УДК 004.89

МУЛЬТИМОДАЛЬНОЕ РАСПОЗНАВАНИЕ И СЕМАНТИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ СТРУКТУРНЫХ СЛАЙДОВ НА ОСНОВЕ КОМПЬЮТЕРНОГО ЗРЕНИЯ

Солодкая М.А.[1]
Научный руководитель – канд. техн. наук Кугаевских А.В.[1]

[1]Университет ИТМО
msolodkaya@itmo.ru

Введение

Современные образовательные материалы требуют не только визуальной привлекательности, но и смысловой точности в подаче информации. Преподаватели сталкиваются с проблемой оценки слайдов не только на предмет соответствия шаблону, но и с точки зрения логики размещения данных: корректно ли диаграмма иллюстрирует текст, не противоречит ли заголовок содержимому изображения, уместно ли расположены блоки относительно друг друга. Существующие системы проверки анализируют либо «сырой» текст, либо картинку изолированно, теряя связь между этими модальностями. Таким образом, возникает потребность в инструменте, способном к целостному анализу слайда как единого смыслового пространства.

Основная часть

Цель данной работы – повышение качества семантической интерпретации с использованием подхода к мультимодальному распознаванию структуры слайдов на основе визуально-языковых моделей (Vision Language Models - VLM). В отличие от классических методов компьютерного зрения, ограниченных детекцией объектов, предлагаемое решение будет использовать возможности VLM для анализа слайда как единого пространства. Актуальность применения VLM в образовательном контексте подтверждается работами, демонстрирующими их потенциал для анализа визуальных данных, включая изображения учебных материалов, и снижение барьеров доступности для преподавателей [1].

Модель не просто идентифицирует наличие заголовка или диаграммы, но и понимает их семантические взаимосвязи, роли в нарративе и соответствие заложенным дидактическим паттернам. Современные исследования в области мультимодального понимания видео-лекций, такие как фреймворк PreMind, уже демонстрируют эффективность использования VLM для сегментации слайдов, извлечения визуального контента и интеграции речи и изображения в единое понимание [2]. Другие работы, например, создание мультимодального «учебника» на основе тысяч часов instructional video, подчеркивают важность качественных, логически связных данных с высоким уровнем согласованности изображений и текста для обучения VLM решению емких и интенсивных задач.

Особенность предлагаемого подхода — контекстный инжиниринг через промпт-дизайн и fine-tuning VLM, позволяющий модели отвечать на вопросы вида: «Соответствует ли подпись под графиком его визуальному тренду?», «Является ли данный блок основным текстом или выноской?», «Формируют ли элементы слайда логическую последовательность для чтения?», «Корректно ли размещены элементы относительно друг друга с точки зрения восприятия?». Применимость VLM в подобных задачах активно исследуется, например, в работах по созданию датасетов для многомодальных диалогов, где баланс данных оказывается критичнее их простого объема [3]. Это позволяет выявлять не только формальные ошибки вёрстки, но и глубинные смысловые рассогласования в учебных материалах.

Разрабатываемый модуль станет основой интеллектуального ассистента в платформе помощи преподавателю, способного предоставлять контекстные рекомендации по улучшению композиции и семантической цельности слайдов.

Выводы

Предлагаемый подход позволяет выявить глобальные взаимосвязи за счёт использования контекста, в то время как классические нейронные сети позволяют выделить только локальный контекст. Результаты работы будут полезны профессорско-преподавательскому составу, специалистам в области педагогического дизайна и исследователям, работающим над задачами анализа документов и визуально-текстовых корреляций с применением больших мультимодальных моделей.

Литература

1. Realizing Visual Question Answering for Education. (2025) Chungnam National University. [Электронный ресурс] URL: https://library.cnu.ac.kr/eds/detail/eric_EJ1467530?briefLink=%2Feds%2Fbrief%2FdiscoveryResult%3Fst%3DKWRD%26service_type%3Dbrief%26si%3DAU%26q%3D%2522Xiaoming%2BZhai%2522%26 (дата обращения: 10.01.2026).
2. Wei K. et al. Premind: Multi-agent video understanding for advanced indexing of presentation-style videos //arXiv preprint arXiv:2503.00162. – 2025.
3. Yan D. et al. Mmcr: Advancing visual language model in multimodal multi-turn contextual reasoning //arXiv preprint arXiv:2503.18533. – 2025.