Implementation Details.

We use an S3D-G network pre-trained by Miech *et al.* [42,43] on Howto100M [43] as our visual feature extraction for visual-semantic action grounding. We use a CLIP model with the ViT-B/32 [20] as its backbone network. ConceptNet was used as our source of commonsense knowledge for neuro-symbolic reasoning, and ConceptNet Numberbatch [54] was used as the semantic representation for action grounding. The MCMC-based inference from KGL [4] was used as our reasoning mechanism. For experiments on Charades-Ego, where gaze information was unavailable, center bias [35] was used to approximate gaze locations. The mapping function, defined in Section 3.3, was a 1-layer feedforward network trained with the MSE loss for 100 epochs with a batch size of 256 and learning rate of 10^{-3}. Generalization errors on unseen actions were used to pick the best model. Two iterations of posterior-based action refinement were performed per video. Experiments were conducted on a desktop with a 32-core AMD ThreadRipper and an NVIDIA Titan RTX.

Why temporal smoothing?

Since we predict frame-level activity interpretations to account for gaze transitions, we first perform temporal smoothing to label the entire video clip before training the mapping function $\psi(g_i^a, f_V)$ to reduce noise in the learning process. For each frame in the video clip, we take the five most common actions predicted at the *activity* level (considering the top-10 predictions) and sum their energies to consolidate activity predictions and account for erroneous predictions. We then repeat the process for the entire clip, i.e., get the top-5 actions based on their frequency of occurrence at the frame level and consolidated energies across frames. These five actions provide targets for the mapping function $\psi(g_i^a, f_V)$, which is then trained with the MSE function. We use the top-5 action labels as targets to limit the effect of frequency bias. For example, some actions, such as *clean*, can possess a high affinity to many objects and hence be the most commonly predicted action for a frame. Hence, 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.

Why posterior-based refinement?

Since our predictions are made on a per-frame basis, it does not consider the overall temporal coherence and visual dynamics of the clip. Hence, there can be contradicting predictions for the actions done over time. Similarly, when setting the action priors to 1, we consider all actions equally plausible and do not restrict the action labels through grounding, as done for objects in Section 3.1. Hence, we iteratively update the action priors for the energy computation to re-rank the interpretations based on the clip-level visual dynamics. This prior could be updated to consider predictions from other models, such as EGO-VLP [36]

through prompting mechanisms similar to our neuro-symbolic object grounding. We iteratively refine the activity labels and update the visual-semantic action grounding modules simultaneously by alternating between posterior update and action grounding until the generalization error (i.e., the performance on unseen actions) saturates, which indicates overfitting. We empirically verify this in Section 4.2, where we observe the impact of this posterior update on activity recognition performance and the resulting loss of generalization.

Datasets

The GTEA Gaze dataset consists of 14 subjects performing activities composed of 10 verbs and 38 nouns across 17 videos. The gaze information is collected using Tobii eye-tracking glasses at 30 frames per second. The Gaze Plus dataset has 27 nouns and 15 verbs from 6 subjects performing 7 meal preparation activities across 37 videos. The gaze information is collected at 30 frames per second for both datasets using SMI eye-tracking glasses. Charades-Ego contains 7,860 videos containing 157 activities. Following prior work [36], we use the 785 egocentric clips in the test set for evaluation. EpicKitchens-100 is a large dataset comprising several hours of egocentric videos with 300 nouns (objects) and 97 verbs (actions). We use the validation set to evaluate our approach following prior works [65].

Visualizations of generated interpretations

Evidence-based grounding. Some examples of evidence-based grounding through ConceptNet are shown in Figure 3. As can be seen, each concept can have multiple evidence generators derived from ConceptNet using its ego-graph and limiting edges to those that express *compositional* properties such as `IsA, UsedFor, HasProperty` and `SynonymOf`. Using ego-graph helps preserve the contextual information within the semantic locality of the object to filter high-order noise induced by regular k-hop neighborhoods. The derived concepts provide additional context for verifying the presence of a given object in the video by querying CLIP as a noisy object oracle. Note that we do not visualize the edge label to avoid clutter. Each edge is qualified by a semantic assertion and is quantified by a value between -2 and 2 to express its strength. This provides a more explainable representation that enhances the final interpretation generated, as discussed next.

Semantically rich interpretations. The final interpretations generated by the approach are shown in Figure 4. It can be seen that the final interpretation is semantically rich and has concepts that are not directly in the scene but are compositionally relevant, either through affordance or object-level evidence. We visualize two different interpretations for affordance-based reasoning for the activity "cut fork" in (a) and (b), where it can be seen that the approach can generate graphs of varying structures. It also shows the impact of the noise in the knowledge graph that can introduce irrelevant concepts into the reasoning

process. We anticipate that having an additional reasoning step can reduce the impact of noise. Interestingly, we see that combinations of similar verbs and nouns such as "pour honey" (c) and "pour ketchup" (d) result in different ungrounded generators, indicating that the affordance of each concept is used for reasoning, resulting in semantically rich graphical interpretations. We anticipate that learning customized embeddings from these graphs can result in a better grounding of novel compositional concepts such as actions and activities.

Baselines. Since it is the most comparable, deep learning-only baseline, we choose ActionDecomposition [65] as our primary ZSL baseline for the GTEA datasets. It decouples verb and noun recognition and uses similar feature extractors for noun+verb prediction. However, they only report leave-one-class-out cross-validation, which assumes access to other examples from "seen" classes during training, which is not a fair comparison with our approach. We do not require any labels during training. Given the list of possible objects and actions in the videos, we ground the objects using CLIP, infer plausible actions using affordance-based reasoning, and predict the final activities driven purely by prior knowledge encoded in ConceptNet. Besides CLIP (trained only on objects), ALGO is not trained with any labels. Zero-shot models and "foundation models" are trained on considerable amounts of *video* data *with action/activity/object* labels and learn verb+noun associations from these labels. While "chain-of-thought prompting" or other decomposition approaches can possibly improve their generalization, no such models exist. We report the supervised/zero-shot/VLMs performance to illustrate that ALGO performs competitively despite not requiring large-scale video pretraining and labels (especially for actions and activities). While there is a gap between their performance, note that we do not have access to any data about actions (verbs) or activities, while VLMs and ZS models are trained exclusively on videos with such information. We learn to recognize actions with no labels and show that VLMs can benefit from ALGO (ALGO+LAVILA and ALGO+EGOVLP).

Action Recognition and Metrics. While activity recognition is the focus of the work, we would like to point out that the action (verb) recognition is also done in an open-world, inferred without labeled data. The performances reported for "Action" in Tables 1-3 represent ALGO's ability to infer the verb from contextual information. The top-5 performance of ALGO (obj/action/activity) on Gaze (40.75/34.89/37.82 vs. KGL's 32.39/10.73/18.78) and Gaze+ (42.77/53.88/48.33 vs. KGL's 24.64/37.99/27.53) as well as the exact match activity accuracy (Gaze - ALGO: 1.8, ALGO+LAVILA: 3.5, KGL: 0.3, EGOVLP: 0.9, LAVILA: 2.1, Gaze+ - ALGO: 3.4, ALGO+LAVILA: 8.3, KGL: 1.1, EGOVLP: 4.1, LAVILA: 7.9) show consistent improvements over the baselines across all metrics. While VLMs perform better than ALGO on exact match, which we attribute to their ability to identify the verb correctly, augmenting them with ALGO consistently improves their performance.

Search space and knowledge source. We use ConceptNet as the source of knowledge, over LLM or word embedding approaches, due to its ability to support probabilistic reasoning and an interpretable internal mechanism (see [4]).

While this does limit its vocabulary, it serves us well for reasoning over action-object affordance and affinity. We could replace ConceptNet priors with word embedding techniques such as GloVe or GPT 4, and the framework would still function without any modifications. Preliminary analysis with GloVe embedding on GTEA Gaze results in object/action/activity accuracy of 12.86/14.53/13.70, which outperforms KGL/KGL+CLIP/LAVILA/EGOVLP. More complex embedding could improve this performance at the cost of interpretability. Additionally, we show in Fig 2a that ConceptNet can be augmented with ChatGPT to move beyond its vocabulary. **The vocabulary of ConceptNet is not the same as the search space for CLIP/LAVILA/EGOVLP.** The search space is the space of prompts one provides to VLMs to select from and is usually predefined in zero-shot inference. Open-world inference requires building this search space for VLMs to infer labels. We propose (including prior-driven prompting) to construct such a search space without brute-force search over all combinations of verbs/nouns.

Qualitative Analysis Our analysis shows that the performance across the datasets varies and primarily stems from how the verbs and nouns are defined in the dataset. For example, Gaze and Gaze+ have activity labels with less ambiguous verb-noun combinations than EK100. For example, "clean", "put", and "take", common verbs in EK100, can apply to every single object and have very high affinity in ConceptNet, leading to a higher likelihood of prediction. This is one of the failure modes of ALGO and is one of our future research directions.

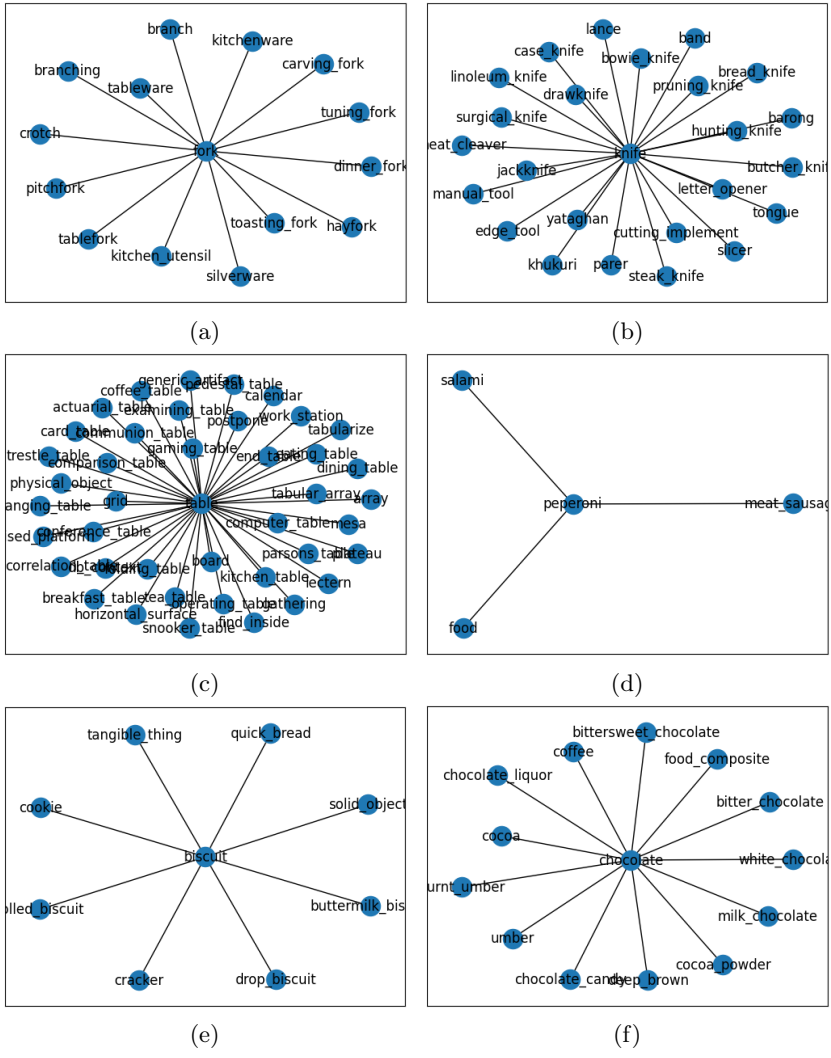

Fig. 3: Visualization of alternative concepts that were tested for grounding concepts in the video such as (a) fork, (b) knife, (c) table, (d) pepperoni, (e) biscuit, and (f) chocolate. These are automatically derived from ConceptNet and have semantic assertions quantifying how they are related.

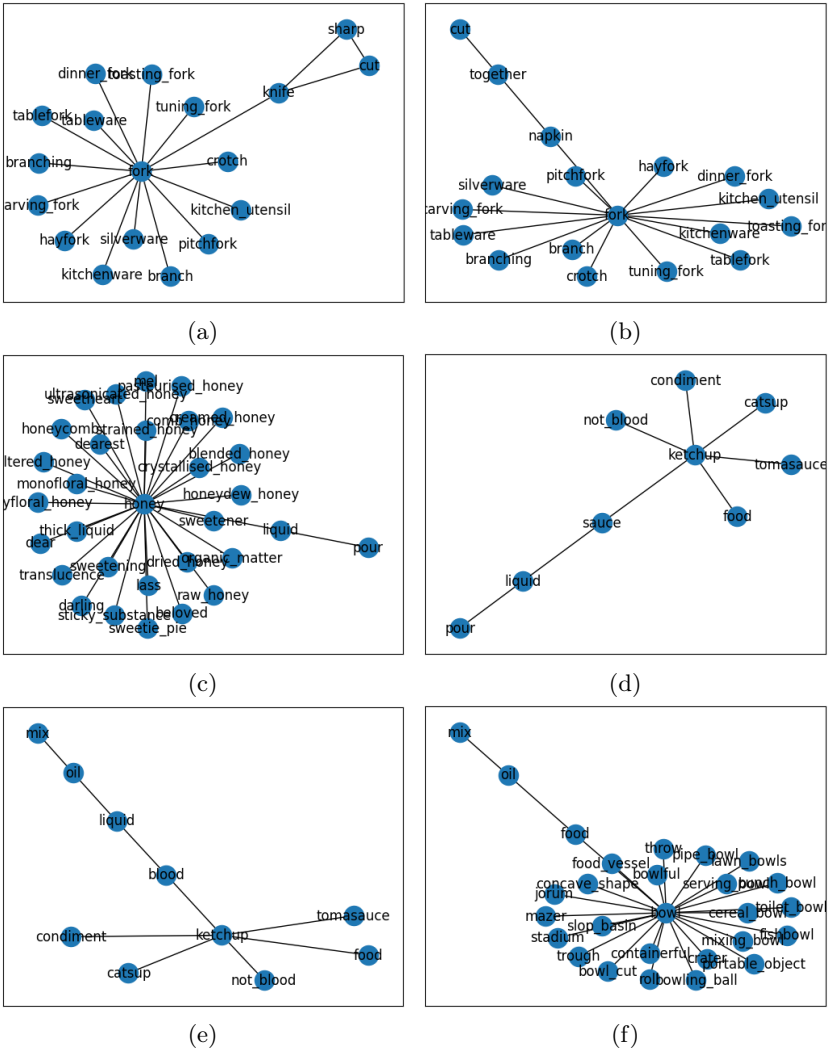

Fig. 4: Visualization of final interpretations for videos containing the activity (a) cut fork (top interpretation), (b) cut fork (second best interpretation), (c) pour honey, (d) pour ketchup, (e) mix ketchup, and (f) mix bowl. These are automatically derived from ConceptNet and have semantic assertions quantifying how they are related.