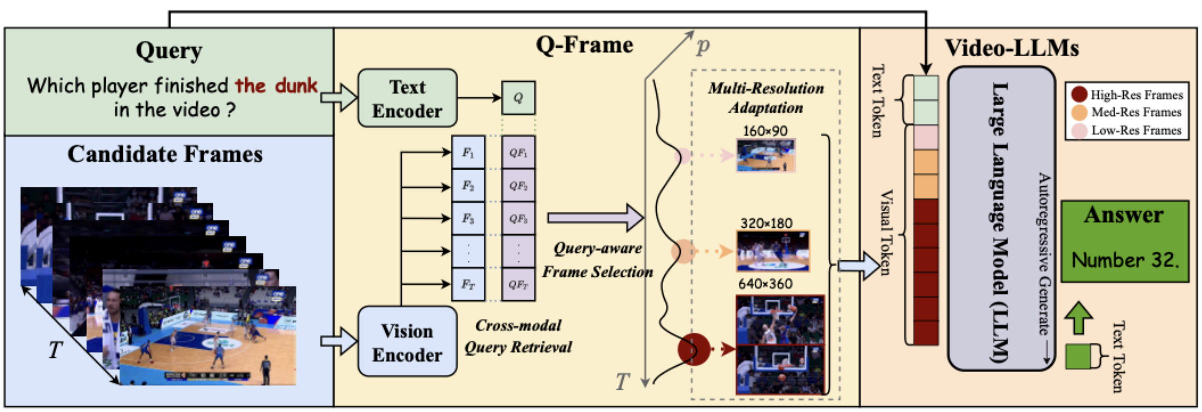
THESIS

Title: Probabilistic Query-Aware Frame Sampling for Efficient Video Understanding with Multimodal LLMs

Description:

Recent advances in Multimodal Large Language Models (MLLMs), such as GPT-5, Gemini-2.5, and VideoLLaMA3, have shown impressive capabilities in understanding vision contents (such as images, videos) through natural language queries. However, processing full-length videos remains computationally challenging due to the quadratic growth of visual tokens with video length. To address this, current systems often employ sampling techniques to limit the number of frames fed into the model. Common approaches include uniform sampling, adaptive temporal sampling i.e. selecting frames at regular intervals or based on video duration, hierarchical chunking i.e. breaking a video into chunks and retrieving selected chunks for a query [1-3].

Although effective, these approaches ignore query relevance, treating all frames as equally informative regardless of their semantic importance to the specific user query. For example, in answering “When does the car start moving?”, only a small subset of frames around the relevant event truly matter. This inefficiency motivates a query-aware sampling strategy that focuses on frames most likely to contribute to accurate answers.



An example workflow of a recent work in frame selection [4].

This thesis aims to develop a probabilistic frame selection mechanism that dynamically selects the most relevant frames in a video given a textual query. Specifically, the problem can be splitted into two tasks: (1) model frame relevance as a probability distribution conditioned on the query, (2) sample frames according to this distribution under a limited token budget.

Research Objectives:

- Develop a probabilistic model for query-conditioned frame relevance estimation.
- Design and implement a query-aware frame sampler using Bayesian or importance-sampling principles.
- Evaluate the proposed sampler against uniform and adaptive baselines on public video QA benchmarks.
- Analyze trade-offs between accuracy, diversity, and computational efficiency.

References:

[1] Yi Wang et al. InternVideo2.5: Empowering Video MLLMs with Long and Rich Context Modeling. 2025. arXiv: 2501.12386 [cs.CV]. url: https://arxiv.org/abs/2501.12386
[2] Xinyue Hao et al. "Principles of Visual Tokens for Efficient Video Understanding". In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2025, pp. 21254–21264
[3] Xi Tang et al. "Adaptive keyframe sampling for long video understanding". In: Proceedings of the Computer Vision and Pattern Recognition Conference. 2025, pp. 29118–29128.
[4] Zhang, Shaojie, et al. "Q-Frame: Query-aware Frame Selection and Multi-Resolution Adaptation for Video-LLMs." arXiv preprint arXiv:2506.22139 (2025).

Prerequisites:

- Strong programming skills in Python and experience with PyTorch, Transformers, or similar frameworks.
- Solid background in probability and statistics, including Bayesian inference and probabilistic modeling.
- Familiarity with computer vision and machine learning concepts (transformers, attention mechanisms).
- Courses such as Advanced Probabilistic Machine Learning, Deep Learning and LLMs.

Contact persons:

- Tianru Zhang, tianru.zhang@it.uu.se
- Ekta Vats (webpage), ekta.vats@it.uu.se

(Please include your CV, academic transcripts and a motivation letter)

Starting date: Jan 2026, can be discussed
No. of students: 1-2