Prijatelj, Samuel Grieggs, Jin Huang, Dawei Du, Ameya Shringi, Christopher Funk, Adam Kaufman, Eric Robertson, and Walter J. Scheirer. Human Activity Recognition in an Open World. *Journal of Artificial Intelligence Research*, 81:935–971, 2024. arXiv:2212.12141 [cs]. 3

[40] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. 2018. 3

[41] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. 2019. 3

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, 2021. arXiv:2103.00020 [cs]. 3

[43] M. S. Ryoo, Brandon Rothrock, and Larry Matthies. Pooled motion features for first-person videos. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 896–904, Boston, MA, USA, 2015. IEEE. 3

[44] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022. 3

[45] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 1134–1141. IEEE, 2018. 2

[46] Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and Observer: Joint Modeling of First and Third-Person Videos, 2018. arXiv:1804.09627 [cs]. 3, 6

[47] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, 2017. 3, 4, 6

[48] Robyn Speer, Joshua Chin, and Catherine Havasi. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), 2017. Number: 1. 4, 6

[49] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. LSTA: Long Short-Term Attention for Egocentric Action Recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9946–9955, Long Beach, CA, USA, 2019. IEEE. 3

[50] Alessandro Suglia, Claudio Greco, Katie Baker, Jose L. Part, Ioannis Papaioannou, Arash Eshghi, Ioannis Konstas, and Oliver Lemon. AlanaVLM: A Multimodal Embodied AI Foundation Model for Egocentric Video Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11101–11122, Miami, Florida, USA, 2024. Association for Computational Linguistics. 3

[51] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, 2023. arXiv:2302.13971 [cs]. 3

[52] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer,

Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023. arXiv:2307.09288 [cs]. 3

[53] Xiaohan Wang, Linchao Zhu, Heng Wang, and Yi Yang. Interactive Prototype Learning for Egocentric Action Recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8148–8157, Montreal, QC, Canada, 2021. IEEE. 3

[54] Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138, 1994. 6

[55] Xing Xi, Yangyang Huang, Zhijie Zhong, and Ronghua Luo. UMB: Understanding Model Behavior for Open-World Object Detection. 2024. 3

[56] Boshen Xu, Ziheng Wang, Yang Du, Zhinan Song, Sipeng Zheng, and Qin Jin. Do Egocentric Video-Language Models Truly Understand Hand-Object Interactions?, 2025. arXiv:2405.17719 [cs]. 3

[57] Zihui Sherry Xue and Kristen Grauman. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. *Advances in Neural Information Processing Systems*, 36:53688–53710, 2023. 3

[58] Hanrong Ye, Haotian Zhang, Erik Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Bowen Zhang, Haoxuan You, Dan Xu, Zhe Gan, Jiasen Lu, and Yinfei Yang. MM-Ego: Towards Building Egocentric Multimodal LLMs, 2024. arXiv:2410.07177 [cs]. 3

[59] Heng Zhang, Daqing Liu, Qi Zheng, and Bing Su. Modeling video as stochastic processes for fine-grained video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2225–2234, 2023. 3

[60] Yun C. Zhang, Yin Li, and James M. Rehg. First-Person Action Decomposition and Zero-Shot Learning. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 121–129, 2017. 3

[61] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning Video Representations from Large Language Models, 2022. arXiv:2212.04501 [cs]. 3, 5, 6

[62] Yang Zhou, Bingbing Ni, Richang Hong, Xiaokang Yang, and Qi Tian. Cascaded Interactional Targeting Network for Egocentric Video Analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1904–1913, Las Vegas, NV, USA, 2016. IEEE. 3