УДК 004.8

# Применение мультимодальных языковых моделей к задаче визуального вопросно-ответного анализа на видеоданных

В. Р. Семенов[1,2]

[1]Московский физико-технический институт (национальный исследовательский университет)

[2]Лаборатория интеллектуального транспорта НКБ ВС

Задача визуального вопросно-ответного анализа (VideoQA) заключается в генерации ответа на текстовый вопрос, заданный пользователем, на основе визуального содержания видеоролика. В рамках данной работы рассматривается подход, в котором мультимодальные языковые модели (MLLM), применяются для анализа и интерпретации семантического содержания ключевых сцен видеоролика.

Предлагаемый подход состоит из трех основных этапов:

1) Извлечение ключевых кадров: используется эмбеддинг-сопоставление между кадрами и вопросом на основе CLIP для отбора наиболее релевантных сцен. Это позволяет значительно сократить избыточность визуальных данных и сфокусировать анализ на содержательно значимых кадрах.
2) Построение сцены через пространственно-семантический граф: с использованием модели сегментации SAM2 [5] выполняется детекция объектов и консистентная идентификация между кадрами. Далее, с помощью MLLM анализируются действия и пространственные отношения между объектами для построения графа сцены. Для повышения интерпретируемости исследуются различные способы подачи визуальной информации: только маски, только текстовые подписи и их комбинации.
3) Генерация ответа: на основе построенного графа сцены и вопроса формируется ответ.

Проведено сравнение нескольких MLLM (таблица 1), таких как Ferret [1], Shikra [2], SPHINX-Tiny [3] и Qwen2.5-VL [4], на небольшом датасете изображений с размеченными отношениями. Полученные языковые метрики (EM, BLEU, METEOR, Bert-Score) позволили выделить лучшую модель Queen 2.5-VL. В данном эксперименте наиболее репрезентативной оказалась метрика Bert-Score.

Таблица 1. Сравнение MLLM

| | Exact Match | BLEU | METEOR | BERTScore |
|---|---|---|---|---|
| **Ferret** | 0.0565 | 0.0011 | 0.0535 | 0.7793 |
| **Shikra** | 0.0983 | 0.0000 | 0.0530 | 0.7825 |
| **Sphinx-Tiny** | 0.0491 | 0.0042 | 0.0783 | 0.8351 |
| **Qwen2.5-VL** | 0.1130 | 0.0018 | 0.1033 | 0.8041 |

Для данной модели дополнительно исследовалось влияние формата подачи данных об объектах на точность предсказания отношения междуними, вопрос

содержал список из 50 возможных ответов (таблица 2). Рассмотрены варианты нумерации объектов на изображении, указания названия объектов в вопросе и совмещение обоих подходов, которое позволило лучше всего выбрать ответ.

Таблица 2. сравнение формата подачи данных об объектах

| | top1 acc | top2 acc | top3 acc |
|---|---|---|---|
| номера на объектах | 0.0513 | 0.1026 | 0.1282 |
| названия объектов | 0.1795 | 0.3077 | 0.359 |
| номера + названия | 0.2564 | 0.4359 | 0.4872 |

Полученные результаты демонстрируют потенциал использования MLLM в задачах VideoQA и обосновывают необходимость комплексного подхода, объединяющего визуальные эмбеддинги, сегментационные модели и языковые механизмы.

**Благодарности**

Дмитрий Юдин (Заведующий лабораторией интеллектуального транспорта МФТИ - НКБ ВС),
Сергей Линок (Научный работник лаборатории интеллектуального транспорта МФТИ - НКБ ВС).

**Литература**

[1] You H., Zhang H., Gan Z., Du X., Zhang B., Wang Z., Cao L., Chang S.-F., Yang Y. Ferret: Refer and Ground Anything Anywhere at Any Granularity // arXiv preprint arXiv:2310.07704, 2023.

[2] Chen K., Zhang Z., Zeng W., Zhang R., Zhu F., Zhao R. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic // arXiv preprint arXiv:2306.15195, 2023.

[3] Lin Z., Liu C., Zhang R., Gao P., Qiu L., Xiao H., Qiu H., Lin C., Shao W., Chen K., Han J., Huang S., Zhang Y., He X., Li H., Qiao Y. SPHINX: The Joint Mixing of Weights, Tasks, and Visual Embeddings for Multi-modal Large Language Models // arXiv preprint arXiv:2311.07575, 2023.

[4] Wang P., Bai S., Tan S., Wang S., Fan Z., Bai J., Chen K., Liu X., Wang J., Ge W., Fan Y., Dang K., Du M., Ren X., Men R., Liu D., Zhou C., Zhou J., Lin J. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution // arXiv preprint arXiv:2409.12191, 2024.

[5] Ravi N., Gabeur V., Hu Y.-T., Hu R., Ryali C., Ma T., Khedr H., Rädle R., Rolland C., Gustafson L., Mintun E., Pan J., Alwala K.V., Carion N., Wu C.-Y., Girshick R., Dollár P., Feichtenhofer C. SAM 2: Segment Anything in Images and Videos // arXiv preprint arXiv:2408.00714, 2024.