Applying LLM to question/answer problem based on sequence of frames

Proposed Enhancements & Recent Advances

Semenov Vadim, MIPT, Cognitive Modeling Lab, March 2025

Core research question:

How can we leverage LLMs to enhance question answering over sequences of frames by incorporating temporal and spatial context for robust place understanding in dynamic or large-scale environments?

My personal task:

1) Filter key frames.
2) Construct relationships between relevant nodes for the question using a multi-modal large language model (MLLM).
3) Predict the answer to the question using the graph with a large language model (LLM).

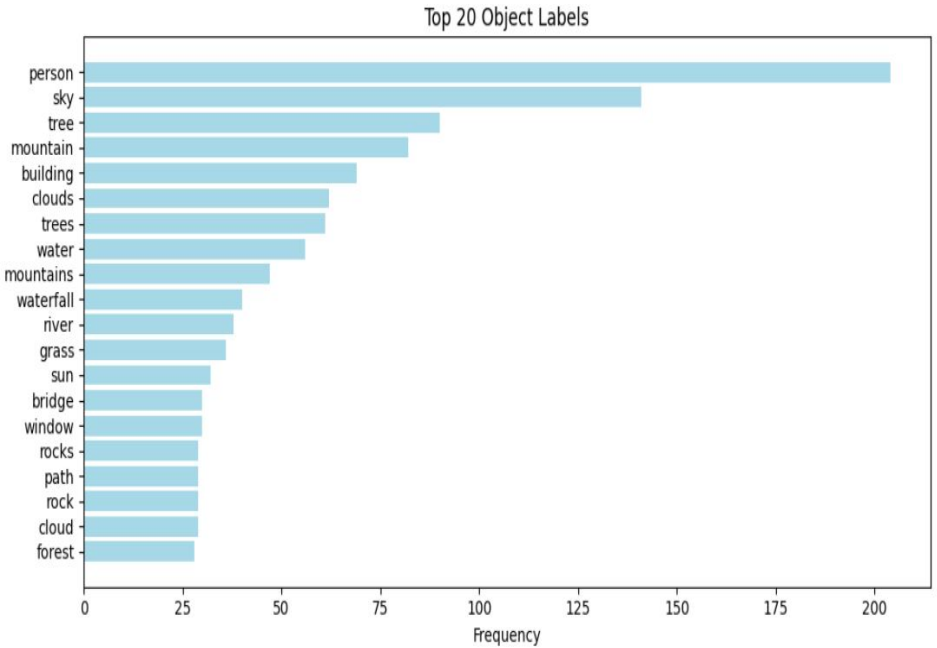
Planned results:

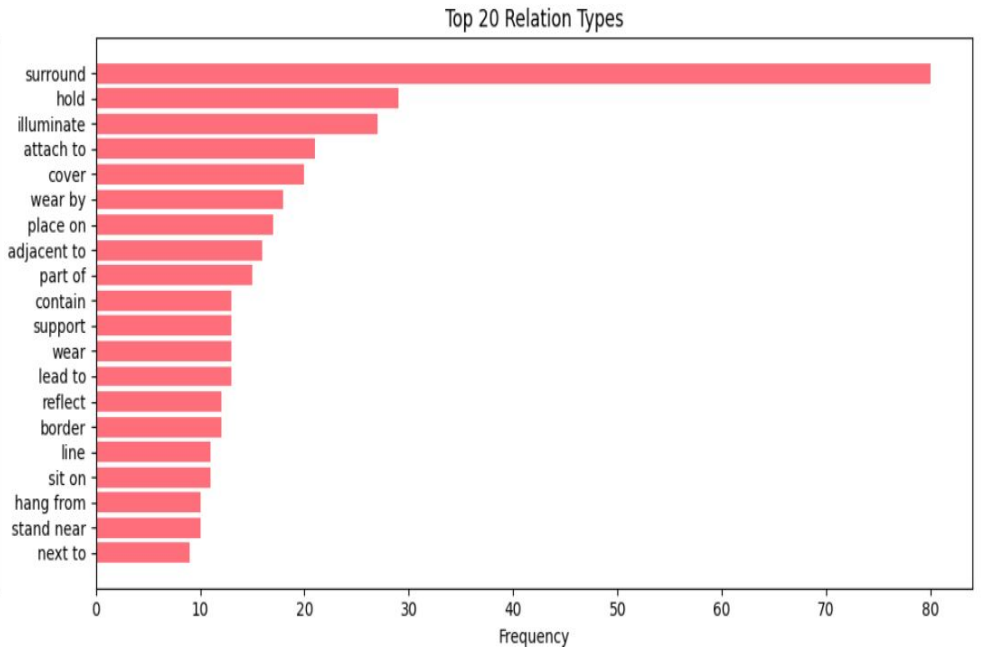
- **Paper Submission**: Aim to write and submit a scientific paper by the end of the semester
- **Technical Innovations**: Enhance the visual scene graph construction pipeline using MLLMs to extract semantic relations between key entities. Explore adding spatial/temporal edges to better capture context, and consider handling dynamic objects or incorporating richer semantic categories.
- **Evaluation:** Test the approach on benchmark datasets (GQA, Visual Genome), and evaluate using standard metrics such as EM, BLEU ROUGE METEOR BERTScore

Current status:

- **(done) Literature Gathered:** created a table summarizing article title, link, publication date, authors' affiliations, citations, code availability, GitHub stars, key differences & possible weaknesses, datasets, metrics, etc.
- **(done) Initial Summaries:** Ferret, Shikra, Kosmos-2.5, Osprey, Sphinx, etc..
- **(done) Datasets Gathered:** GQA(Graph Question Answering), 3DSRBanch, GRIT, Visual Genome
- **(new) Dataset Preparation:** LAION-SG (100 images, 400 questions) with images, graphs, and added bounding boxes, human verification.
- **(new) Model Evaluation Script:** works: Shikra, Sphinx, Qwen2.5-VL; issues: Ferret, Kosmos-2.5, Osprey

LAION-SG dataset





Evaluation metrics

model	GPU RAM (GB)	speed (s/it)	parallel	**Exact Match**	**BLEU**	**ROUGE-L**	**METEOR**	**BERTScore**
Ferret								
Shikra	14.5	0.49	yes	0.0000	0.0	0.0167	0.0109	0.8855
Kosmos-2	7.4	0.17	yes					
SPHINX-Tiny	9.8	0.22	no	0.0025	0.0	0.0242	0.0131	0.8986
Osprey								
Qwen2-VL	16.3	01.04	yes	0.0000	0.0	0.0204	0.0125	0.8910

Q/A

References:

- MLLM Review