DOI: 10.15514/ISPRAS-2026-38(6)-33

Мультимодальный графовый подход к генерации с дополнительным поиском

Д.Е. Косогоров, ORCID: 0009-0005-7118-6724 <dmitry.kosogorov@icc.ru>
Д.Е. Копылов, ORCID: 0009-0000-6348-4004 <it-daniil@yandex.ru>
И.И. Охотин, ORCID: 0009-0003-9600-913X <ilia.ohotin@yandex.ru>
А.И. Поддубный, ORCID: 0000-0002-2431-8692 <poddubnyyiv@yandex.ru>
К.В. Тобола, ORCID: 0009-0006-1014-451X <kirilltobola@icc.ru>
Н.О. Дородных, ORCID: 0000-0001-7794-4462 <nikidorny@icc.ru>
А.А. Михайлов, ORCID: 0000-0003-4057-4511 <mikhailov@icc.ru>

Институт динамики систем и теории управления имени В.М. Матросова СО РАН, Россия, 664033, г. Иркутск, ул. Лермонтова, д. 134.

Аннотация. В настоящее время метод генерации с дополнительным/расширенным поиском (Retrieval-Augmented Generation – RAG) стал эффективным решением фундаментальных проблем, которые присущи всем большим языковым моделям. Данный метод повышает возможности LLM за счет извлечения релевантных внешних знаний и их встраивания в контекст модели. В данной статье предложен новый мультимодальный графовый подход RAG для сложно-структурированных PDF-документов. Традиционные RAG-системы игнорируют пространственную структуру документов, страдают от проблемы так называемого «модального разрыва» и не моделируют связи между отдельными фрагментами документа. Разработанное решение объединяет регионно-ориентированное разбиение на фрагменты, специализированный табличный кодировщик на базе дообученной мультимодальной модели и интеграцию структурного и семантического графов в единый конвейер извлечения и генерации. Предварительная экспериментальная оценка на наборах данных MMLongBench-Doc и LongDocURL подтвердила превосходство метода разбиения на фрагменты с использованием регионов над классическим методом RAG по всем метрикам оценки на этапе извлечения (поиска). Результаты подтверждают гипотезу о необходимости явного моделирования сложных межмодальных зависимостей. Подход открывает перспективы для создания надежных систем анализа корпоративной, научной и технической документации.

Ключевые слова: обработка естественного языка; генерация с дополнительным поиском; мультимодальные большие языковые модели; сложно-структурированные документы; векторные представления; структурный граф документа; семантический граф связности.

Для цитирования: Косогоров Д.Е., Копылов Д.Е., Охотин И.И., Поддубный А.И., Тобола К.В., Дородных Н.О., Михайлов А.А. Мультимодальный графовый подход к генерации с дополнительным поиском. Труды ИСП РАН, том 38, вып. 6, часть 2, 2026 г., стр. 199–226. DOI: 10.15514/ISPRAS–2026–38(6)–33.

Благодарности: Работа выполнена в рамках государственного задания Министерства науки и высшего образования Российской Федерации (Методы и технологии обучения представлений данных и самообучения для задач обработки изображений и естественного языка. Тема № 124051500067-7.

A Multimodal Graph-Based Approach for Retrieval Augmented Generation

D.E. Kosogorov, ORCID: 0009-0005-7118-6724 <dmitry.kosogorov@icc.ru>
D.E. Kopylov, ORCID: 0009-0000-6348-4004 <it-daniil@yandex.ru>
I.I. Okhotin, ORCID: 0009-0003-9600-913X <ilia.ohotin@yandex.ru>
A.I. Poddubny, ORCID: 0000-0002-2431-8692 <poddubnyyiv@yandex.ru>
K.V. Tobola, ORCID: 0009-0006-1014-451X <kirilltobola@icc.ru>
N.O. Dorodnykh, ORCID: 0000-0001-7794-4462 <nikidorny@icc.ru>
A.A. Mikhailov, ORCID: 0000-0003-4057-4511 <mikhailov@icc.ru>

Matrosov Institute for System Dynamics and Control Theory of the Russian Academy of Sciences, 134, Lermontov st., Irkutsk, 664033, Russia.

Abstract. Retrieval-Augmented Generation (RAG) has become an effective solution to fundamental problems inherent in all Large Language Models (LLM). This method enhances LLM by extracting relevant external knowledge and embedding it in the model's context. This paper proposes a new multimodal graph-based approach, RAG, for complexly structured PDF documents. Traditional RAG systems ignore the spatial structure of documents, suffer from the so-called "modal gap" problem, and do not model the relationships between individual document fragments. The developed solution combines region-oriented chunking, a specialized table encoder based on a pre-trained multimodal model, and the integration of structural and semantic graphs into a single retrieval and generation pipeline. Preliminary experimental evaluation using the MMLongBench-Doc and LongDocURL benchmarks confirmed the superiority of the region-based chunking method over the classic RAG method across all evaluation metrics at the retrieval stage. The results support the hypothesis of the need for explicit modeling of complex intermodal dependencies. This approach opens up prospects for creating reliable systems for analyzing corporate, scientific, and technical documentation.

Keywords: natural language processing; retrieval-augmented generation; large language models; large multimodal language models; complexly structured documents; embedding.

For citation: Kosogorov D.E., Kopylov D.E., Okhotin I.I., Poddubny A.I., Tobola K.V., Dorodnykh N.O., Mikhailov A.A. A Multimodal Graph-Based Approach for Retrieval Augmented Generation. Trudy ISP RAN/Proc. ISP RAS, vol. 38, issue 6, 2026, pp. 199-226 (in Russian). DOI: 10.15514/ISPRAS-2026-38(6)-33.

Acknowledgements. The research was carried out within the state assignment of Ministry of Science and Higher Education of the Russian Federation (Methods and technologies of representation learning and self-supervised learning for image processing and natural language tasks. Theme no. 124051500067-7.

1. Введение

Большие языковые модели (*Large Language Model, LLM*), основанные на архитектуре трансформер, достигли значительных успехов во многих задачах, связанных с обработкой естественного языка [1]. Благодаря предварительному обучению на обширных высококачественных наборах данных, LLM могут изучать широкий спектр языковых шаблонов, структур и фактических знаний. Они могут генерировать человекоподобный текст с высокой степенью семантической связности, а также демонстрировать высокие результаты в задачах ответа на вопросы (*Question Answering*), что свидетельствует об их способности понимать и решать различные задачи [2-3]. Однако, несмотря на их впечатляющие возможности, LLM по-прежнему сталкиваются со значительными ограничениями. Одна из основных проблем заключается в их низкой производительности в предметных узкоспециализированных областях или задачах, требующих специфических знаний. Хотя LLM часто обучаются на разнообразных и обширных наборах данных, такие наборы могут не охватывать всю глубину знаний, необходимых для решения предметной задачи. Также следует отметить, что у LLM внутри себя нет механизмов обновления информации в

реальном времени. Это может быть особенно проблематично в таких областях, как медицина, юриспруденция, финансы и другие технические области, где точность и актуальность знаний должны быть приоритетными. Другой очень важной проблемой LLM являются «галлюцинации». При обработке запросов, выходящих за рамки знаний, полученных в ходе обучения, или требующих самой актуальной информации, LLM могут генерировать заведомо неверные ответы, вводящие пользователя в заблуждение. Минимизация случаев галлюцинаций важна для повышения надежности LLM в предоставлении точной и контекстно-релевантной информации в различных областях [3].

В последнее время метод генерации с расширенным поиском (*Retrieval-Augmented Generation, RAG*) стал эффективным решением отмеченных выше проблем, повышая возможности LLM за счет извлечения релевантных внешних знаний [4-5]. Существующие системы RAG обычно работают в два этапа: поиск (*retrieval*) и генерация (*generation*) [6]. На этапе поиска цель состоит в том, чтобы быстро найти релевантные знания, семантически схожие с запросом, в большой коллекции документов. Поскольку релевантные знания часто разбросаны по различным частям документов, каждый документ предварительно разбивается на несколько фрагментов. На этапе генерации извлеченные фрагменты документов объединяются с запросом для формирования расширенного ввода. Этот расширенный ввод предоставляет LLM дополнительный контекст, включающий внешние знания. Кроме того, RAG позволяет LLM динамически интегрировать самую актуальную информацию на этапе вывода. Эта возможность гарантирует, что ответы модели основаны не только на статических, предварительно обученных знаниях, но и постоянно обновляются с учетом актуальных и релевантных данных. Извлекая и используя внешние знания, RAG обосновывает генерируемые ответы фактической информацией, тем самым значительно снижая вероятность возникновения галлюцинаций [7]. Однако классические RAG-систем, как правило, сосредоточены на базах знаний, построенных на основе простого текста и LLM, предварительно обученных только на тексте, игнорируя другие богатые источники знаний, такие как таблицы, изображения, или видео. В реальных условиях люди естественным образом взаимодействуют с такими мультимодальными данными, например, просматривая отчеты в формате PDF (Portable Document Format), сочетающие в себе текст, таблицы, формулы и изображения в смешанной форме. Анализируя изображение вместе с текстом, пользователь может лучше понять контекст информации, что в свою очередь повышает удовлетворенность качеством ответов.

На данный момент объем мультимодальных данных в мире постоянно растет, охватывая текстовые документы, изображения, таблицы и видео. Это создает критическую потребность в интеллектуальных системах, способных эффективно извлекать и синтезировать знания из разнородных источников информации. Перенос возможностей LLM, работающих изначально с простым текстом, на область мультимодальных данных является активной областью научных исследований. Эффективное включение мультимодальных данных имеет важное значение для повышения возможностей LLM. Так в последние годы разработка мультимодальных генеративных моделей (*Multimodal Large Language Model, MLLM*, или иначе, *Large Vision Language Model, LVLM*), способные одновременно понимать и обрабатывать визуальную (изображения, видео) и текстовую информацию, продемонстрировала дополнительные возможности их применения в различных сценариях [8-10].

Недавние исследования [11] показали, что RAG с мультимодальными данными превосходит традиционный RAG, основанный только на текстовой информации. За счет улучшения возможностей генерации MLLM путем извлечения внешних мультимодальных знаний, система мультимодального RAG может дополнительно повысить возможности и качество ответов на вопросы, тем самым еще больше уменьшая проблемы, связанные с галлюцинациями. Однако анализ современных работ выявляет три ключевые проблемы в области мультимодального RAG:

Во-первых, большинство существующих систем страдают от так называемого «модального разрыва» (*modality gap*), то есть тенденции кластеризации векторных представлений по модальностям вместо семантического сходства [12-13]. Это приводит к тому, что текстовые запросы систематически тяготеют к текстовым блокам, игнорируя релевантный визуальный контент, даже если ответ содержится именно в изображении (например, диаграмме).

Во-вторых, традиционные подходы на этапе поиска обрабатывают документы как коллекцию независимых фрагментов (*chunks*), не моделируя структурные и семантические связи между ними. Это особенно критично для длинных документов, где информация часто распределена по нескольким страницам, а понимание контекста требует интеграции знаний из различных частей документа.

В-третьих, существующие методы не учитывают структуру документов, что ограничивает их применимость для задач, требующих точной навигации по документу. Так ни один из проанализированных подходов [12-22] не моделирует явно структурные отношения между элементами документа.

Таким образом, в данной работе предлагается новый мультимодальный графовый подход RAG, который решает указанные проблемы через интеграцию структурного графа и семантического графа связности в один общий конвейер (*pipeline*) извлечения и генерации.

Основной вклад данной работы заключается в следующем:

- Предложен новый способ разбиения исходных PDF-документов на отдельные модальные сегменты (регионы) с учетом иерархической структуры документа совместно со специализированной стратегией векторизации табличных данных. Это позволяет преодолеть «модальный разрыв» и формировать плотные векторные представления, сохраняющие двумерную топологию, семантику заголовков и отношения между ячейками.
- Впервые осуществлена интеграция структурного и семантического графа в общий конвейер извлечения и генерации. В отличие от плоских векторных индексов, такой подход позволяет учитывать контекстную вложенность фрагментов документа и межмодальные зависимости, что существенно повышает релевантность поиска в сложно-структурированных PDF-документах.
- Проведена предварительная экспериментальная оценка на эталонных тестовых наборах длинных мультимодальных документов, демонстрирующая перспективность метода сегментации документов на этапе извлечения (поиска), особенно для многостраничных и кросс-модальных запросов. Разработанная система реализована в виде открытой масштабируемой микросервисной архитектуры, что обеспечивает воспроизводимость результатов и готовность к практическому применению.

Статья организована следующим образом: Раздел 2 представляет всесторонний обзор современного состояния исследований в области RAG, в том числе мультимодальных подходов. Раздел 3 детально описывает архитектуру предлагаемого решения и детали реализации. Раздел 4 представляет предварительную экспериментальную оценку и сравнительный анализ полученных результатов. В заключении формулируются итоговые выводы и планы будущей работы.

2. Современное состояние исследований

2.1 Технология RAG

Retrieval-Augmented Generation (RAG) – это архитектура, которая объединяет поиск информации (*retrieval*) с генерацией ответа (*generation*) [4-6]. Она позволяет LLM повысить точность ответов на вопросы, опираясь не только на свои внутренние знания, но и на

актуальные, внешние данные. На практике системы RAG следуют трехэтапному конвейеру, который изображен на Рис. 1.

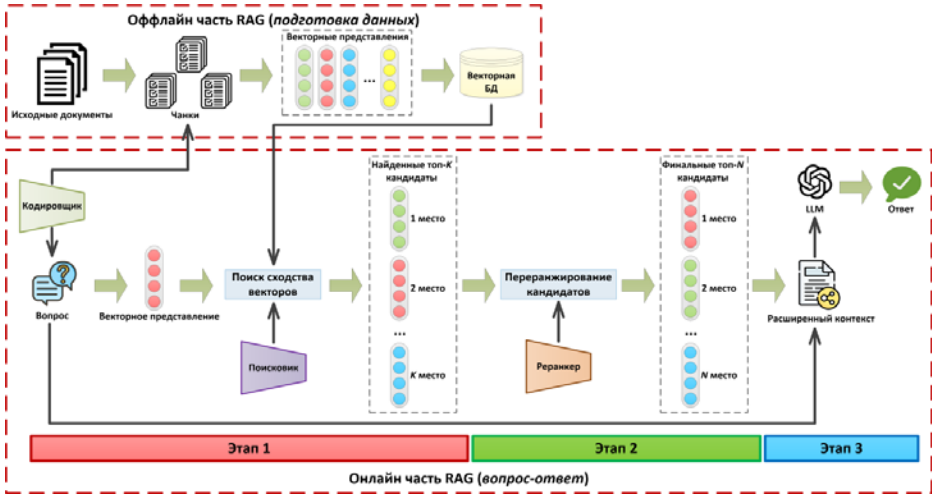


Рис. 1. Концептуальная схема традиционного трехэтапного подхода RAG.
Fig. 1. Conceptual diagram of the traditional three-stage RAG approach.

Фаза подготовки данных (*оффлайн часть RAG-системы*) – внешние знания (например, внутренние корпоративные документы, базы данных, технические спецификации) разбиваются на фрагменты (*chunks*), которые кодируются в векторные представления с помощью специальных кодировщиков (например, BERT, E5, BGE) и сохраняются в векторной базе данных (например, FAISS, Milvus, ChromaDB, Pinecone) с применением индексации [6]. В частности, популярным является применение ANN-индексов с использованием метода приближенного поиска ближайшего соседа (Approximate Nearest Neighbor). При такой подготовке данных критически важно выбрать оптимальную гранулярность (слишком мелкие фрагменты теряют контекст, слишком крупные вносят шум). Поэтому классические RAG-системы, которые строятся на векторной базе, обычно фокусируются на векторных представлениях и разных стратегиях разбиения на фрагменты (*chunking*), например, семантического или структурного разбиения. Однако такие решения, как правило, игнорируют структурные связи между различными блоками внутри сложно-структурированных документов.

Фаза «вопрос-ответ» (*онлайн часть RAG-системы*):

Этап 1: Пользователь задает вопрос, который кодируется в векторное представление (*embedding*). Далее с помощью специального механизма поиска (*retriever*) из векторной базы данных извлекаются топ-*k* самых похожих фрагментов. При этом обычно применяются ANN-методы, которые используются для быстрого поиска ближайших векторов в больших базах данных. Вместо точного сравнения запроса со всеми векторами, ANN-индекс позволяет найти приблизительно ближайшие векторы (топ-*k*) за меньшее время, что делает его эффективным для приложений, требующих высокой производительности, таких как системы рекомендаций или поиск изображений. Помимо этого используются методы, как плотного поиска (*dense retrieval*), встраивающие запрос пользователя и исходные документы в непрерывное векторное пространство по определенному критерию, например, семантической близости [23], так и разреженного поиска (*sparse retrieval*), который обычно опирается на сопоставление инвертированного индекса с исходными данными. Например, многие исследования непосредственно применяют TF-IDF или BM25 для поиска на уровне

отдельных отрывков текстов [24-25]. Современные системы, например ITER-RTGEN [26], часто комбинируют оба подхода (*hybrid retrieval*), чтобы захватить как семантическую близость, так и точные термины.

Этап 2: После механизм переранжировки (*reranker*) для повышения точности пересортирует эти *k*-фрагментов, выбирая топ-*n* самых релевантных.

Этап 3: Выбранные *n*-фрагментов формируют расширенный контекст, который вместе с исходным вопросом подается в LLM (*generator*), который использует эту контекстную информацию для получения более обоснованных и согласованных ответов.

Недавние достижения в области фреймворков (*framework*) RAG, такие как: поиск с помощью планирования [27], агентный RAG [28] и итеративное уточнение на основе обратной связи [29], еще больше улучшили как компоненты поиска, так и компоненты генерации в RAG-системах.

На сегодняшний день, основными парадигмами поиска в системах RAG являются:

1) **Текстовый поиск (*text retrieval*)** – все элементы исходного документа представляются в виде текста, при этом таблицы линеаризуются в текст, а изображения либо пропускаются, либо используются только подрисуночные подписи, либо тоже переводятся в текст с использованием специальных моделей (*image2text*). Весь полученный таким образом текст разбивается на фрагменты, после чего используются модели для векторизации фрагментов текста, например, text-embedding-3-small [30].
2) **Поиск по изображению (*image retrieval*)** – все элементы исходного документа представляются в виде растровых изображений (одно изображение может выступать в виде фрагмента). Далее используются модели для векторизации изображений, например, colqwen2.5 [31].
3) **Мультимодальный комбинированный поиск (*multi-modal text+image retrieval*)** – применение к исходным документам отдельно двух выше упомянутых парадигм. Например, берутся топ-5 (или больше) результатов текстового поиска и топ-5 результатов поиска изображений. Полученные отдельные векторные представления текста и изображения подаются в целевую модель совместно, без какого-либо объединения (слияния). В частности, в работе [32] именно этот подход стал самым эффективным по результатам экспериментов.
4) **Мультимодальный объединенный (совместный) поиск (*multi-modal joined retrieval*)** – применяются специальные техники и модели объединения (слияния) векторных представлений исходных элементов документов (текста и изображений) в один общий вектор. Ключевая цель – это создать единое векторное пространство, в котором можно вычислять семантическое сходство между любыми модальностями. На сегодняшний день известны такие подходы как: General Multimodal Embedder (GME) [33], VLM2Vec-V2 [34], Uni-Retrieval [35], SeBe [36] и другие. В общем виде создание мультимодальных (унифицированных) векторных представлений (*unified embeddings*) подразумевает: (1) Использование MLLM в качестве основы. Так большинство работ используют современные MLLM (например, Qwen3-VL, LLaVA, Phi-3.5V) в качестве базовой модели (*backbone*). Эти модели уже обучены на парах «изображение-текст» и способны принимать последовательности в виде текста и изображения/видео. Вместо генерации текста, финальное скрытое состояние токена [EOS] (или аналогичного специального токена маркера) используется как векторное представление всего входа. (2) Использование контрастного обучения (*contrastive learning*) или инструкций (*instruction tuning*) для дообучения базовых моделей. Также может применяться метод LoRA (*Low-Rank Adaptation*) для эффективной тонкой настройки (*fine-tuning*) модели, который замораживает основные веса модели и обучает лишь небольшие дополнительные

матрицы вместо полного дообучения. Это особенно важно при работе с большими MLLM.

Современные тенденции уходят от классического «наивного» RAG к модульным, адаптивным и многошаговым подходам. Например, подход FLARE [37] использует гипотезы о будущем токене для активации поиска, а Self-RAG [38] дополняет систему механизмами саморефлексии, при которой модель критикует релевантность найденного контекста и качество собственного ответа. В работе [39] предложен принципиально новый подход xRAG к сжатию контекста в RAG-системах, основанный на модальном слиянии текстовых и векторных представлений. xRAG позволяет заменить весь извлеченный документ (фрагмент) всего одним токеном, который является проекцией его векторного представления в пространство языковой модели. Это позволяет достичь рекордного коэффициента сжатия (до 178 раз) и значительного ускорения вывода (*inference*) без потери качества. Однако xRAG пока ограничен простыми задачами и обработкой только одного документа. Также появляются комплексные работы [40], где исследуются отдельные аспекты RAG-процесса. В частности, рассматривается влияние размера базы знаний (хранилища данных) и включения документов, представленных на различных языках, на общую производительность RAG-системы. Кроме этого исследуются такие вопросы как: (а) повышение точности модели путем расширение исходного запроса, и (б) включение контрастивных примеров в контекстное обучение модели при формировании ответов RAG-системой.

Таким образом, за последнюю пару лет появились тысячи новых подходов. На Рис. 2 представлена обобщенная таксономия, которая иллюстрирует эволюцию и взаимосвязи между разными подходами RAG, выделяя ключевые ветви (направления) такие как [41]: классический «наивный» RAG, расширенный и модульный RAG, агентный RAG, графовый RAG, мультимодальный RAG, а также их конвергенцию в единую мультимодальную агентную технологию RAG. При этом техники графового RAG [42] могут быть наложены на любое направление для поддержки реляционного и многошагового рассуждения. Далее подробнее рассмотрим направление работ, связанных с мультимодальным RAG.

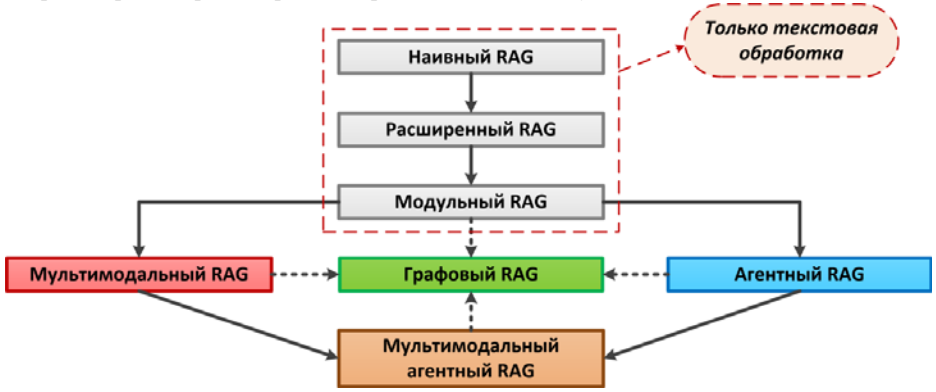


Рис. 2. Обобщенная таксономия подходов RAG.
Fig. 2. Generalized taxonomy of RAG approaches.

2.2 Мультимодальный RAG

Технология RAG прошла значительный путь эволюции от текст-ориентированных систем к полноценным мультимодальным фреймворкам. С появлением MLLM (например, GPT-4V, LLaVA, Qwen3-VL) исследователи начали интегрировать визуальный контент в конвейеры RAG. Сами по себе MLLM хоть и демонстрируют способность понимать изображения и текст совместно, однако они не предназначены для эффективного поиска в крупных корпусах

документов и не сохраняют структурные связи. Напротив, существующие модели для структурного анализа документов, такие как LayoutLM [43] и DocLLM [44], хоть и могут учитывать макет (*layout*) документа и структурные связи для лучшего понимания, однако они напрямую не применяются к RAG-системам ориентированных на поиск и генерацию. Таким образом, мультимодальный RAG стал быстро развивающейся областью, которая включает в себя различные источники знаний, такие как изображения, видео, аудио и сложно-структурированные документы, для обогащения входного контекста MLLM и повышения качества генерации за счет использования кросс-модальных сигналов [14].

Так, работы вроде MuRAG [15] продемонстрировали потенциал мультимодального извлечения, однако их подходы ограничивались простым объединением текстовых и визуальных векторных представлений без учета семантических отношений между модальностями.

Подход ColPali [16] представляет революционную концепцию «извлечения в пространстве зрения» (*Retrieval in Vision Space*), где документы обрабатываются как изображения без предварительного извлечения текста. Модель использует специальный механизм для вычисления сходства между запросом и страницами документа, достигая высокой точности на наборе данных ViDoRe при значительно меньшей задержке по сравнению с традиционными конвейерами RAG. Однако этот подход имеет существенное ограничение: он обрабатывает страницы независимо, не моделируя связи между ними, что критично для многостраничных документов с распределённой информацией.

Подходы MDocAgent [17] и HM-RAG [18] используют специализированных агентов для обработки различных аспектов документа. MDocAgent включает пять агентов (общий, критический, текстовый, визуальный, суммирующий), работающих совместно через параллельный извлечение текста и изображений. Эксперименты показывают улучшение на 12% по сравнению с предыдущими решениями на наборе данных MMLongBench-Doc [45]. HM-RAG расширяет эту концепцию до трехуровневой иерархической архитектуры с декомпозицией запроса, применением мульти-извлечения на основе векторок, графов и веб-поиска, а также механизмом консистентного голосования. Система достигает высокой точности на наборе данных ScienceQA [46], однако страдает от высокой задержки из-за последовательной работы агентов.

Подходы UniversalRAG [12] и UniRAG [13] решают проблему так называемого «модального разрыва» (*modality gap*) через динамическую маршрутизацию запросов к наиболее подходящим модально-специфичным блокам информации. Вместо объединения всех модальностей в единое векторное пространство, система предсказывает оптимальную комбинацию модальностей и уровней детализации для каждого запроса.

Метод мультимодального разбиения документа на фрагменты (*vision-guided chunking*) для RAG-систем, представленный в работе [19], использует MLLM для обработки сложных PDF-документов настраиваемыми пакетами страниц с сохранением контекста между пакетами, что обеспечивает точную обработку таблиц, охватывающих несколько страниц, встроенных визуальных элементов и процедурного контента. Такое мультимодальное разбиение сложных документов на пакеты сохраняет семантическую целостность и структурную связность, что значительно улучшает качество извлечения информации в RAG. Однако обработка больших документов требует значительных вычислительных ресурсов, а решение в целом сильно зависит от качества и возможностей исходной MLLM. Кроме того, очень длинные таблицы (занимающие несколько страниц) не обрабатываются, а сложные схемы и диаграммы могут быть не полностью распознаны.

В работе [20] авторы предлагают новый легковесный фреймворк SV-RAG для понимания многостраничных визуально насыщенных документов с помощью MLLM. Ключевая идея – вместо подачи всех страниц документа в модель (что приводит к перегрузке памяти и шуму от нерелевантного контента), использовать саму MLLM в качестве мульти-модального

поиска релевантных страниц, а затем генерировать ответ только на основе этих страниц. Это решает фундаментальную проблему масштабируемости MLLM на длинных документах. Фреймворк состоит из двух модулей, использующих единую предварительно обученную MLLM с двумя наборами адаптеров LoRA для поиска и генерации ответа.

В [21] представлена новая система VDocRAG, которая может напрямую понимать различные документы и модальности в едином формате изображения, предотвращая потерю информации, возникающую при разборе документов для получения текста. Для повышения производительности предложено решать две задачи самообучения, которые адаптируют большие модели визуально-языковой обработки для поиска путем сжатия визуальной информации в плотные токен-представления и их сопоставления с текстовым содержимым документов.

Система Pythia-RAG [22] строит единый мультимодальный граф знаний из текстовых и визуальных триплетов на основе исходных мультимодальных данных. Затем она использует специальный алгоритм Штейнера о минимальном дереве для извлечения топологически связного подграфа на этапе поиска. Это позволяет достичь высокой точности на наборе данных ScienceQA [46] и в целом снижает галлюцинаций модели. Однако Pythia-RAG фокусируется на интеграции внешних знаний через графы, а не на моделировании структурных связей внутри документа. Кроме того, подход требует сложного конвейера извлечения триплетов и построения глобального графа знаний, что ограничивает его масштабируемость.

Одним из основных недостатков всех рассмотренных подходов является отсутствие полноценного учета структуры документа. Как правило, подходы не учитывают в полной мере структуру документов (не моделирует явно межстраничные или межблочные связи, оглавление или иерархию разделов, связи текста с таблицами и рисунками и т.п.). Это критическое ограничение для задач навигации по документу, особенно для технической документации (инструкций, спецификаций), юридических и финансовых отчетов и научной литературы. Другим очень важным недостатком является ограниченная поддержка межмодальной интеграции. Большинство подходов обрабатывают модальности параллельно, но не моделируют явные связи между ними. Например, визуальный элемент на одной странице может быть связан с текстовым описанием на другой странице, но ни один из существующих методов не учитывает такие межстраничные связи. Также, несмотря на усилия по выравниванию модальностей (UniversalRAG, ColPali и др.), проблема кластеризации векторных представлений по модальностям вместо семантического сходства остается нерешенной. Это особенно критично для гибридных запросов, требующих интеграции информации из разных модальностей.

3. Предлагаемый подход

3.1 Общая архитектура подхода

Предлагаемый подход представляет собой мультимодальный графовый фреймворк RAG, который интегрирует структурный граф документа и семантический граф связности в единый конвейер извлечения и генерации. Архитектура системы состоит из пяти основных этапов, три из которых приходятся на оффлайн фазу подготовки данных (Рис. 3), а два других на онлайн фазу ответов на вопросы (Рис. 4). Далее подробнее рассмотрим этапы подхода.

Этап 1: Разбиение на фрагменты с использованием регионов документов. Традиционные подходы к процессу разбиения исходных документов на набор отдельных блоков текста (фрагментов) основаны на линейной логике, где последовательность символов режется по достижении определенного предела токенов или по границам предложений/абзацев. Однако сложно-структурированные документы (например, PDF-отчеты, научные статьи, слайды презентаций) обладают двумерной визуальной структурой. Человек обычно воспринимает

визуально-структурные регионы документа как единый объект при просмотре, например, таблица целиком, блок диаграммы с подписью, карточка товара, колонка текста или выделенная врезка. Поэтому традиционные методы разбиения документа на фрагменты представляются неэффективными для такого рода данных. На данном этапе осуществляется сегментация исходных сложно-структурированных документов с выделением отдельных структурных элементов (регионов/сегментов) таких как: «*Текст*» (*Text*), «*Заголовок*» (*Header*), «*Таблица*» (*Table*), «*Изображение*» (*Image*), «*Нижний колонтитул изображения*» (*Image_Footer*), «*Подпись к изображению*» (*Image_Caption*), «*Нижний колонтитул таблицы*» (*Table_Footer*), «*Подпись к таблице*» (*Table_Caption*), «*Формула*» (*Equation*). При этом под элементом «*Изображение*» понимается любой рисунок, который встречается в тексте. Например, это может быть схема, график или диаграмма. Каждый такой регион представляет собой отдельный фрагмент. Выделение таких фрагментов-регионов может происходить на основе специальных эвристик (близость расположения, общий фон, соединительные линии) или методов машинного обучения (графовые нейросети, MLLM). Применение разбиения на фрагменты с использованием регионов документа позволяет повысить эффективность работы механизма поиска для смешанных (текст + графики), структурных (сравнение карточек товаров) и табличных запросов.

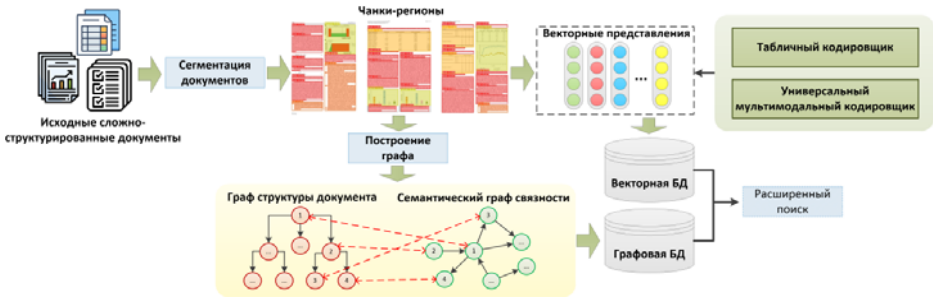


Рис. 3. Архитектура предлагаемого мультимодального графового подхода RAG в части оффлайн фазы подготовки данных.
Fig. 3. Architecture of the proposed multimodal graph-based RAG approach regarding the offline data preparation phase.

Этап 2: Векторизация мультимодальных данных. Извлеченные из документов фрагменты-регионы преобразуются в единое векторное пространство с использованием мультимодальной модели векторизации. Например, нумерованный список сериализуется как последовательный текст перед векторизацией. Однако простое преобразование таблицы в текст (линеаризация) является малоэффективным решением [47], при котором теряется пространственная структура и связи между ячейками, а применение обычной текстовой векторизации к таблицам не всегда хорошо передают семантику числовых данных и отношений между столбцами. На данном этапе вводится дополнительная специализированная модели векторизации для табличных данных, которая способна распознавать и понимать двумерную структуру таблицы. В качестве базовой модели может использоваться малоразмерная мультимодальная модель векторизации, которая дообучается с применением техники контрастного обучения (*contrastive learning*) [48] на различных табличных задачах (семантическая интерпретация таблиц, расширение таблиц, поиск и проверка фактов на табличных данных, связывание сущностей в столбцах и т.п.) [49]. При этом набор данных для обучения формируется за счет синтеза и агрегации открытых крупномасштабных табличных корпусов (например, TURL, HiTab, FeTaQA, WikiSQL, WikiTQ, TabFact, GitTables, WebTables, SOTAB [49]), нормализованных к единому визуально-текстовому формату. В процессе обучения модель учится явно кодировать

двумерную топологию таблицы, иерархию строк/столбцов, а также контекстные зависимости между заголовками и значениями ячеек. В результате формируется специализированный табличный кодировщик, способный генерировать высококачественные векторные представления, устойчивые к структурному шуму и сохраняющие реляционную семантику. Это позволяет повысить точность ответов на вопросы, требующие извлечения данных из таблиц.

Этап 3: Построение структурного графа и семантического графа связности. Большинство современных RAG-систем строят индекс на основе плоского набора фрагментов и используют косинусное расстояние для поиска в векторном пространстве. Такой подход игнорирует два важных измерения сложно-структурированного документа:

1) *иерархическая структура* (отношения «родитель–потомок», «часть–целое»), которая определяет контекстную вложенность фрагментов документа;
2) *семантические связи*, выходящие за рамки иерархии (например, перекрестные ссылки, тематическая близость, цитирование внутри документа).

Предполагается, что построение двух таких графов и их последующая интеграция позволит механизму поиска (извлечения) учитывать не только содержимое самого фрагмента, но и его положение в документе, а также неочевидные смысловые связи с другими фрагментами.

Таким образом, на данном этапе вначале строится ориентированный граф структуры документа, где вершины – это выделенные фрагменты-регионы, а ребра – это иерархические связи между этими регионами (раздел содержит подразделы, подраздел содержит текст и т.п.). При этом таблица становится одним узлом графа. Дополнительно можно создать узлы для отдельных строк или ячеек, связанных с родительской таблицей ребрами типа «содержится в». Это позволяет точечно извлекать только нужные строки, если таблица очень большая.

Далее строится семантический граф связности на основе:

- Семантической близости фрагментов-регионов. Для этого вычисляется косинусное сходство между всеми фрагментами-регионами. Таким образом, вершина в таком графе – это фрагмент-регион, а ребра – это веса сходства. Изначально это полный граф (каждая вершина связана со всеми другими вершинами). Однако он может быть урезан, путем отбрасывания ребер, веса которых ниже определенного порогового значения (например, 0.7).
- Тематического моделирования (кластеризации) фрагментов-регионов. Этот метод фокусируется не на локальной семантической близости чанков-регионов, а на глобальной структуре тем, выявляемой в исходном наборе документов. Основная цель – это сгруппировать фрагменты-регионы по смысловым темам. При этом использовать кластеры как узлы графа, а при поиске (извлечении) сначала определить релевантные темы, затем извлекать фрагменты-регионы из соответствующих кластеров. Таким образом, вначале все фрагменты-регионы кластеризуются. После для каждого кластера генерируется тематическая метка и строится его векторное представление (центроид).

Оба построенных графа интегрируются в один общий граф, позволяющий расширить начальный контекст генеративной модели за счет найденных соседей.

Этап 4: Поиск и извлечение через граф. На данном этапе запрос векторизуется и находится ближайшая вершина (фрагмент-регион) или несколько вершин (топ-*k* фрагментов-регионов) по векторному признаку. Затем выполняется обход в построенном графе: из стартовых вершин выбираются соседи (*1-hop*) по ребрам обоих типов с учетом весов. Например, можно отобрать вершины по иерархии до родительского раздела, чтобы получить заголовок, или перейти по семантическому ребру к смежному фрагменту, дополняющему ответ. При этом также осуществляется удаление дубликатов (если фрагмент-регион является соседом для

нескольких фрагментов-регионов из топ-*k*, он включается в финальный контекст только один раз). Собранный таким образом контекст подается в генеративную модель с исходным запросом.

Такой подход позволяет добиться контекстной полноты, где ответ будет опираться не только на прямое совпадение, но и на окружение (структурное и смысловое), а также устойчивости к неточным запросам (если запрос плохо сформулирован, но тематически связан с каким-то фрагментом, семантические ребра помогут найти нужную информацию). Более того, такой интерпретируемый поиск может помочь объяснить, почему найден тот или иной фрагмент-регион через цепочку ребер.

Этап 5: Генерация ответа. Синтез финального ответа на основе расширенного контекста с использованием MLLM.

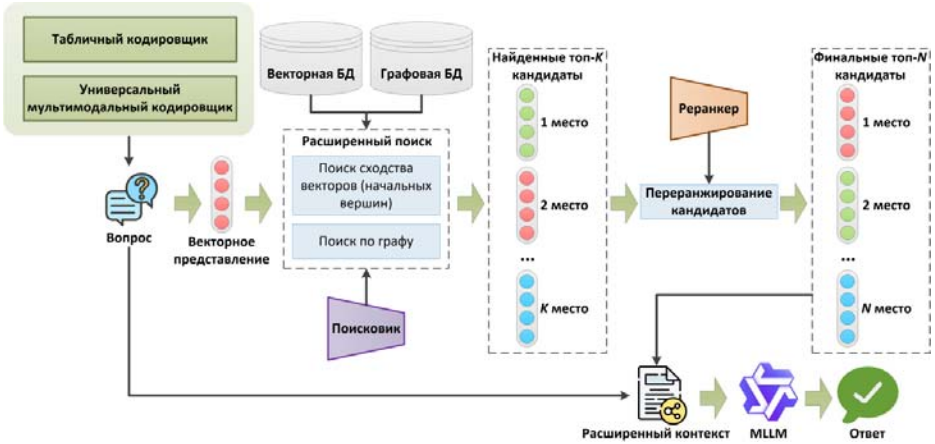


Рис. 4. Архитектура предлагаемого мультимодального графового подхода RAG в части онлайн фазы ответов на вопросы.
Fig. 4. Architecture of the proposed multimodal graph-based RAG approach regarding the online question-answering phase.

3.3 Детали реализации подхода

На данный момент реализован прототип предлагаемого подхода в виде набора взаимодействующих между собой сервисов (фреймворка). Для извлечения содержимого из разнообразных типов сложно-структурированных документов, представленных в формате PDF, используется библиотека MinerU [50], которая преобразует их в машиночитаемые форматы (Markdown или JSON). Библиотека использует наборы моделей PDF-Extract-Kit и OCR-распознаватели для сегментации документов (обнаружение и анализ макета, распознавание формул и таблиц и т.д.) и определения порядка чтения. При этом для хранения извлекаемых «сырых» регионов (неструктурированных данных, изображений) применяется высокопроизводительное объектное хранилище с открытым исходным кодом – MinIO [51].

В качестве мультимодальной модели векторизации используется малоразмерная модель – Qwen3-VL-Embedding 2B [52], поддерживающая смешанные модальные входные данные (текст, изображения, скриншоты, видео) для кросс-модального понимания.

Для единого хранения получаемых векторных представлений фрагментов-регионов применяется высокопроизводительная векторная база данных – Qdrant [53], которая позволяет также хранить граф структуры документа и семантический граф связности. Qdrant включает мощную поисковую систему для быстрого поиска по семантическому сходству векторов с высокой точностью. Qdrant также поддерживает определение семантической

близости через косинусное сходство между вершинами (фрагментов-регионов) хранимого графа.

Для тематического моделирования (кластеризации) фрагментов-регионов применяется подход GraphRAG [54], который использует LLM для построения графового индекса в два этапа: (1) сначала для получения графа знаний сущностей (*knowledge graph*) [55] из исходных документов, (2) затем для предварительной генерации сводных данных для всех кластеров тесно связанных сущностей.

В качестве основной MLLM для генерации ответов на вопросы используется средне-размерная модель – Qwen3-VL-32B [56], которая на сегодняшний день является одной из самых мощных моделей обработки визуальной и языковой информации в серии Qwen. Эта же модель используется для построения графового индекса GraphRAG.

Все перечисленные компоненты фреймворка инкапсулированы в изолированные Docker-контейнеры. Данное решение обеспечивает детерминированность экспериментального окружения, независимое горизонтальное масштабирование отдельных модулей и упрощает развертывание на гетерогенных вычислительных средах. Взаимодействие между контейнерами организовано через асинхронных REST API, что минимизирует задержки при передаче промежуточных векторных представлений и графовых данных.

4. Экспериментальная оценка

В данном разделе представлена предварительная экспериментальная оценка производительности разработанного подхода, включающего только этап разбиения на фрагменты с использованием регионов документов.

В качестве базового решения для сравнения предлагаемого подхода выбран классический подход «наивного» RAG [6], реализованный на базе фреймворка LlamaIndex, где документ разбивается на фрагменты по фиксированной длине в 512 токенов. При этом применялось строгое архитектурное ограничение, где фрагмент не мог выходить за пределы одной страницы. Данное условие обеспечивало однозначную идентификацию страницы-источника на этапе формализации метрик извлечения. Перекрытие (*overlap*) между смежными фрагментами не применялось. Векторизация осуществлялась при помощи мультимодальной модели Qwen3-VL-Embedding 2B, а индексация и поиск выполнялись через SimpleVectorStore с использованием метрики косинусного сходства.

Предлагаемый подход и данное базовое решение оценивались отдельно на этапах извлечения (поиска) информации и генерации итогового ответа. При этом в предлагаемом подходе размер фрагмента не является жестко заданной величиной, а определяется естественными границами структурных регионов документа (текстовые блоки, таблицы, изображения, формулы и т.п.). Меж-документные связи на текущем этапе не моделируются, поскольку экспериментальная задача ограничена анализом изолированных PDF-файлов.

Параметры семантического графа связности, включая количество извлекаемых сущностей, типов отношений и тематических кластеров в GraphRAG, в рамках данной сессии оценки не применялись, так как механизм обхода графа находился в стадии интеграции.

Все эксперименты проводились на кластере, в конфигурацию которого входят два 16-ти ядерных процессора Intel Xeon Gold 6326 «Ice Lake» 2.9 GHz, четыре графических процессора NVIDIA A100 80 GB PCIe и 2 TB оперативной памяти DDR4-3200.

4.1 Наборы данных для тестирования

Для всесторонней оценки предложенного подхода используются два тестовых набора данных: MMLongBench-Doc [45] и LongDocURL [57].

MMLongBench-Doc представляет собой комплексный набор данных, специально разработанный для оценки способности моделей понимать длинные мультимодальные

документы. Набор данных включает 135 документов в формате PDF, охватывающих 7 предметных областей: *научные отчеты* (25.2%), *финансовые отчеты* (8.1%), *академические статьи* (19.3%), *брошюры* (11.1%), *руководства* (16.3%), *административные и промышленные документы* (7.4%), а также *учебные материалы* и *воркшопы* (12.6%). Ключевой особенностью набора является значительная длина документов: в среднем 47.5 страниц (медиана 28) и 21214 текстовых токенов (медиана 12179), что создает реалистичные условия для оценки масштабируемости систем извлечения информации.

Набор содержит 1082 вопросов, аннотированных экспертами, и распределенных по трем категориям в соответствии с требованиями к источнику информации:

1) *одностраничные вопросы* (494 / 45.7%) – требуют информации с одной страницы документа для оценки локализации релевантного контента;
2) *многостраничные вопросы* (365 / 33.7%) – требуют интеграции информации с нескольких страниц (28.6% из них требуют более двух страниц), что позволяет оценить способность системы к межстраничному рассуждению;
3) *вопросы без ответов* (223 / 20.6%) – не имеют подтверждающих доказательств в документе, что критически важно для количественной оценки галлюцинаций модели.

Каждый вопрос требует информации из одного или нескольких источников: *чистый текст* (35.5%), *структурный макет документа* (13.9%), *таблицы* (25.4%), *диаграммы/графики* (20.7%) и *остальные изображения* (35.4%). Такое распределение обеспечивает сбалансированную оценку мультимодальных возможностей системы. Ответы представлены в четырех форматах: *строка* (29.1%), *целое число* (34.8%), *число с плавающей точкой* (18.5%) и *список* (17.6%), что позволяет оценивать точность генерации в различных сценариях.

Для каждого вопроса сохранена метаинформация: идентификаторы страниц-источников доказательств, типы модальностей, необходимых для ответа, и ожидаемый формат ответа. Эта информация используется для точной оценки этапа извлечения (поиска) и генерации ответа.

LongDocURL представляет собой наиболее комплексный на сегодняшний день набор данных для оценки задач понимания, рассуждения и локализации на длинных мультимодальных документах. Набор включает 396 документов со средней длиной от 50 до 150 страниц, охватывая более 33 000 страниц в общей сложности. Это делает его значительно более масштабным по сравнению с MMLongBench-Doc и создает экстремальные условия для оценки производительности систем извлечения информации.

Ключевое отличие LongDocURL заключается в трехуровневой таксономии задач:

1) *понимание* (*Understanding*) – базовое извлечение фактов из документа;
2) *рассуждение* (*Reasoning*) – многошаговые логические выводы на основе распределенной информации;
3) *локализация* (*Locating*) – точное определение местоположения информации в документе, включая кросс-элементную локализацию (например, связь между таблицей и соответствующим графиком)

Набор содержит 2 325 высококачественных вопрос-ответных пар, сгенерированных через полуавтоматический конвейер с последующей многоэтапной верификацией. Вопросы распределены по типам источников информации: *чистый текст*, *структурный макет*, *диаграммы и изображения*, *таблицы*. Особое внимание уделено созданию сложных кросс-элементных задач, которые составляют 37.1% всего набора данных. Особое внимание уделено кросс-элементным задачам (37.1% от общего числа), где ответ требует интеграции информации как минимум из двух разных типов элементов (например, таблица + диаграмма или текст + изображение), что создает дополнительные вызовы для мультимодальных RAG-систем.

4.3 Метрики оценки

Следует отметить, что наборы данных MMLongBench-Doc [45] и LongDocURL [57] напрямую не относится к оценке качества RAG-систем. Поэтому была предложена собственная система оценки, основанная на хорошо известных стандартных метриках из области RAG. Так, в качестве основной метрики оценки этапа извлечения/поиска (*retrieval*) использовалась: точное совпадение (*exact match*) и полнота (*recall*).

Вначале, формализуем набор данных MMLongBench-Doc:

$$Q = \{q_1, \dots, q_{1082}\}, q_i = \langle a_i^{gt}, P_i^{gt}, T_i^{gt} \rangle, i = \overline{1{,}1082},$$

где Q – множество вопросов набора данных MMLongBench-Doc; q_i – i-вопрос, сформулированный на естественном языке; a_i^{gt} – корректный (истинный) ответ на i-вопрос (*ground truth*); P_i^{gt} – множество страниц исходного документа, на которых содержится правильный ответ a_i^{gt}, при этом: $P_i^{gt} = \{p_1^{gt}, \dots, p_n^{gt}\}$, где p_j^{gt} – j-страница, содержащая ответ; T_i^{gt} – множество типов фрагментов страницы (текст, обобщенная структура, диаграмма, таблица, изображение), где содержится правильный ответ a_i^{gt}, при этом $T_i^{gt} = \{t_1^{gt}, \dots, t_n^{gt}\}$, где t_k^{gt} – k-тип фрагмента страницы, содержащий ответ.

Формальное представление результата этапа извлечения (поиска):

$$R = \{r_1, \dots, r_k\}, r_i = \langle c_i, p_j^{rt}, t_n^{rt} \rangle, i = \overline{1, k},$$

где R – множество результатов извлечения (топ-k); r_i – i-фрагмент-регион; c_i – содержимое i-фрагмента-региона; p_j^{rt} – j-страница документа, содержащая извлеченный i-фрагмент-регион; t_n^{rt} – n-тип извлеченного i-фрагмента-региона.

Таким образом, точное совпадение является строгой метрикой и формализуется следующим образом:

$$exact\ match = \begin{cases} 1, if\ \forall r_i \in R: p_j^{rt} \in P_i^{gt}\ \wedge\ t_n^{rt} \in T_i^{gt} \\ otherwise\ 0 \end{cases},$$

точное совпадение подразумевает, что среди множества извлеченных топ-k фрагментов-регионов существует такое подмножество, которое полностью пересекается с подмножеством истинных (релевантных) страниц P_i^{gt} и соответствующими типами T_i^{gt} (в этом случае считается 1, и 0 в противном случае).

Формальное описание полноты по страницам:

$$recall_{page} = \frac{P}{PN},$$

где P – количества корректных (истинных) страниц с найденными релевантными фрагментами-регионами; PN – общее количество корректных (истинных) страниц для данного вопроса.

Формальное описание полноты по типам:

$$recall_{type} = \frac{T}{TN},$$

где T – количества корректных (истинных) типов, относящихся к найденным релевантным фрагментам-регионам; TN – общее количество корректных (истинных) типов фрагментов страниц для данного вопроса.

Дополнительными метриками оценки качества этапа извлечения/поиска (*retrieval*) являлись: доля нахождений релевантной информации (*Hit Rate*) и характеристика эффективности информационного поиска, зависящая от порядкового положения (ранга) первого релевантного результата (*MRR*).

Формальное описание доли нахождений релевантной информации:

$$Hit\ Rate = \frac{RN}{QN},$$

где RN – количество вопросов, на которые получен хотя бы один корректный фрагмент-регион (извлеченный фрагмент-регион принадлежит правильной странице и соответствует истинному типу); QN – общее количество вопросов.

Формальное описание характеристики эффективности информационного поиска, зависящего от порядкового положения (ранга) первого релевантного результата:

$$MRR = \frac{1}{QN} \sum_{i=1}^{QN} \frac{1}{rank_i}, i = \overline{1{,}1082},$$

где QN – общее количество вопросов; $rank_i$ – порядковое положение (ранг) первого релевантного результата извлечения (фрагмента-региона) в топ-k для i-вопроса.

Для оценки качества генерации итогового ответа моделью используются стандартные метрики точности (*Accuracy*) и F-меры (*F1 score*), предложенные в рамках набора данных MMLongBench-Doc и LongDocURL. При этом они рассчитываются комплексно, в зависимости от типа данных (строка, целое число, число с плавающей точкой, список).

4.3 Конфигурация текстовых инструкций

Для обеспечения согласованной работы компонентов извлечения и генерации, а также минимизации вероятности галлюцинаций, в предлагаемом подходе применяются три специализированных текстовых подсказок-инструкций (*prompts*). Их структура разработана с учетом протоколов оценки выбранных целевых наборов данных и требований к форматированию вывода.

Инструкция 1: Извлечение итогового ответа для MMLongBench-Doc. Данная текстовая инструкция настраивает модель форматировать ответ строго в одном из четырех типов (*Integer*, *Float*, *String* или *List*). При отсутствии доказательной базы в контексте предписывается возврат маркера «*Not answerable*», а при невозможности визуального распознавания – «*Fail to answer*». Эта инструкция также включает *few-shot* примеры для стабилизации вывода.

Инструкция 2: Извлечение итогового ответа для LongDocURL. Данная текстовая инструкция расширяет логику предыдущей инструкции механизмом верификации видимости. Модель обязана фильтровать информацию, извлеченную из изображений или страниц, не входящих в заданный диапазон видимости вопроса (например, «*между 7-м и 20-м изображением*»). Ответ оборачивается в специальные токены *<concise_answer>* и *<answer_format>*, что упрощает автоматический разбор результатов.

Инструкция 3: Дополнительная инструкция генерации ответа. Данная текстовая инструкция задает роль интеллектуального ассистента с жестким ограничением, где ответ должен формироваться исключительно на основе предоставленного контекста (Рис. 5). Эта инструкция также включает правила цитирования фрагментов, указания на противоречия в источниках и сохранения языка вопроса. При отсутствии информации предписывается прямой ответ: «*В предоставленном контексте нет информации для ответа на этот вопрос*». Данный тип инструкции использовался отдельно в экспериментах для корректировки генерации финального ответа.

Вы — интеллектуальный ассистент для анализа документов. Ваша задача — отвечать на вопросы пользователя, основываясь ИСКЛЮЧИТЕЛЬНО на предоставленном контексте (*текст и изображения*).

ПРАВИЛА ОТВЕТА:
1. Используйте ТОЛЬКО информацию из предоставленного контекста
2. Если ответ не найден в контексте, честно сообщите: «В *предоставленном контексте нет информации для ответа на этот вопрос*»
3. Для изображений, таблиц и диаграмм внимательно анализируйте визуальную информацию вместе с подписями
4. Цитируйте конкретные фрагменты контекста при формулировке ответа
5. Будьте точны и лаконичны
6. Если контекст содержит противоречивую информацию, укажите на это
7. Сохраняйте язык ответа таким же, как язык вопроса

Рис. 5. Дополнительная текстовая инструкция-подсказка для генерации ответа.
Fig. 5. An additional prompt for generating the answer.

4.4 Основные результаты

В Табл. 1 представлены результаты производительности предлагаемого подхода с использованием метода разбиения документа на фрагменты по регионам и структурного графа в сравнении с классическим подходом RAG на этапе извлечения/поиска (*retrieval*) 10 и 20 кандидатных фрагментов.

Табл. 1. Результаты экспериментальной оценки этапа извлечения/поиска (retrieval) для предлагаемого подхода в сравнении с классическим RAG. При этом (*) – тестируется только предлагаемый метод разбиения документа на фрагменты по регионам; (+) – тестируется предлагаемый подход в конфигурации: метод разбиения документа на фрагменты по регионам + структурный граф.
Table 1. Results of the experimental evaluation of the retrieval stage for the proposed approach compared to classical RAG approach. (*) indicates testing of only the proposed region-based chunking method, while (+) indicates testing of the proposed approach in the following configuration: region-based chunking method + structural graph.

Подход	Топ-*k*	EM	Recall-*p*	Recall-*t*	Hit Rate	MRR
MMLongBench-Doc						
Классический RAG	10	0.105	0.229	0.184	0.188	0.216
*Предлагаемый подход**		0.183	0.309	0.302	0.367	0.328
Классический RAG	20	0.125	0.259	0.197	0.201	0.238
*Предлагаемый подход**		0.241	0.392	0.374	0.448	0.355
Предлагаемый подход+		**0.348**	**0.510**	**0.515**	**0.589**	**0.512**
LongDocURL						
Классический RAG	20	0.154	0.356	0.293	0.398	0.342
*Предлагаемый подход**		0.375	0.581	0.611	0.691	0.590

Предложенный метод разбиения документа на фрагменты по регионам демонстрирует статистически значимое превосходство над классическим «наивным» методом RAG по всем ключевым метрикам этапа извлечения/поиска (*retrieval*). При формировании топ-10

кандидатов наблюдается прирост метрики точного совпадения (*Exact Match*) на 74.3%, а при топ-20 более чем двукратное улучшение (с 12.5% до 34.8%). Аналогичная динамика фиксируется для метрик полноты по страницам (*Recall-p*) и типам (*Recall-t*), что свидетельствует о способности регионно-ориентированного разбиения документа на фрагменты более точно локализовывать релевантные текстовые блоки в сложно-структурированных документах.

Особого внимания заслуживают результаты на более сложном наборе данных LongDocURL, характеризующемся экстремальной длиной документов (до 150 страниц). Предложенный метод обеспечивает значение Hit Rate на уровне 69.1%, что указывает на высокую устойчивость подхода к распределению информации на больших отдалениях внутри документа. Значения MRR подтверждают, что релевантные регионы ранжируются ближе к началу списка кандидатов, что критически важно для снижения нагрузки на последующие этапы конвейера и минимизации шума в контексте генеративной модели.

Таким образом, экспериментально подтверждено, что учет сложной многомерной структуры документа на этапе сегментации позволяет преодолеть ограничения линейного разбиения документа на фрагменты и существенно повысить качество первичного отбора контекста для мультимодальных запросов.

В Табл. 2 представлены результаты оценки качества предлагаемого подхода в разной конфигурации на этапе генерации итогового ответа (*generation*), используя топ-20 извлечений с предыдущего этапа. Для объективной оценки конкурентоспособности разработанного подхода было также проведено сопоставление с современными решениями (*State-Of-The-Art – SOTA*), основанными на LVLM и представленными на используемых наборах данных. Результаты этого сравнения представлены в Табл. 3.

В настоящее время, на наборе данных MMLongBench-Doc передовые мультимодальные RAG-подходы с агентной маршрутизацией и мощными LVLM демонстрируют значения точности до 47.9%. На более сложном наборе данных LongDocURL, характеризующемся экстремальной длиной документов, современные подходы достигают метрик точности порядка 57-64%. Наш прототип показывает результат точности в 43% и 45.4% соответственно. Полученный результат является ожидаемым и методологически обоснованным, так как в рамках текущего эксперимента прототип системы использовал только компонент регионно-ориентированного разбиения документа на фрагменты, тогда как механизмы интеграции структурного и семантического графов, являющиеся ключевым инновационным вкладом работы, были отключены. Следовательно, генеративная модель получала расширенный контекст без явного моделирования иерархических и кросс-модальных зависимостей, что ограничивало потенциал подхода. В то же время, на наборе данных LongDocURL предложенный подход демонстрирует абсолютное значение точности в 45.4%, которое указывает на то, что даже частичное тестирование реализации архитектуры (сегментации) обеспечивает устойчивое преимущество в сценариях, требующих навигации по длинным мультимодальным документам. В итоге, полученные результаты позволяют сформулировать важную гипотезу – полное развертывание конвейера, включающее обход графа связности на этапе извлечения, должно обеспечить синергетический эффект и привести к статистически значимому приросту метрик генерации. Проверка данной гипотезы запланирована в рамках следующего этапа исследований.

Для всесторонней оценки эффективности предложенного подхода также был проведен статистический анализ производительности системы на наборе данных MMLongBench-Doc по трем ключевым аспектам, отражающим различные измерения сложности задачи понимания длинных мультимодальных документов. Такой многофакторный анализ позволяет выявить не только общую эффективность подхода, но и его сильные и слабые стороны в конкретных сценариях применения.

Табл. 2. Результаты экспериментальной оценки этапа генерации ответа (generation) для предлагаемого подхода в сравнении с классическим RAG. При этом (*) – тестируется только предлагаемый метод разбиения документа на фрагменты по регионам; (+) – тестируется предлагаемый подход в конфигурации: метод разбиения документа на фрагменты по регионам + структурный граф; (++) – тестируется предлагаемый подход в конфигурации: метод разбиения документа на фрагменты по регионам + структурный граф + дополнительная текстовая инструкция для генерации ответа.
Table 2. Results of the experimental evaluation of the answer generation stage for the proposed approach compared to classical RAG approach. (*) denotes testing of the proposed region-based chunking method alone; (+) denotes testing of the proposed approach in the configuration: region-based chunking method + structural graph; and (++) denotes testing of the proposed approach in the configuration: region-based chunking method + structural graph + additional answer generation prompt.

Подход	Accuracy	F1
MMLongBench-Doc		
Классический RAG	0.381	0.351
*Предлагаемый подход**	0.320	0.307
Предлагаемый подход+	0.365	0.369
Предлагаемый подход++	**0.430**	**0.385**
LongDocURL		
Классический RAG	**0.595**	–
*Предлагаемый подход**	0.454	–

Табл. 3. Результаты экспериментальной оценки предлагаемого подхода в конфигурации: метод разбиения документа на фрагменты по регионам + структурный граф + дополнительная текстовая инструкция, на этапе генерации ответа (generation) в сравнении с современными SOTA-решениями.
Table 3. Results of the experimental evaluation of the proposed approach configured as region-based chunking + structural graph + additional answer generation prompt at the answer generation stage, compared to state-of-the-art (SOTA) solutions.

Подход/Модель	Accuracy	F1
MMLongBench-Doc		
Claude-3 Opus	0.174	0.181
Gemini-1.5-Pro	0.282	0.206
MDocAgent (top-4) [17]	0.315	–
GPT-4V	0.324	0.312
CLIP + Qwen2.5-VL-32B	0.367	–
GPT-4o	0.428	**0.449**
MLDocRAG [58]	**0.479**	–
Предлагаемый подход	0.430	0.385
LongDocURL		
CLIP + Qwen2.5-VL-32B	0.345	–
Qwen-VL-Max	0.495	–
MLDocRAG [58]	0.508	–
Gemini-1.5-Pro	0.509	–
MDocAgent (top-4) [17]	0.578	–
GPT-4o	**0.645**	–
Предлагаемый подход	0.454	–

1) *категория вопроса по количеству источников информации* (Табл. 4) – оценивалась способность подхода к локализации релевантного контента на одной странице (одностраничные вопросы – «*Single-page*»), интеграции распределенной информации из нескольких страниц документа (многостраничные вопросы – «*Cross-page*») и корректному распознаванию отсутствия ответа в документе (вопросы без ответов – «*Unanswerable*»);
2) *тип источника информации* (Табл. 5) – в данном случае, дополнительно оценивалась эффективность интеграции различных модальностей (текст, структурный макет, таблицы, диаграммы/графики, изображения) в единую семантическую структуру;
3) *тип документа* (Табл. 6) – проверялась робастность метода к типам документов, обладающих разной структурой (научные отчеты, финансовые отчеты, академические статьи, брошюры, руководства, административные и промышленные документы, учебные материалы и воркшопы).

Для получения количественной оценки для всех трех аспектов использовалась метрика точности (*Accuracy*), адаптированная к формату ответа (строка, целое число, число с плавающей точкой, список) в соответствии с протоколом оценки набора данных MMLongBench-Doc.

Табл. 4. Результаты экспериментальной оценки точности (accuracy) этапа генерации ответа (generation) в зависимости от категории вопроса на наборе данных MMLongBench-Doc и в сравнении с классическим RAG. При этом (*) – тестируется только предлагаемый метод разбиения документа на фрагменты по регионам; (+) – тестируется предлагаемый подход в конфигурации: метод разбиения документа на фрагменты по регионам + структурный граф; (++) – тестируется предлагаемый подход в конфигурации: метод разбиения документа на фрагменты по регионам + структурный граф + дополнительная текстовая инструкция для генерации ответа.
Table 4. Results of the experimental evaluation (accuracy) for the answer generation stage, categorized by question type, using the MMLongBench-Doc dataset and compared against standard RAG. (*) denotes testing only the proposed region-based chunking method; (+) denotes testing the proposed approach with the configuration: region-based chunking + structural graph; (++) denotes testing the proposed approach with the configuration: region-based chunking + structural graph + additional answer generation prompt.

Подход	Single-page	Cross-page	Unanswerable
Классический RAG	0.380	0.293	0.518
*Предлагаемый подход**	0.396	**0.241**	0.287
Предлагаемый подход+	**0.474**	0.293	0.241
Предлагаемый подход++	0.414	0.237	**0.770**

Наиболее высокие показатели точности достигаются для одностраничных вопросов (*Single-page*), что подтверждает эффективность регионно-ориентированного разбиения документа на фрагменты для задач локальной навигации. Для многостраничных вопросов (*Cross-page*) наблюдается снижение метрики до 24.1%, что объясняется ограниченной реализацией общего конвейера. Так, без использования механизма обхода структурного и семантического графа связности система не может в полной мере агрегировать информацию, распределенную по различным разделам документа. Значение для вопросов без ответов указывает на умеренную способность модели корректно идентифицировать отсутствие релевантного контекста, что является важным аспектом борьбы с галлюцинациями.

Предложенный подход демонстрирует наибольшее преимущество для табличных данных (47.7% точности), что напрямую подтверждает гипотезу о целесообразности специализированной векторизации табличных структур. Сохранение двумерной топологии и семантики заголовков позволяет модели более точно интерпретировать запросы, требующие извлечения числовых значений или сравнения атрибутов. Наименьшие показатели зафиксированы для изображений и диаграмм, что отражает сохраняющуюся сложность

кросс-модального выравнивания и указывает на направление для дальнейшей оптимизации мультимодального кодировщика.

Табл. 5. Результаты экспериментальной оценки точности (accuracy) этапа генерации ответа (generation) в зависимости от того, в каком источнике информации он находится, на наборе данных MMLongBench-Doc и в сравнении с классическим RAG. При этом (ПП) – тестируется только предлагаемый метод разбиения документа на фрагменты по регионам; (ПП+) – тестируется предлагаемый подход в конфигурации: метод разбиения документа на фрагменты по регионам + структурный граф; (ПП++) – тестируется предлагаемый подход в конфигурации: метод разбиения документа на фрагменты по регионам + структурный граф + дополнительная текстовая инструкция для генерации ответа.*

Table 5. Results of the experimental evaluation (accuracy) for the answer generation stage based on the information source using the MMLongBench-Doc dataset and comparing the results with classical RAG approach. (PM) denotes testing only the proposed region-based chunking method; (PM+) denotes testing the proposed approach in the configuration: region-based chunking method + structural graph; (PM++) denotes testing the proposed approach in the configuration: region-based chunking method + structural graph + additional answer generation prompt.*

Источник информации	Классический RAG	ПП*	ПП+	ПП++
Диаграмма/график (*Chart*)	0.254	0.230	**0.378**	0.295
Таблица (*Table*)	**0.528**	0.477	0.493	0.456
Текст (*Pure-text*)	0.404	0.362	**0.415**	0.365
Структурированный макет (*Generalized-text*)	**0.328**	0.241	0.243	0.233
Изображение (*Figure*)	0.144	0.201	**0.316**	0.241

Табл. 6. Результаты экспериментальной оценки точности (accuracy) этапа генерации ответа (generation) в зависимости типа документа на наборе данных MMLongBench-Doc и в сравнении с классическим RAG. При этом (ПП) – тестируется только предлагаемый метод разбиения документа на фрагменты по регионам; (ПП+) – тестируется предлагаемый подход в конфигурации: метод разбиения документа на фрагменты по регионам + структурный граф; (ПП++) – тестируется предлагаемый подход в конфигурации: метод разбиения документа на фрагменты по регионам + структурный граф + дополнительная текстовая инструкция для генерации ответа.*

Table 6. Results of the experimental evaluation (accuracy) for the answer generation stage, categorized by document type using the MMLongBench-Doc dataset and compared against standard RAG. (PP) denotes testing only the proposed region-based chunking method; (PP+) denotes testing the proposed approach in the configuration: region-based chunking + structural graph; (PP++) denotes testing the proposed approach in the configuration: region-based chunking + structural graph + additional answer generation prompt.*

Тип документа	Классический RAG	ПП*	ПП+	ПП++
Научные отчеты (*Research report*)	0.358	0.319	0.348	**0.478**
Учебные материалы и воркшопы (*Tutorial/Workshop*)	0.126	0.307	0.323	**0.363**
Академические статьи (*Academic paper*)	0.464	0.317	0.414	**0.468**
Руководства (*Guidebook*)	**0.421**	0.275	0.377	0.378
Брошюры (*Brochure*)	0.320	0.251	0.307	**0.395**
Административные и промышленные документы (*Administration/Industry documents*)	**0.440**	0.338	0.343	0.428
Финансовые отчеты (*Financial report*)	**0.572**	0.445	0.429	0.431

Подход показывает стабильную производительность независимо от предметной области, однако наибольший прирост наблюдается для документов с высокой плотностью структурированных данных – это финансовые отчеты (44.5%) и административных документов (33.8%). Это подтверждает, что регионно-ориентированный способ разбиения

документа на фрагменты наиболее эффективен для документов с выраженной иерархической составляющей. Относительно более низкие значения для брошюр (25.1%) и руководств (27.5%) могут быть обусловлены неформализованной структурой данных и преобладанием свободного текста, где преимущества структурного подхода проявляются в меньшей степени.

В целом, можно констатировать, что предложенный подход обеспечивает избирательное, но статистически значимое улучшение качества извлечения и генерации, особенно в сценариях работы с табличными данными и структурированными отчетами. Полученные результаты подтверждают архитектурные решения, заложенные в основу предлагаемого решения, и обосновывают целесообразность полномасштабной реализации графового компонента для достижения максимального потенциала подхода.

5. Заключение

В данной работе представлен новый мультимодальный графовый подход RAG, решающий ключевые проблемы современных систем генерации с дополнительным поиском, а именно игнорирование пространственной структуры документов, проблему модального разрыва и отсутствие явного моделирования семантических связей между фрагментами документа. Предложенный подход интегрирует регионно-ориентированное разбиение документа на фрагменты, специализированный табличный кодировщик, а также объединенный граф, отражающий как иерархическую структуру документа, так и смысловую близость фрагментов. Реализация архитектуры в виде контейнеризованной микросервисной системы обеспечивает ее детерминированность, горизонтальную масштабируемость и готовность к промышленному внедрению.

Проведенная предварительная экспериментальная оценка на эталонных наборах данных MMLongBench-Doc и LongDocURL продемонстрировала статистически значимое преимущество разработанного метода разбиения документа на фрагменты по регионам над классическим «наивным» методом RAG. Прирост всех метрик этапа извлечения (*EM, Recall, Hit Rate, MRR*) подтвердил эффективность учета двумерной визуальной структуры, а многофакторный анализ генерации выявил наибольшую точность при работе с табличными данными и финансовыми отчетами. Отмечено, что текущие показатели этапа генерации ограничены частичной реализацией графового компонента, что оставляет значительный резерв для дальнейшего улучшения качества ответов.

В рамках будущих исследований планируется полномасштабная интеграция алгоритма обхода семантического и структурного графов на этапе поиска, что позволит существенно повысить релевантность извлечения в сценариях с многостраничными и кросс-модальными запросами. Результаты работы подтверждают перспективность графово-ориентированного моделирования документов и создают прочную основу для следующего поколения интеллектуальных систем анализа сложно-структурированной информации.

Список литературы

[1]. Chang Y., Wang X., Wang J., Wu Y., Yang L., Zhu K., Chen H., Yi X., Wang C., Wang Y., Ye W., Zhang Y., Chang Y., Yu P.S., Yang Q., Xie X. A Survey on Evaluation of Large Language Models. ACM Transactions on Intelligent Systems and Technology, vol. 15, no. 3, 2024, pp. 1-45. DOI: 10.1145/3641289.

[2]. Hadi M.U., Tashi Q.A., Qureshi R., Shah A., Muneer A., Irfan M., Zafar A., Shaikh M.B., Akhtar N., Hassan S.Z., Shoman M., Wu J., Mirjalili S., Shah M. A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage. TechRxiv, 2025, pp. 1-56. DOI: 10.36227/techrxiv.23589741.v8.

[3]. Liu Y.-Y., Zheng Z., Zhang F., Feng J.-C., Fu Y.-Y., Zhai J.-D., He B.-S., Zhang X., Du X.-Y. A comprehensive taxonomy of prompt engineering techniques for large language models. Frontiers of Computer Science, vol. 20, 2026, 2003601. DOI: 10.1007/s11704-025-50058-z.

[4]. Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W., Rocktäschel T., Riedel S., Kiela D. Retrieval-augmented generation for knowledge-intensive NLP tasks. Proc. the 34th International Conference on Neural Information Processing Systems (NIPS'20), Vancouver, Canada, no. 793, 2020, pp. 9459-9474. DOI: 10.5555/3495724.3496517.

[5]. Borgeaud S., Mensch A., Hoffmann J., Cai T., Rutherford E., Millican K., Driessche G., Lespiau J.-B., Damoc B., Clark A., Casas D.L., Guy A., Menick J., Ring R., Hennigan T., Huang S., Maggiore L., Jones C., Cassirer A., Brock A., Paganini M., Irving G., Vinyals O., Osindero S., Simonyan K., Rae J.W., Elsen E., Sifre L. Improving language models by retrieving from trillions of tokens. Proc. the International Conference on Machine Learning (ICML), Online, 2021, pp. 1-43.

[6]. Fan W., Ding Y., Ning L., Wang S., Li H., Yin D., Chua T.-S., Li Q. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. Proc. the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'24), Barcelona, Spain, 2024, pp. 6491-6501. DOI: 10.1145/3637528.3671470.

[7]. Wu S., Xiong Y., Cui Y., Wu H., Chen C., Yuan Y., Huang L., Liu X., Kuo T.-W., Guan N., Xue J.C. Retrieval-Augmented Generation for Natural Language Processing: A Survey. SSRN, 2025, pp. 1-60. DOI: 10.2139/ssrn.5163979.

[8]. Liu H., Li C., Wu Q., Lee Y.J. Visual Instruction Tuning. Advances in Neural Information Processing Systems (NeurIPS 2023), vol. 36, 2023, pp. 34892-34916.

[9]. Li J., Li D., Savarese S., Hoi S. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. Proc. the 40th International Conference on Machine Learning (ICML'23), Honolulu, USA, vol. 814, 2023, pp. 19730-19742. DOI: 10.5555/3618408.3619222.

[10]. Zhang D., Yu Y., Dong J., Li C., Su D., Chu C., Yu D. MM-LLMs: Recent Advances in MultiModal Large Language Models. Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 2024, pp. 12401-12430. DOI: 10.18653/v1/2024.findings-acl.738.

[11]. Abootorabi M.M., Zobeiri A., Dehghani M., Mohammadkhani M., Mohammadi B., Ghahroodi O., Baghshah M.S., Asgari E. Ask in Any Modality: A Comprehensive Survey on Multimodal Retrieval-Augmented Generation. Findings of the Association for Computational Linguistics: ACL 2025, Vienna, Austria, 2025, pp. 16776-16809. DOI: 10.18653/v1/2025.findings-acl.861.

[12]. Yeo W., Kim K., Jeong S., Baek J., Hwang S.J. UniversalRAG: Retrieval-Augmented Generation over Corpora of Diverse Modalities and Granularities. 2025, pp. 1-28. DOI: 10.48550/arXiv.2504.20734.

[13]. Sharifymoghaddam S., Upadhyay S., Chen W., Lin J. UniRAG: Universal Retrieval Augmentation for Large Vision Language Models. Proc. the Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, 2025, pp. 2026-2039. DOI: 10.18653/v1/2025.findings-naacl.108.

[14]. Hu Z., Iscen A., Sun C., Wang Z., Chang K.-W., Sun Y., Schmid C., Ross D.A., Fathi A. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. Proc. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, Canada, 2023, pp. 23369-23379. DOI: 10.1109/CVPR52729.2023.02238.

[15]. Chen W., Hu H., Chen X., Verga P., Cohen W. MuRAG: Multimodal retrievalaugmented generator for open question answering over images and text. Proc. the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), Abu Dhabi, UAE, 2022, pp. 5558-5570. DOI: 10.18653/v1/2022.emnlp-main.375.

[16]. Faysse M., Sibille H., Wu T., Omrani B., Viaud G., Hudelot C., Colombo P. ColPali: Efficient Document Retrieval with Vision Language Models. Proc. the 2025 International Conference on Learning Representations (ICLR), Singapore, 2025, pp. 1-26.

[17]. Han S., Xia P., Zhang R., Sun T., Li Y., Zhu H., Yao H. MDocAgent: A Multi-Modal Multi-Agent Framework for Document Understanding. 2025, pp. 1-16. Available at: https://arxiv.org/abs/2503.13964, accessed 03.04.2026.

[18]. Liu P., Liu X., Yao R., Liu J., Meng S., Wang D., Ma J. HM-RAG: Hierarchical Multi-Agent Multimodal Retrieval Augmented Generation. Proc. the 33rd ACM International Conference on Multimedia (MM'25), Dublin, Ireland, 2025, pp. 2781-2790. DOI: 10.1145/3746027.3754761.

[19]. Tripathi V., Odapally T., Das I., Allu U., Ahmed B. Vision-Guided Chunking Is All You Need: Enhancing RAG with Multimodal Document Understanding. 2025, pp. 1-21. Available at: https://arxiv.org/abs/2506.16035, accessed 03.04.2026.

[20]. Chen J., Zhang R., Zhou Y., Yu T., Dernoncourt F., Gu J., Rossi R.A., Chen C., Sun T. SV-RAG: LoRA-Contextualizing Adaptation of MLLMs for Long Document Understanding. Proc. the 2025 International Conference on Learning Representations (ICLR), Singapore, 2025, pp. 1-16.

[21]. Tanaka R., Iki T., Hasegawa T., Nishida K., Saito K., Suzuki J. VDocRAG: Retrieval-Augmented Generation over Visually-Rich Documents. Proc. the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, USA, 2025, pp. 24827-24837. DOI: 10.1109/CVPR52734.2025.02312.

[22]. Ali Z., Huang Y., Khan A., Qi G., Zhang Y., Feng J., Deng C., Kefalas P. Pythia-RAG: Retrieval-augmented generation over a unified multimodal knowledge graph for enhanced QA. Knowledge-Based Systems, vol. 336, 2026, 115200. DOI: 10.1016/j.knosys.2025.115200.

[23]. Karpukhin V., Oguz B., Min S., Lewis P., Wu L., Edunov S., Chen D., Yih W. Dense Passage Retrieval for Open-Domain Question Answering. Proc. the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 2020, pp. 6769-6781. DOI: 10.18653/v1/2020.emnlp-main.550.

[24]. Zhong Z., Lei T., Chen D. Training Language Models with Memory Augmentation. Proc. the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), Abu Dhabi, UAE, 2022, pp. 5657-5673. DOI: 10.18653/v1/2022.emnlp-main.382.

[25]. Zhou S., Alon U., Xu F.F., Wang Z., Jiang Z., Neubig G. DocPrompting: Generating Code by Retrieving the Docs. Proc. the International Conference on Learning Representations (ICLR'2023), Kigali, Rwanda, 2023, pp. 1-19.

[26]. Ren Y., Cao Y., Guo P., Fang F., Ma W., Lin Z. Retrieve-and-Sample: Document-level Event Argument Extraction via Hybrid Retrieval Augmentation. Proc. the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Toronto, Canada, vol. 1, 2023, pp. 293-306. DOI: 10.18653/v1/2023.acl-long.17.

[27]. Lee M., An S., Kim M.-S. PlanRAG: A Plan-then-Retrieval Augmented Generation for Generative Large Language Models as Decision Makers. Proc. the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), Mexico City, Mexico, 2024, pp. 6537-6555. DOI: 10.18653/v1/2024.naacl-long.364.

[28]. An Z., Ding X., Fu Y.-C., Chu C.-C., Li Y., Du W. Golden-Retriever: High-Fidelity Agentic Retrieval Augmented Generation for Industrial Knowledge Base. 2024, pp. 1-11. Available at: https://arxiv.org/abs/2408.00798, accessed 03.04.2026.

[29]. Liu Y., Peng X., Zhang X., Liu W., Yin J., Cao J., Du T. RA-ISF: Learning to Answer and Understand from Retrieval Augmentation via Iterative Self-Feedback. Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 2024, pp. 4730-4749. DOI: 10.18653/v1/2024.findings-acl.281.

[30]. text-embedding-3-small. Available at: https://huggingface.co/collections/tsystems/colqwen2-multimodal-embedding, accessed 03.04.2026.

[31]. ColQwen2 Multimodal Embedding. Available at: https://askwonder.com/research/number-google-sheets-users-worldwide-eoskdoxav, accessed 03.04.2026.

[32]. Peng X., Qin C., Chen Z., Xu R., Xiong C., Wu C.-S. UNIDOC-BENCH: A Unified Benchmark for Document-Centric Multimodal RAG. 2026, pp. 1-39. Available at: https://arxiv.org/abs/2510.03663, accessed 03.04.2026.

[33]. Zhang X., Zhang Y., Xie W., Li M., Dai Z., Long D., Xie P., Zhang M., Li W., Zhang M. Bridging Modalities: Improving Universal Multimodal Retrieval by Multimodal Large Language Models. Proc. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, USA, 2025, pp. 9274-9285. DOI: 10.1109/CVPR52734.2025.00866.

[34]. Meng R., Jiang Z., Liu Y., Su M., Yang X., Fu Y., Qin C., Chen Z., Xu R., Xiong C., Zhou Y., Chen W., Yavuz S. VLM2Vec-V2: Advancing Multimodal Embedding for Videos, Images, and Visual Documents. 2025, pp. 1-18. Available at: https://arxiv.org/abs/2507.04590, accessed 03.04.2026.

[35]. Jia Y., Wu X., Hao L., Zhang Q., Hu Y., Zhao S., Fan W. Uni-Retrieval: A Multi-Style Retrieval Framework for STEM's Education. Proc. the 63rd Annual Meeting of the Association for Computational Linguistics (ACL), Vienna, Austria, 2025, pp. 10182-10197. DOI: 10.18653/v1/2025.acl-long.502.

[36]. Chen B., Khare A., Kumar G., Akula A., Narayana P. Seeing Beyond: Enhancing Visual Question Answering with Multi-Modal Retrieval. Proc. the 31st International Conference on Computational Linguistics: Industry Track (COLING), Abu Dhabi, UAE, 2025, pp. 410-421.

[37]. Jiang Z., Xu F.F., Gao L., Sun Z., Liu Q., Dwivedi-Yu H., Yang Y., Callan J., Neubig G. Active Retrieval Augmented Generation. Proc. the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), Singapore, 2023, pp. 7969-7992. DOI: 10.18653/v1/2023.emnlp-main.495.

[38]. Asai A., Wu Z., Wang Y., Sil A., Hajishirz H. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. Proc. the International Conference on Learning Representations (ICLR), Vienna, Austria, 2024, pp. 1-30.

[39]. Cheng X., Wang X., Zhang X., Ge T., Si-Qing Chen S.-Q., Wei F., Zhang H., Zhao D. xRAG: Extreme Context Compression for Retrieval-augmented Generation with One Token. Proc. the 38th Conference on Neural Information Processing Systems (NeurIPS'24), Vancouver, Canada, 2024, pp. 109487-109516. DOI: 10.5555/3737916.3741392.

[40]. Li S., Stenzel L., Eickhoff C., Bahrainian S.A. Enhancing Retrieval-Augmented Generation: A Study of Best Practices. Proc. the 31st International Conference on Computational Linguistics (COLING), Abu Dhabi, UAE, 2025, pp. 6705-6717.

[41]. Perlić V., Lebailly S., Malvone V., Nguyen V.-T., Urard P. MMA-RAG: A Survey on Multimodal Agentic Retrieval-Augmented Generation. SSRN, 2025, pp. 1-24. DOI: 10.2139/ssrn.5630330.

[42]. Peng B., Zhu Y., Liu Y., Bo X., Shi H., Hong C., Zhang Y., Tang S. Graph Retrieval-Augmented Generation: A Survey. ACM Transactions on Information Systems, vol. 44, no. 2, 2025, pp. 1-52. DOI: 10.1145/3777378.

[43]. Xu Y., Li M., Cui L., Huang S., Wei F., Zhou M. LayoutLM: Pre-training of Text and Layout for Document Image Understanding. Proc. the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'20), New York, USA, 2020, pp. 1192-1200. DOI: 10.1145/3394486.3403172.

[44]. Wang D., Raman N., Sibue M., Ma Z., Babkin P., Kaur S., Pei Y., Nourbakhsh A., Liu X. DocLLM: A Layout-Aware Generative Language Model for Multimodal Document Understanding. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Bangkok, Thailand, 2024, pp. 8529-8548. DOI: 10.18653/v1/2024.acl-long.463.

[45]. Ma Y., Zang Y., Chen L., Chen M., Jiao Y., Li X., Lu X., Liu Z., Ma Y., Dong X., Zhang P., Pan L., Jiang Y.-G., Wang J., Cao Y., Sun A. MMLONGBENCH-DOC: benchmarking long-context document understanding with visualizations. Proc. the 38th International Conference on Neural Information Processing Systems (NIPS'24), Vancouver, Canada, 2024, pp. 95963-96010. DOI: 10.5555/3737916.3740957.

[46]. ScienceQA. Available at: https://scienceqa.github.io/leaderboard.html, accessed 03.04.2026.

[47]. Zheng M., Feng X., Si Q., She Q., Lin Z., Jiang W., Wang W. Multimodal Table Understanding. Proc. the 62nd Annual Meeting of the Association for Computational Linguistics (ACL'2024), Bangkok, Thailand, 2024, pp. 9102-9124. DOI: 10.18653/v1/2024.acl-long.493.

[48]. Hu H., Wang X., Zhang Y., Chen Q., Guan Q. A comprehensive survey on contrastive learning. Neurocomputing, vol. 610, 2024, 128645. DOI: 10.1016/j.neucom.2024.128645.

[49]. Zhang T., Yue X., Li Y., Sun H. TableLlama: Towards Open Large Generalist Models for Tables. Proc. the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Mexico City, Mexico, 2024, pp. 6024-6044. DOI: 10.18653/v1/2024.naacl-long.335.

[50]. MinerU. Available at: https://github.com/opendatalab/mineru, accessed 03.04.2026.

[51]. MinIO. Available at: https://www.min.io/, accessed 03.04.2026.

[52]. Qwen3-VL-Embedding-2B. Available at: https://huggingface.co/Qwen/Qwen3-VL-Embedding-2B, accessed 03.04.2026.

[53]. Qdrant - Vector Search Engine. Available at: https://qdrant.tech, accessed 03.04.2026.

[54]. Edge D., Trinh H., Cheng N., Bradley J., Chao A., Mody A., Truitt S., Metropolitansky D., Ness R.O. From Local to Global: A GraphRAG Approach to Query-Focused Summarization. 2025, pp. 1-26. Available at: https://arxiv.org/abs/2404.16130, accessed 03.04.2026.

[55]. Hogan A., Blomqvist E., Cochez M., d'Amato C., De Melo G., Gutierrez C., Gayo J. E. L., Kirrane S., Neumaier S., Polleres A., Navigli R., Ngomo A.-C. N., Rashid S. M., Rula A., Schmelzeisen L., Sequeda J., Staab S., Zimmermann A. Knowledge Graphs. Springer Nature Switzerland, 2021, 237 p. DOI: 10.1007/978-3-031-01918-0.

[56]. Qwen3-VL-32B-Instruct. Available at: https://huggingface.co/Qwen/Qwen3-VL-32B-Instruct, accessed 03.04.2026.

[57]. Deng C., Yuan J., Bu P., Wang P., Li Z.-Z., Xu J., Li X.-H., Gao Y., Song J., Zheng B., Liu C.-L. LongDocURL: a Comprehensive Multimodal Long Document Benchmark Integrating Understanding, Reasoning, and Locating. Proc. the 63rd Annual Meeting of the Association for Computational Linguistics (ACL), Vienna, Austria, 2025, pp. 1135-1159. DOI: 10.18653/v1/2025.acl-long.57.

[58]. Zhang Y., Wu Y. MLDocRAG: Multimodal Long-Context Document Retrieval Augmented Generation. 2026, pp. 1-15. Available at: https://arxiv.org/abs/2602.10271, accessed 07.07.2026.

Информация об авторах / Information about authors

Дмитрий Евгеньевич КОСОГОРОВ – аспирант Института динамики систем и теории управления им. В.М. Матросова Сибирского отделения РАН (ИДСТУ СО РАН) с 2025 года и младший научный сотрудник ИДСТУ СО РАН с 2026 года. Сфера научных интересов: большие языковые модели, генерация с дополнительным поиском.

Dmitry Evgenievich KOSOGOROV is a postgraduate student at the Matrosov Institute of System Dynamics and Control Theory named SB RAS (ISDCT SB RAS) since 2025 and junior researcher at ISDCT SB RAS since 2026. Research interests: large language models, and retrieval-augmented generation.

Даниил Евгеньевич КОПЫЛОВ – аспирант Иркутского государственного университета и младший научный сотрудник ИДСТУ СО РАН с 2026 года. Сфера научных интересов: прикладная математика, анализ данных.

Daniil Evgenievich KOPYLOV is a postgraduate student of Irkutsk State University and junior researcher at ISDCT SB RAS since 2026. Research interests: applied mathematics, and data analysis.

Илья Игоревич ОХОТИН – аспирант ИДСТУ СО РАН и младший научный сотрудник ИДСТУ СО РАН с 2026 года. Сфера научных интересов: большие языковые модели, работа с табличными данными, распознавание структуры таблиц; обработка заголовков таблиц.

Ilya Igorevich OKHOTIN id a postgraduate student and junior researcher at ISDCT SB RAS since 2026. Research interests: large language models, tabular data processing, table structure recognition, and table header processing.

Иван Александрович ПОДДУБНЫЙ – аспирант ИДСТУ СО РАН с 2024 года и младший научный сотрудник ИДСТУ СО РАН с 2026 года. Сфера научных интересов: табличные рассуждения, ответы на вопросы по табличным данным, верификация фактов в таблицах, табличные представления, большие языковые модели, глубокое обучение.

Ivan Aleksandrovich PODDUBNYY is a postgraduate student at ISDCT SB RAS since 2024 and junior researcher at ISDCT SB RAS since 2026. Research interests: table reasoning, table question answering, table fact verification, table representation, large language models, and deep learning.

Кирилл Владимирович ТОБОЛА – аспирант ИДСТУ СО РАН с 2024 года и младший научный сотрудник ИДСТУ СО РАН с 2026 года. Сфера научных интересов: табличные представления, большие языковые модели для реляционных данных, извлечение данных из табличных источников.

Kirill Vladimirovich TOBOLA is a postgraduate student at ISDCT SB RAS since 2024 and junior researcher at ISDCT SB RAS since 2026. Research interests: table embedding, large language models for relational data, and data extraction from tabular sources.

Никита Олегович ДОРОДНЫХ – кандидат технических наук ИДСТУ СО РАН с 2018 года, и старший научный сотрудник ИДСТУ СО РАН с 2021 года. Сфера научных интересов: автоматизация создания интеллектуальных систем и баз знаний, получение знаний на основе преобразования концептуальных моделей и электронных таблиц, графы знаний и семантические технологии, автоматическое понимание табличной информации, табличные рассуждения, большие языковые модели.

Nikita Olegovych DORODNYKH – Cand. Sci. (Tech.) at ISDCT SB RAS since 2018 and senior associate researcher at ISDCT SB RAS since 2021. Research interests: computer-aided development of intelligent systems and knowledge bases, knowledge acquisition based on the transformation of

conceptual models and tables, knowledge graphs and semantic web, table understanding, table reasoning, and large language models.

Андрей Анатольевич МИХАЙЛОВ – кандидат технических наук ИДСТУ СО РАН с 2017 года, заведующий молодежной лабораторией ИДСТУ СО РАН с 2024 года. Сфера научных интересов: методы и технологии обучения представлений и извлечения информативных признаков на основе самообучения без ручной разметки данных, применяемые для задач автоматического понимания и анализа электронных документов.

Andrey Anatolyevich MIKHAILOV – Cand. Sci. (Tech.) at ISDCT SB RAS since 2017, Head of the Youth Laboratory at ISDCT SB RAS since 2024. Research interests: methods and technologies for representation learning and informative feature extracting based on self-supervised learning without manual data labeling, applied to the tasks of automatic understanding and analysis of electronic documents.