From Pixels to Graphs: using Scene and Knowledge Graphs for HD-EPIC VQA Challenge

Agnese Taluzzi[1]* Davide Gesualdi[1]* Riccardo Santambrogio[1] Chiara Plizzari[1]
Francesca Palermo[2] Simone Mentasti[1] Matteo Matteucci[1]

[1]Politecnico di Milano [2]EssilorLuxottica

Abstract

This report presents SceneNet and KnowledgeNet, our approaches developed for the HD-EPIC VQA Challenge 2025. SceneNet leverages scene graphs generated with a multi-modal large language model (MLLM) to capture fine-grained object interactions, spatial relationships, and temporally grounded events. In parallel, KnowledgeNet incorporates ConceptNet's external commonsense knowledge to introduce high-level semantic connections between entities, enabling reasoning beyond directly observable visual evidence. Each method demonstrates distinct strengths across the seven categories of the HD-EPIC benchmark, and their combination within our framework results in an overall accuracy of 44.21% on the challenge, highlighting its effectiveness for complex egocentric VQA tasks.

1. Introduction

Egocentric videos present unique challenges due to rapid camera motion from head and body movements and a limited, shifting field of view. These factors lead to frequent occlusions and partial observations, complicating reliable reasoning about interactions and temporal dynamics, especially when critical context lies outside the visible frame [6, 9]. Visual question answering (VQA) on egocentric video suffers from these difficulties. First, reasoning must be performed over long sequences of frames, resulting in large amounts of data and tokens that must be efficiently processed. Second, most state-of-the-art multi-modal large language models (MLLMs), such as Gemini [7], are pretrained on extensive datasets that primarily consist of third-person or non-egocentric content, limiting their ability to generalize to the unique characteristics of egocentric video [1, 14]. These present a significant obstacle for current models, which often rely on frame-level representations and lack effective mechanisms to incorporate structured semantic knowledge for deeper reasoning about object affordances, state changes, and causal dynamics.

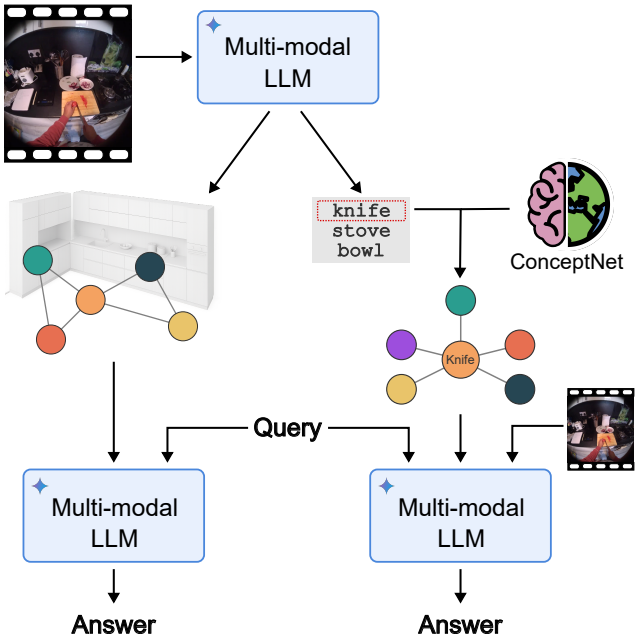

Figure 1. **Overview of our VQA method.** Given an input video, we use a multi-modal large language model (LLM) to construct two structured representations: (1) a scene graph capturing spatial relationships between objects (SceneNet), and (2) a knowledge graph for each object based on ConceptNet [23] (KnowledgeNet). These structured representations, together with a natural language question, are fed into the multi-modal LLM to generate an answer.

To address these challenges, we turn to neuro-symbolic AI, a composite framework that seeks to merge neural network-based methods with symbolic knowledge-based approaches [3]. It is founded on the premise of attaining the complementary benefits of both approaches, integrating the neural network capabilities of direct training from raw data and robustness against faults in the underlying data with the symbolic reasoning capabilities that enable them to reason about abstract concepts, extrapolate from limited data and generate explainable results [5, 10, 15, 26]. This makes

*Equal Contribution.

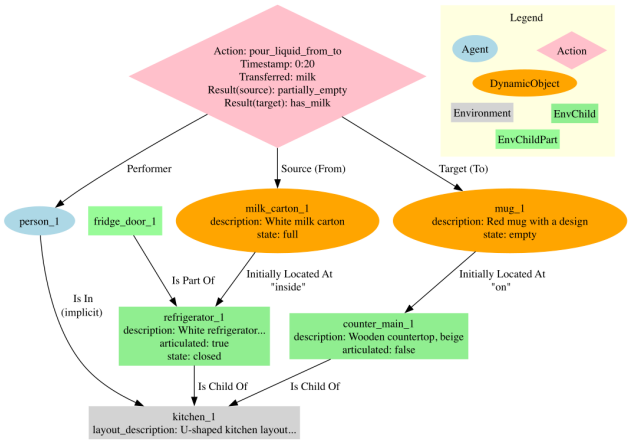

Figure 2. **SceneNet.** Example SceneNet graph illustrating an initial state with various entities (Agent, Environment, Dynamic Objects) and their relationships, followed by a "pour milk" action.

them especially suitable for challenging domains like egocentric video analysis, where semantic understanding, commonsense reasoning, and generalization to the egocentric domain are essential.

In this work, we propose a framework that addresses key challenges in egocentric VQA by incorporating two complementary neuro-symbolic abstractions: *scene graphs* and *commonsense knowledge graphs*, instantiated through dedicated modules we call **SceneNet** and **KnowledgeNet**, respectively. Scene graphs, originally introduced for static image understanding [12] and extended to dynamic video domains [17, 20, 25], encode objects, attributes, spatial relationships, and interactions within frames in a structured symbolic format, enabling reasoning over abstracted scene representations. These capabilities are leveraged by SceneNet to capture fine-grained spatial and relational information. Commonsense knowledge graphs such as *ConceptNet* [23] provide external semantic knowledge about typical object interactions, affordances, and causal events [4, 11, 13, 16, 21], which KnowledgeNet integrates to support inference beyond direct visual evidence. An overview of the approach is provided in Figure 1.

We detail the data processing pipeline, scene graph extraction methods, graph-based reasoning strategies, and provide experimental results across various configurations.

2. Our Approach

Our approach uses an MLLM conditioned on question, video (optional), and graph-based inputs: either SceneNet (visual-spatial relations) or KnowledgeNet (semantic associations from ConceptNet). While both graphs are rooted in MLLM-extracted objects from the video or question, they differ in structure and purpose: SceneNet is visually grounded, whereas KnowledgeNet expands from a root object using ConceptNet [22] to capture implicit, external knowledge semantically related to the object.

The two graph construction pipelines are illustrated in Section 2.1 and Section 2.2 respectively.

2.1. SceneNet

2.1.1. Graph Generation

SceneNet models video segments as structured *scene graphs* (objects, attributes, spatial/temporal relations).

We extract scene graphs from egocentric videos using prompted generation with an MLLM guiding the model to produce structured JSON outputs. According to this schema, a target scene graph $\mathcal{G}$ is to be structured as a tuple:

$$\mathcal{G} = (N, E_B, A) \tag{1}$$

where N is the set of nodes (entities), E_B is a set of binary edges representing direct, structural relationships, and A is a set of action relationships (hyperedges) representing interactions derived from the event timeline.

The conceptual design of the entities N and their structural relationships E_B draws inspiration from detailed 3D scene representations [2], adapted here for 2D video. The modeling of interactions A as hyperedges is inspired by approaches utilizing situation hyper-graphs to capture multi-entity events for video understanding [24].

The following paragraphs describe the expected constitution of N, E_B, and A as per the provided schema.

Core Entities (Nodes, N). Nodes $n \in N$ include: `Agent` (human actor), `Environment` (the kitchen), `EnvironmentChild` (fixed structural elements or large appliances, e.g., countertops, refrigerator), `EnvironmentChildPart` (movable sub-components like fridge doors), and `DynamicObjects` (movable or manipulated items, e.g. mugs, food items). Each has type-specific attributes (e.g. 'description', 'initial_state', 'articulated'–true if the object has movable parts).

Direct Relationships (Binary Edges, E_B). Binary edges $e_b \in E_B$ define hierarchical (`Contains/Is Child Of`, `Has Part/Is Part Of`) and spatial (`InitiallyLocatedAt`) links. `CreatedFrom` edges track when a dynamic object is originated from another one (e.g., when an object is sliced into smaller parts)

Interactions (Action Relationships / Hyperedges, A). Action hyperedges $a \in A$ represent multi-faceted, timestamped interactions, primarily defined by an `action` (e.g., "take_object", "pour_from_to", "stir"). Each action connects an `agent` (the performer) to various optional participant entities such as `source`, `target`, `location`, `tool`, or any newly created dynamic objects whose involvement can vary depending on the specific action. Key descriptive properties of the interaction, like the 'resulting_state_of_source' and 'resulting_state_of_target', are also captured within the hyperedge.

Figure 3. **KnowledgeNet.** Simplified knowledge graph rooted at the concept node *dishwasher*, prior to domain filtering.

Figure 2 provides a visual example of the scene graph representation used in SceneNet.

2.2. KnowledgeNet

To facilitate deeper reasoning about objects affordances, we construct symbolic *commonsense knowledge graphs* rooted in objects identified in video clips, using ConceptNet as our external knowledge source. ConceptNet provides concept nodes, each representing a word or short phrase in natural language, typically a common noun or verb in its undisambiguated form, linked through directed, labeled edges, also referred to as *assertions*. These nodes and assertions are used to expand object-centric graphs. We outline our graph generation pipeline below.

Object Recognition and Root Node Selection. Objects are first detected using zero-shot object recognition with an MLLM, prompted with a dedicated template. Let $\mathcal{O} = \{o_1, \ldots, o_m\}$ be the set of all detected object mentions in a scene. Each object o_i serves as the root node for a knowledge graph $\mathcal{G}_i$.

Knowledge Graph Definition. Each object-centric knowledge graph $\mathcal{G}_i$ is represented as a labeled, directed, and attributed graph defined as:

$$\mathcal{G}_i = (V_i, R, E_i), \tag{2}$$

where V_i is the set of concept nodes (e.g., `kitchen`, `cupboard`, ...), R is a predefined set of semantic relations (e.g., `used for`, `part of`, `has property`), and $E_i \subseteq V_i \times R \times V_i$ is the set of directed, labeled edges, referred to as assertions.

Each assertion $e \in E_i$ is a relational triple of the form:

$$e = (\mathrm{a}, \mathrm{r}, \mathrm{b}), \tag{3}$$

where $\mathrm{a}, \mathrm{b} \in V_i$, $\mathrm{a} \neq \mathrm{b}$, and $\mathrm{r} \in R$ denotes the semantic relation linking the two nodes (e.g., `cupboard is used for storing dishes`).

Each graph $\mathcal{G}i$ is a subgraph of the global ConceptNet knowledge graph. That is, $V_i \subseteq V_{\text{CN}}$, $R \subseteq R_{\text{CN}}$, and $E_i \subseteq E_{\text{CN}}$, where $V_{\text{CN}}, R_{\text{CN}}, E_{\text{CN}}$ represent the full set of nodes, relations, and assertions in ConceptNet, respectively.

Graph Construction. Graphs are constructed through a breadth-first expansion from the root node o_i up to depth $d = 3$. We initialize the node and edge sets as:

$$V_i^{(0)} = \{o_i\}, \quad E_i^{(0)} = \emptyset \tag{4}$$

At each step $k \in \{0, \ldots, d-1\}$, we expand the edge set using all ConceptNet assertions involving nodes from the current layer and valid relations:

$$\tilde{E}_i^{(k+1)} = E_i^{(k)} \cup \left\{ (v, r, v') \in E_{\text{CN}} \,\middle|\, v \in V_i^{(k)},\ r \in R \right\} \tag{5}$$

To prevent redundancy, symmetric relations (i.e., `similar to` and `synonym`) are ignored for $k > 0$.

Each assertion e in ConceptNet is associated with a confidence score, which we normalize to the interval $[0, 1]$ using a QuantileTransformer [18]:

$$w : E_{CN} \to [0, 1] \tag{6}$$

We retain only edges with $w(e) > 0.7$. When multiple edges connect the same node pair (v, v'), only the one with the highest weight is kept:

$$E_i^{(k+1)} = \bigcup_{(v,v') \in \mathcal{P}} \left\{ \arg \max_{e \in \{(v,r,v') \in \tilde{E}_i^{(k+1)}\}} w(e) \right\} \tag{7}$$

where

$$\mathcal{P} = \left\{ (v, v') \mid \exists r : (v, r, v') \in \tilde{E}_i^{(k+1)},\ w(e) > 0.7 \right\} \tag{8}$$

Finally, we update the node set:

$$V_i^{(k+1)} = V_i^{(k)} \cup \{v' \mid (v, r, v') \in E_i^{(k+1)}\} \tag{9}$$

Path Extraction. Graphs are serialized into semantic paths, each formally defined as a sequence of the form:

$$P = \{v_0, \mathrm{r}_1, v_1, \mathrm{r}_2, \ldots, \mathrm{r}_k, v_k\}, \tag{10}$$

starting from the root $\mathcal{V}_i^{(0)} = \{\mathrm{v}_0\} \subseteq \mathcal{O}$ and traversing through labeled relations. These are rendered into natural language templates (e.g., *"cupboard is used for storing dishes"*) for inclusion in the MLLM prompt.

Domain Filtering. To mitigate lexical ambiguity and out-of-context expansion, we filter the textual paths by measuring their semantic relevance to a kitchen context. Let $S = \{s_1, \ldots, s_n\}$ be a set of reference sentences describing kitchen-relevant affordances. Each path P is embedded as $\mathbf{p} = \text{Enc}(P)$ using a Sentence Transformer model [19][1]. Relevance is computed by:

$$\text{sim}(P) = \frac{1}{|S|} \sum_{s \in S} \cos(\mathbf{p}, \text{Enc}(s)) \tag{11}$$

[1]Specifically `NovaSearch/jasper_en_vision_language_v1` [27].

where $\cos(\cdot, \cdot)$ denotes the cosine similarity between the path embedding $\mathbf{p}$ and the encoded sentence $\text{Enc}(s)$. Paths are ranked by cosine similarity and up to 30 top-ranked paths are retained for each object (or all paths if fewer than 30 exist), ensuring contextual coherence.

2.3. Inference

We perform inference by prompting the MLLM with the question, the encoded graph (SceneNet or KnowledgeNet), and optionally video (see Section 3 for details), to select one multiple-choice answer.

3. Experiments

In this section, we explain the details for graph generation and the different input configurations used. We then report the results in Section 3.1.

Implementation Details. All experiments use Gemini 2.0 Flash [8] for both graph generation and inference. For SceneNet graph generation, videos were sampled at 1 FPS and 480p resolution, then split into segments of up to 400 seconds. A scene graph (see Section 2.1) was generated independently for each segment. To ensure temporal consistency, timestamps were adjusted globally across segments (e.g., "0:10" in a segment starting at "6:40" becomes "6:50"). Malformed JSONs (11.7%) were wrapped in a "raw_output" field. For gaze prediction, which involves anticipating future events, only the preceding 400s were used. For KnowledgeNet graph generation (see Section 2.2), Gemini was prompted to identify relevant kitchen objects from bounding boxes, timestamps, or visual context. The model returned a plain Python list of object names, guided by question-specific prompts. Gemini Flash 2.0 supports a maximum input duration of 45 minutes. To meet this, we applied a temporal divisor that adaptively accelerated videos (sampled at 1 FPS), with a minimum duration of one second. All videos were standardized to a 2400-second (40-minute) processing window to ensure compatibility.

Configuration Details. During inference, for any given question, we use one of these input configurations:

- **Video Only**: The model receives the natural language question and the raw video, without any structured graph representations. If the question input specified a particular clip or image, that corresponding video segment or frame was provided.
- **SceneNet (S-Net)**: The model receives the question along with the full set of (globally timestamp-normalized) scene graphs corresponding to all segments of the referenced video. Raw video was not used in this configuration.
- **KnowledgeNet (K-Net)**: The model receives as input the question, the video and the set of textualized semantic paths from K-Net.

For both SceneNet and KnowledgeNet configurations, if the input explicitly specified an image, or if the question included a single `<TIME>` tag, that corresponding frame was extracted from the video. If a bounding box tag (`<BBOX>`) was also included in the question, this bounding box was drawn onto the extracted image. This image was then provided to the MLLM alongside the respective graph representation (S-Net or K-Net) and the question to aid in visual disambiguation or grounding.

Category	Video Only	SceneNet	KnowledgeNet
3D Perception	29.38	**41.84** ↑12.46	34.29 ↑4.91
Action	**48.05**	31.97 ↓16.08	47.86 ↓0.19
Gaze	30.45	23.75 ↓ 6.70	**31.45** ↑1.00
Ingredient	45.17	39.67 ↓ 5.50	**45.83** ↑0.66
Nutrition	33.67	34.67 ↑ 1.00	**38.00** ↑4.33
Object Motion	17.53	**30.17** ↑12.64	19.36 ↑1.83
Recipe	40.75	**63.63** ↑22.88	49.50 ↑8.75
Overall	35.00	37.96 ↑2.96	**38.04** ↑3.04

Table 1. Per-category accuracy (%) across configurations. Green ↑ indicates improvement over the Video Only baseline; red ↓ indicates a decrease.

3.1. Results

Per-category results. Table 1 presents per-category results for the Video Only, SceneNet, and KnowledgeNet models. S-Net achieves a 2.9% improvement in overall performance w.r.t. Video Only, with 12% improvement in 3D perception and Object Motion categories and 22% improvement on Recipe. These results suggest that scene graphs help the model better capture spatio-temporal dynamics especially useful for modeling objects changing location through time. SceneNet performs well across most categories, though slightly underperforms in *action*, *gaze*, and *ingredient*, where scene graphs may offer limited value due to the need for fine-grained visual or compositional cues. K-Net improves performance over Video Only by 3% overall, with consistent gains across all categories except Action. Notable gains appear in Recipe and Nutrition categories, which benefit from external procedural knowledge beyond visual detail.

Micro-category Analysis. Figure 4 shows the performance of all configurations across all micro-categories within the broader category groups. The results highlight that different approaches excel in different contexts. In particular, some methods consistently outperform others across all micro-categories within specific macro-categories. For example, S-Net shows consistent improvements across all micro-categories of 3D Perception, while K-Net achieves similar consistent gains in the Recipe category.

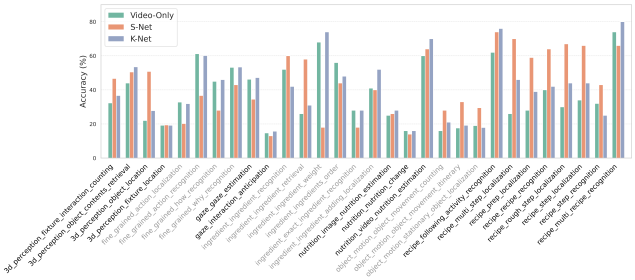

Figure 4. Accuracy (%) for each micro-category.

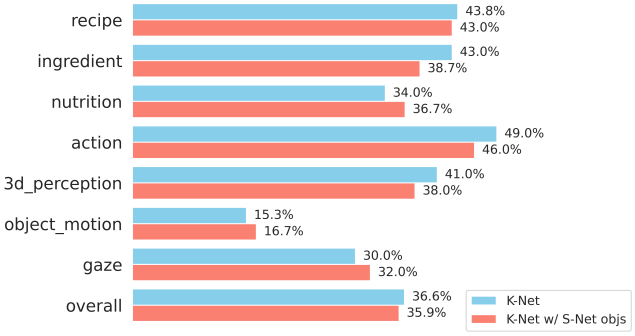

Figure 5. Per-category accuracy (%) for K-Net w/ video and K-Net w/ video + SceneNet objects.

Model	**S-Net**		**K-Net**		**S+K-Net**	
Video	✓	✗	✓	✗	✓	✗
3D Perception	34.50	**42.50**	**41.00**	31.00	**37.50**	36.50
Action	**40.50**	33.00	**49.00**	22.00	**43.50**	25.50
Gaze	27.00	**30.00**	**30.00**	23.00	**25.00**	21.00
Ingredient	**40.66**	38.67	**43.00**	24.00	**41.00**	22.67
Nutrition	30.66	**36.00**	**34.00**	28.67	**34.67**	30.67
Object Motion	26.00	**29.33**	15.33	**31.33**	17.33	**31.33**
Recipe	49.75	**62.75**	**43.75**	30.00	**46.75**	33.25
Overall	35.58	**38.89**	**36.58**	27.14	**35.11**	28.70

Table 2. Per-category accuracy (%) for S-Net, K-Net, and their combination (S+K-Net).

3.2. Ablations

We conducted ablation studies using 50 samples per micro-category. Detailed results are presented in Table 2.

S-Net (w/ and w/o video). We first evaluated the impact of adding raw video to S-Net. Overall, S-Net without video outperformed the combined S-Net with video setup, suggesting that S-Net's structured information is often sufficient, while raw video can introduce noise or complicate fusion. However, for inherently visual categories like Action and Ingredient, S-Net with video showed clear gains, consistent with the strengths of the video-only baseline. Despite these cases, the general trend favored S-Net only.

K-Net (w/ and w/o video). We then compared K-Net with and without raw video. The version without video generally underperformed its video-augmented counterpart, highlighting the importance of grounding symbolic knowledge in visual context to avoid reasoning using just commonsense knowledge not grounded in the video.

S+K-Net. This joint setup generally underperformed compared to using either source alone, likely due to the inclusion of excessive or irrelevant information that hinders reasoning. It also reflects the individual weaknesses of its components, i.e., SceneNet's sensitivity to added video and KnowledgeNet's reliance on visual grounding. These results highlight the challenges of integrating heterogeneous knowledge in VQA. We hypothesize that a more selective, question-aware fusion strategy could improve alignment with task-specific needs and enhance performance.

K-Net (from S-Net objects). We experimented with generating K-Net graphs using objects derived from S-Net entities, aiming for tighter integration through contextually grounded ConceptNet graphs (Figure 5). However, this setup underperformed compared to more targeted KnowledgeNet approaches, likely due to the added noise from including visually grounded but question-irrelevant entities.

3.3. Submission Strategy

For the final challenge submission, corresponding to team name *DeepFrames*, we selected the best-performing method for each micro-category. This ensemble strategy allowed each question to benefit from the input modality best suited to its specific reasoning needs.

4. Conclusion

In this report, we present our submission to the HD-EPIC VQA challenge. We propose extracting graph-based structured representations from video using two modules: SceneNet and KnowledgeNet. Each offers complementary strengths for egocentric video question answering, and we alternate between them based on the question category. A promising direction for future work is to integrate both modules into a unified approach, leveraging their strengths jointly rather than independently.

Acknowledgements. This work was conducted at the Smart Eyewear Lab, a joint research center between Essilor-Luxottica and Politecnico di Milano.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Paul Luc, Antoine Miech, Ian Barr, Yana Hasson, Mohammad Azar, Matthew Botvinick, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022. 1

[2] Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 3d scene

graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5664–5673, 2019. 2

[3] Luís C. Lamb Artur d'Avila Garcez. Neurosymbolic ai: the 3rd wave. *Artificial Intelligence Review*, 56(11):12387–12406, 2023. 1

[4] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, 2019. 2

[5] Minkyu Choi, Harsh Goel, Mohammad Omama, Yunhao Yang, Sahil Shah, and Sandeep Chinchali. *Towards Neuro-Symbolic Video Understanding*, page 220–236. Springer Nature Switzerland, 2024. 1

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. In *International Journal of Computer Vision*, pages 1255–1274, 2021. 1

[7] Yixuan Du, Jiechao Feng, Tianrui Tang, Jiarui Zhang, Jiarui Chen, Yanwei Wang, Zicheng Zhang, Fei Wu, Yutong Xu, Jie Hu, et al. Gemini: Integrating large language models and large vision models for general-purpose multimodal ai. *arXiv preprint arXiv:2303.17580*, 2023. 1

[8] Google. Gemini 2.0 flash. 4

[9] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1

[10] Drew Hudson and Christopher Manning. Learning by asking questions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 11–20, 2019. 1

[11] Gabriel Ilharco, Xiang Lisa Li, Trevor Darrell, Luke Zettlemoyer, Jean-Baptiste Alayrac, et al. Patching open-vocabulary vision models with commonsense. In *CVPR*, 2022. 2

[12] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *CVPR*, 2015. 2

[13] Sanjoy Kundu, Shubham Trehan, and Sathyanarayanan N. Aakur. Algo: Object-grounded visual commonsense reasoning for open-world egocentric action recognition, 2024. 2

[14] Junnan Li, Dongxu Li, Caiming Xiong, and Steven CH Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 1

[15] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. Neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8681–8690, 2019. 1

[16] Bahram Mohammadi, Yicong Hong, Yuankai Qi, Qi Wu, Shirui Pan, and Javen Qinfeng Shi. Augmented commonsense knowledge for remote object grounding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38 (5):4269–4277, 2024. 2

[17] Rohith Peddi, Saurabh, Ayush Abhay Shrivastava, Parag Singla, and Vibhav Gogate. Towards unbiased and robust spatio-temporal scene graph generation and anticipation, 2025. 2

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 3

[19] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019. 3

[20] Ivan Rodin, Antonino Furnari, Kyle Min, Subarna Tripathi, and Giovanni Maria Farinella. Action scene graphs for long-form understanding of egocentric videos, 2023. 2

[21] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3027–3035, 2019. 2

[22] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. *AAAI*, 2017. 2

[23] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge, 2018. 1, 2

[24] Aisha Urooj, Hilde Kuehne, Bo Wu, Kim Chheu, Walid Bousselham, Chuang Gan, Niels Lobo, and Mubarak Shah. Learning situation hyper-graphs for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14879–14889, 2023. 2

[25] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5419, 2017. 2

[26] Kexin Yi, Jiajun Wu, Amar Park, Ramakrishna Vedantam, Mateusz Malinowski, Dhruv Batra, Devi Parikh, and Damien Teney. Neuro-symbolic visual reasoning: Disentangling "visual" from "reasoning". In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4940–4948, 2018. 1

[27] Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. Jasper and stella: distillation of sota embedding models, 2025. 3