Supplementary Material for: 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

1. Introduction

This supplementary document provides additional experimental details, extended qualitative analyses, and in-depth discussion on key factors affecting open-world egocentric activity recognition. We provide further insights into the biases introduced by ConceptNet and Vision-Language Models (VLMs), analyze the limitations of current text embeddings for structured search, and supplement our qualitative results with additional visualizations.

2. Additional Experimental Setup Details

Datasets. ProbRes is evaluated on four egocentric activity recognition datasets spanning varying degrees of openness: GTEA Gaze, GTEA Gaze+, EPIC-Kitchens-100 (EK100), and Charades-Ego. These datasets cover a spectrum from structured kitchen activities to highly unconstrained daily activities, enabling a robust assessment of open-world inference. GTEA Gaze and GTEA Gaze+ consist of meal preparation tasks with generic and ambiguous action-object compositions (e.g., *take bread*, *spread peanut butter*), leading to frequent overlaps in interactions. EK100 presents a more structured action-object space with precise interactions (e.g., *wash knife*, *cut chicken*), making it a suitable benchmark for large-scale inference. Charades-Ego, with activities like *watch television* and *tidy room*, poses a distinct challenge due to its diverse, multi-step activity sequences.

Implementation Details. We use EGOVLP and LAVILA as the VLM backbones for likelihood estimation, with ConceptNet providing structured priors for search guidance. In L2 and L3 settings, search spaces were generated using Gemini 2.0 Flash, with manual vetting to remove redundancies and inconsistencies. Search behavior is controlled by the balance parameter λ and maximum search iterations T, optimized via grid search on a validation set. We set $\lambda = 0.5$ for effective prior-guided exploration before likelihood-driven exploitation. The search iterations are capped at $T = 3000$ for smaller datasets and $T = 1000$ for larger ones to ensure computational efficiency. The reranking weights λ_a and λ_o are optimized within $[0.3, 0.7]$ to balance contributions from actions and objects. Experiments are conducted on an NVIDIA RTX 3090 GPU, with an average inference time of 2 seconds per video.

3. Biases in ConceptNet and VLMs

The structured priors from ConceptNet and the likelihood scores from VLMs play a crucial role in guiding the search process in ProbRes. However, these two components introduce biases that impact search efficiency and performance. Figure 1 highlights the disparities between the two probability distributions.

ConceptNet priors exhibit a relatively uniform probability distribution, assigning moderate probabilities to frequent human interactions. However, these priors are static, domain-agnostic, and fail to adapt dynamically to the video context. Conversely, VLM likelihoods display significant variance, often assigning high confidence to a subset of actions while heavily suppressing others. This behavior suggests that VLMs disproportionately favor activities seen frequently in pretraining corpora while underestimating less common actions, leading to inefficient search trajectories. The disparity between these two distributions can result in search inefficiencies—spending excessive iterations on high-prior but low-likelihood activities before correcting course.

These biases underscore the need for more adaptive prior weighting mechanisms that dynamically adjust to observed video content. Additionally, alternative embedding formulations, such as hyperbolic representations, could better capture hierarchical action-object relationships, reducing inconsistencies in likelihood scores.

4. Limitations of VLM Text Embeddings for Structured Search

A core challenge in open-world recognition is that VLM text embeddings lack semantic coherence, making struc-

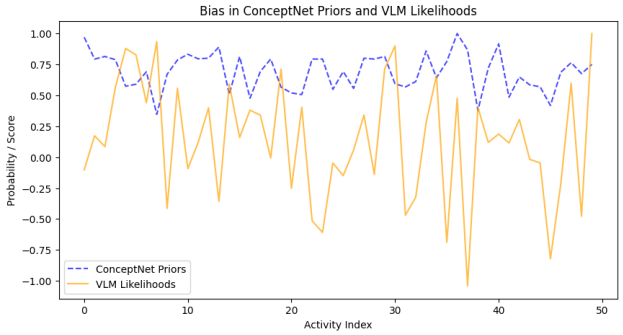

Figure 1. Normalized ConceptNet priors and VLM likelihoods for a sample video with ground truth "Take Fork." The discrepancy between the two distributions reveals biases—ConceptNet favors common, high-frequency activities, while VLM likelihoods fluctuate inconsistently due to unstructured semantic representations.

tured search challenging. We analyze this by projecting the embeddings of activity labels using t-SNE and UMAP.

Global Embedding Structure. Figures 2 (c) and 2 (d) visualize the embeddings of 1000 random activity labels. Ideally, semantically similar activities should form distinct clusters, but our results reveal highly dispersed representations. The silhouette score (0.06) and nearest-neighbor consistency (0.25) indicate weak clustering and poor local structure, reinforcing that the VLM embeddings do not effectively capture compositional relationships between activities.

Intra-Category Semantic Coherence. Figures 2 (a) and 2 (b) focus on activity labels involving the object "tomato." Ideally, related interactions such as "cut tomato" and "slice tomato" should be clustered, while unrelated ones like "throw tomato" should be further apart. However, the silhouette score (0.08) and nearest-neighbor consistency (0.07) reveal weak intra-category coherence. These findings suggest that text embeddings fail to structure actions and objects in a way that facilitates efficient search.

Implications for Open-World Search. These results emphasize two key insights. First, they reinforce the necessity of explicitly structuring the search space, as direct search over VLM embeddings leads to erratic refinements. Second, they highlight the limitations of likelihood-based inference in open-world settings, where unstructured embeddings result in inefficient search paths. Future research should explore hyperbolic embeddings or alternative representations that impose a more semantically meaningful organization of activities.

5. Search Space Construction for L2 and L3

A critical challenge in open-world activity recognition is defining an appropriate search space when no predefined action-object pairs exist. In L2 and L3 settings, where either only the domain (L2) or neither the domain nor valid activities (L3) are known, we constructed the search space using large-scale language models. Specifically, we used **Gemini 2.0 Flash** to generate candidate objects and actions, leveraging its broad commonsense knowledge while applying constraints to ensure a diverse and structured search space.

For **L2 (domain-aware open-world recognition)**, we prompted the model to generate objects typically found in a kitchen and actions that could be performed on those objects. The generated object list included common cooking utensils (e.g., "knife", "spoon"), food items (e.g., "tomato", "bread"), and appliances (e.g., "stove", "microwave"). Actions were generated by explicitly conditioning on these objects, leading to a diverse set of verbs covering manipulation (e.g., "cut", "stir", "slice"), interaction (e.g., "place", "remove"), and state changes (e.g., "heat", "cool"). A total of 231 objects and 159 actions were generated by Gemini and are used as our search space.

For **L3 (fully open-world recognition)**, we removed domain-specific constraints and instead prompted the model to generate a generic list of common objects across various environments. To maintain linguistic validity and ensure broad coverage, we explicitly constrained the output to **WordNet vocabulary terms** and instructed the model to remove redundant variations (e.g., "towel" implicitly covers "beach towel" and "kitchen towel"). Similarly, actions were generated by asking for distinct verbs applicable across objects, resulting in a highly diverse and unconstrained action space. A total of 786 objects and 247 actions were generated by Gemini and are used as our search space.

After generation, **manual vetting** was conducted to remove duplicates and highly ambiguous or irrelevant terms. This process ensured that the search space remained both expansive and representative of realistic open-world scenarios. All concepts are provided as part of the code zip file in the other supplementary materials.

Search space coverage with the groundtruth. Below the details of overlap between the constructed search space and each of the evaluation datasets is presented.

1. **GTEA Gaze** has an overlap of 17 objects and 6 actions with the **L2** search space and an overlap of 22 objects and 7 actions with the **L3** search space.
2. **GTEA Gaze+** has an overlap of 18 objects and 4 actions with the **L2** search space and an overlap of 20 objects and 4 actions with the **L3** search space.
3. **Epic Kitchens 100** has an overlap of 86 objects and 27 actions with the **L2** search space and an overlap of 123 objects and 40 actions with the **L3** search space.

5.1. Prompts Used for Search Space Construction

To ensure reproducibility, we document the exact prompts used to generate the search spaces:

L2 Search Space Generation

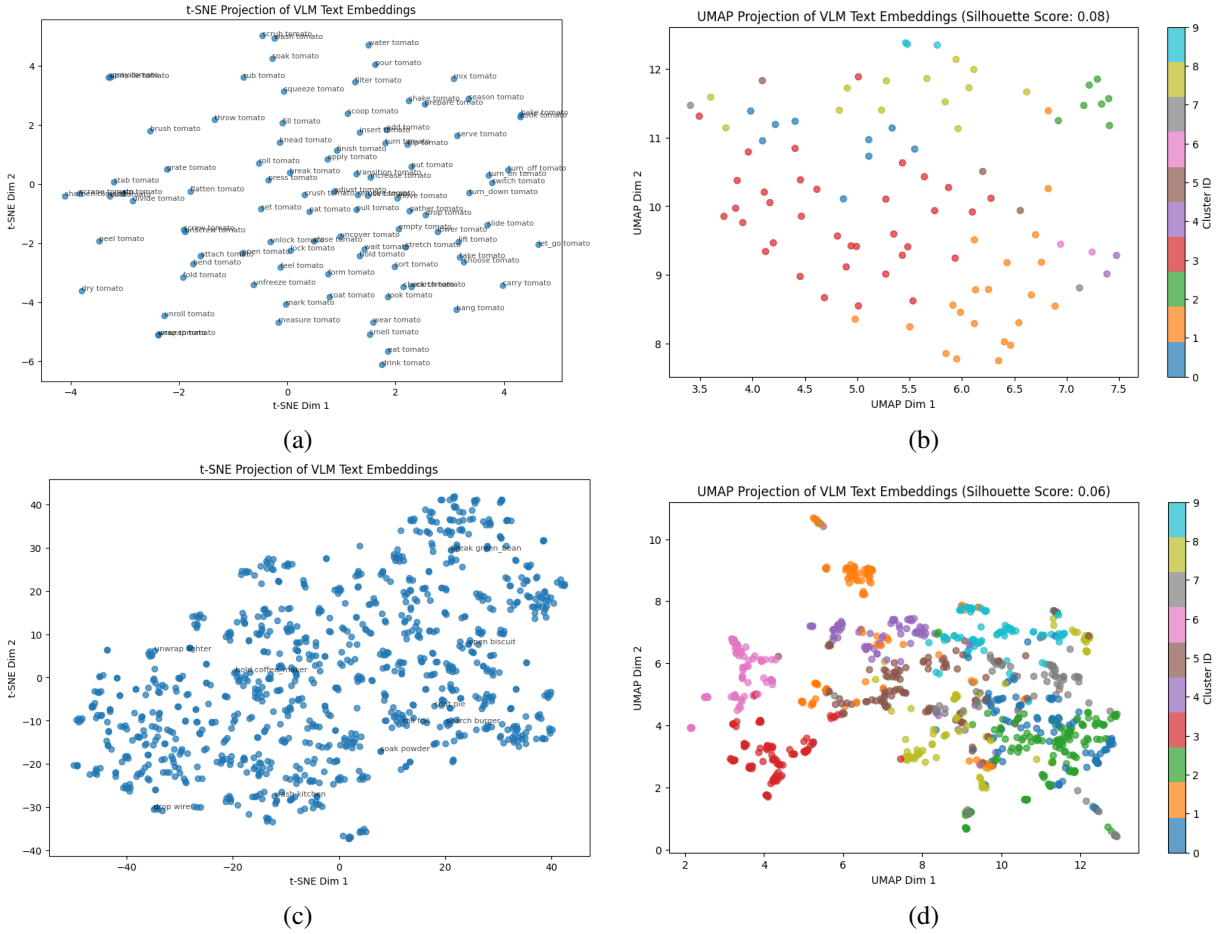

Figure 2. (a) t-SNE and (b) UMAP visualizations of activities involving the object "tomato." (c) t-SNE and (d) UMAP visualizations of 1000 random activities. Weak clustering and dispersion indicate that VLM embeddings lack structured semantic organization, negatively impacting search efficiency.

- **Objects:** *"Give me a list of 1000 common objects that can be found in a kitchen and used for cooking and other activities. The objects must be at most two words. Try to avoid compound words as much as possible. I do not need any categories. Give it as a text file with one item per line."*
- **Actions:** *"Can you generate a list of actions (verbs) that can be performed on the following objects? Be as distinct as possible and generate as many as possible. Ideally, it would be 200+."*

L3 Search Space Generation

- **Objects:** *"Give me a list of 1000 common objects. The objects must be at most two words. Try to avoid compound words as much as possible. I do not need any categories. Give it as a text file with one item per line. Try to keep each one in the WordNet vocabulary. Remove any redundant objects. For example, 'towel' covers kitchen towel, beach towel, etc."*
- **Actions:** *"Can you generate a list of actions (verbs) that can be performed on these objects? Be as distinct as possible and generate as many as possible. Ideally, it would be 300+. I want just the list of actions in the same format as the objects."*

5.2. Implications of Search Space Construction

The automated construction of search spaces ensures **scalability** while introducing **new challenges**:

- **Granularity and Specificity:** L2 search spaces tend to be highly specific due to kitchen-centric constraints, whereas L3 search spaces include broader, more ambiguous terms.
- **Coverage vs. Noise:** While larger search spaces allow for more generalization, they also introduce noise from loosely related concepts, requiring more robust search mechanisms.
- **Impact on Search Efficiency:** The structured nature of L2 search spaces makes search refinement more effective, whereas L3 search spaces lead to greater uncertainty, demanding stronger priors and more adaptive inference.

These findings reinforce the importance of **curating search spaces carefully** to balance generalization and precision in open-world activity recognition.

5.3. Taxonomy of Openness in Open-world Egocentric Activity Understanding

Level 1 refers to activity recognition in egocentric videos, where domain knowledge, atomic object, and verb concepts are known, but we search for a combination of these concepts. Level 2 extends the idea by assuming we have the domain knowledge, but the atomic action and object concepts and the combination of these concepts are unconstrained. In Level 3, the domain knowledge is also constrained along with the atomic concepts and combinations. Level 0 is the traditional closed set recognition of activity where the categories are predefined and fixed.

	Domain Knowledge	**Actions**	**Objects**	**Combinations**
L0	✓	✓	✓	✓
L1	✓	✓	✓	✗
L2	✓	✗	✗	✗
L3	✗	✗	✗	✗

Table 1. Different levels of openness in open-world learning

6. Conclusion

This supplementary analysis provides additional insights into the structural challenges of open-world activity recognition. We demonstrate that ProbRes effectively addresses search inefficiencies by integrating structured priors and likelihood-based reasoning. However, biases in ConceptNet and VLMs introduce inconsistencies, while unstructured text embeddings hinder effective search. Addressing these issues—through adaptive prior weighting and alternative embedding structures—presents a promising direction for improving open-world inference.