Артамонов Артур Алексеевич, магистрант,
Санкт-Петербургский государственный электротехнический
Университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
Artamonov Artur Alekseevich, Master's student,
Saint Petersburg State Electrotechnical University
"LETI" named after V.I. Ulyanov (Lenin)

Белов Александр Михайлович, магистрант,
Санкт-Петербургский государственный электротехнический
Университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
Belov Alexander Mikhailovich, Master's student,
Saint Petersburg State Electrotechnical University
"LETI" named after V.I. Ulyanov (Lenin)

Дашкин Дамир Шах-азамович, магистрант,
Санкт-Петербургский государственный электротехнический
Университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
Dashkin Damir Shakh-azamovich, Master's student,
Saint Petersburg State Electrotechnical University
"LETI" named after V.I. Ulyanov (Lenin)

ПОДХОДЫ К ПОСТРОЕНИЮ VIDEO-LANGUAGE MODELS: АРХИТЕКТУРНЫЕ ПРИНЦИПЫ И АНАЛИЗ СОБЫТИЙ
APPROACHES TO BUILDING VIDEO-LANGUAGE MODELS: ARCHITECTURAL PRINCIPLES AND EVENT ANALYSIS

Аннотация. Video-Language Models (VLM) – класс мультимодальных архитектур, объединяющих компьютерное зрение и большие языковые модели для семантического анализа видео. В статье изложены базовые принципы построения VLM: трёхкомпонентная структура (визуальный энкодер, проектор, языковая модель), методы учёта временной динамики, подходы к обработке длинных видеопоследовательностей и стратегия многостадийного обучения. Рассмотрены причины вычислительных трудностей и архитектурные решения для их преодоления.

Abstract. Video-Language Models (VLMs) are a class of multimodal architectures that integrate computer vision and large language models for the semantic analysis of video data. The paper outlines the fundamental principles of VLM design, including a three-component structure (visual encoder, projector, and language model), methods for capturing temporal dynamics, approaches to processing long video sequences, and a multi-stage training strategy. It also discusses the sources of computational challenges and architectural solutions proposed to address them.

Ключевые слова: Video-Language Models, мультимодальные модели, анализ видео, большие языковые модели, компьютерное зрение.

Keywords: Video-Language Models, multimodal models, video understanding, large language models, computer vision.

Введение. Объём цифрового видео растёт экспоненциально, и для его осмысления требуются средства, выходящие за рамки распознавания отдельных объектов или действий. Video-Language Models позволяют взаимодействовать с видеоконтентом на естественном языке: задавать вопросы о происходящем, искать события по описанию и получать связные объяснения динамики сцены [1]. В основе таких систем лежит синтез визуального восприятия

и лингвистического рассуждения, сопряжённый с необходимостью моделировать временные связи. Цель статьи – объяснить ключевые архитектурные решения, делающие этот синтез возможным.

Основные компоненты архитектуры. Современная VLM состоит из трёх функциональных блоков: визуального энкодера, проектора и большой языковой модели.

Перед подачей в энкодер видео подвергается разреженной выборке кадров, чтобы снизить избыточность, сохранив ключевые моменты. Визуальный энкодер обычно строится на базе трансформерной архитектуры Vision Transformer (ViT) [2], которая разбивает кадр на патчи, превращает их в токены и пропускает через слои самовнимания. Для учёта временной оси используются видео-адаптации: например, TimeSformer факторизует внимание – один слой обрабатывает пространственные связи внутри кадра, другой – временные между кадрами [3]. Альтернативой служат модели, предобученные с помощью маскирования пространственно-временных блоков, таких как VideoMAE [4]. Другой подход – комбинирование нескольких энкодеров для получения устойчивого представления, как в InternVideo [5]. На выходе энкодер даёт последовательность векторных представлений, размер которой может быстро расти с длительностью видео.

Следующий компонент – проектор – приводит визуальные эмбеддинги к формату, понятному языковой модели. Простейший проектор – многослойный перцептрон, отображающий каждый вектор независимо. Более совершенные решения, такие как Q-Former или Perceiver Resampler, вводят небольшой фиксированный набор обучаемых запросов, которые через кросс-внимание извлекают из всего множества визуальных токенов наиболее релевантную информацию [6]. Это позволяет радикально сжимать представление: например, несколько сотен патчевых векторов сводятся к 32 токенам, сохраняя при этом смысловую нагрузку.

Завершает цепочку большая языковая модель (LLM), в качестве которой могут выступать открытые архитектуры наподобие LLaMA [7]. Визуальные токены от проектора объединяются с текстовым запросом и подаются как единая входная последовательность. LLM, предобученная на гигантских текстовых корпусах, интерпретирует визуальные сигналы как «мягкие» подсказки и задействует свои способности к логическому выводу для генерации ответа, описания или поиска нужного фрагмента.

Временной анализ и проблема масштабирования. Понимание видео – это не просто сумма анализа отдельных кадров, а восстановление логики событий: их порядка, длительности и причинно-следственных связей. В VLM временное рассуждение реализуется за счёт добавления позиционного кодирования, которое сообщает модели временную метку или относительное смещение каждого визуального токена. Механизм самовнимания трансформера способен связывать любые пары токенов, независимо от расстояния между ними в кадрах, и тем самым выстраивать полную картину динамики сцены.

Однако этот же механизм создаёт фундаментальную проблему для длинных видео. Вычислительная сложность самовнимания растёт квадратично относительно количества входных токенов. Если для минутного ролика число токенов ещё приемлемо, то для часового оно достигает десятков тысяч, что делает обработку невозможной на стандартном оборудовании.

Для преодоления этого барьера разработаны три группы методов, часто используемые совместно:

- Иерархическое представление. Видео анализируется в несколько уровней: сначала извлекаются признаки из коротких клипов, затем клиповые представления агрегируются в сцены, и только сценовые эмбеддинги поступают в LLM. Такой подход, реализованный, в частности, в системе MovieChat [8], уменьшает длину итоговой последовательности без потери общей структуры повествования.

- Компрессия визуальных токенов. Применяется слияние схожих токенов, адаптивный отбор информативных кадров или упомянутый выше Q-Former с фиксированным числом запросов. В результате LLM получает лишь наиболее значимые элементы, а не весь исходный поток.
- Механизмы долговременной памяти. Даже после сжатия контекстное окно LLM ограничено. Для обработки произвольно длинных видео модель сохраняет компактные состояния пройденных фрагментов в память, а при обработке очередного отрезка считывает релевантные записи. Это позволяет удерживать нить событий на протяжении нескольких часов, аналогично тому, как Transformer-XL работает с длинными текстами.

Стратегия обучения VLM. Обучение VLM обычно разбито на три этапа, что позволяет совместить сильные стороны замороженных компонентов.

Первый этап – предобучение визуального энкодера. Он учится на больших неразмеченных коллекциях видео с помощью контрастивных задач (например, в духе CLIP, где текст является якорем) или самоконтролируемого маскирования. Цель – получить универсальное представление, улавливающее и пространственные, и временные закономерности.

Второй этап – согласование модальностей. Здесь замораживаются веса и энкодера, и LLM, а обучается только проектор. На парах «видео – текст» модель учится отображать визуальные эмбеддинги в те области семантического пространства, которые соответствуют правильным текстовым понятиям. Типичная функция потерь – кросс-энтропия, побуждающая LLM восстанавливать описание видео по поданным визуальным токенам.

Третий этап – инструктивное дообучение. Модель получает смешанные данные: видео, вопрос и целевой ответ. LLM (иногда вместе с проектором) дообучается на задачах следования инструкциям. Часто используются параметроэффективные методы, например LoRA, чтобы адаптироваться к новым задачам без разрушения общих знаний. На этом этапе VLM приобретает способность вести диалог, искать события и рассуждать на основе видеоконтента.

Заключение. Video-Language Models объединяют мощь пространственно-временных визуальных энкодеров с рассуждающей способностью больших языковых моделей. Ключевыми принципами их построения являются модульная архитектура с обучаемым проектором, явное временное кодирование и комбинация иерархических, компрессионных и мемориальных методов для обработки длинных видео. Многостадийное обучение обеспечивает согласование модальностей и гибкую настройку на практические задачи. Дальнейший прогресс связывается с линейным масштабированием внимания, более глубоким причинным анализом событий и внедрением VLM в интерактивные системы реального времени.

Список литературы:

1. Zhang J. Vision-Language Models for Vision Tasks: A Survey / J. Zhang, J. Huang, S. Jin, S. Lu // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2024. – DOI: 10.1109/TPAMI.2024.3355396. – URL: https://ieeexplore.ieee.org/document/10409519 (дата обращения: 21.06.2026).

2. Dosovitskiy A. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale / A. Dosovitskiy, L. Beyer, A. Kolesnikov [et al.] // International Conference on Learning Representations (ICLR). – 2021. – URL: https://openreview.net/forum?id=YicbFdNTTy (дата обращения: 21.06.2026).

3. Bertasius G. Is Space-Time Attention All You Need for Video Understanding? / G. Bertasius, H. Wang, L. Torresani // Proceedings of the International Conference on Machine Learning (ICML). – 2021. – P. 813–824.

4. Tong Z. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training / Z. Tong, Y. Song, J. Wang, L. Wang // Advances in Neural Information Processing Systems (NeurIPS). – 2022. – URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/416f9cb3276121c42e70286309a5d2e7-Abstract-Conference.html (дата обращения: 21.06.2026).

5. Wang Y. InternVideo: General Video Foundation Models via Generative and Discriminative Learning / Y. Wang, K. Li, Y. Li [et al.] // arXiv preprint. – 2022. – URL: https://arxiv.org/abs/2212.03191 (дата обращения: 21.06.2026).

6. Li J. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models / J. Li, D. Li, S. Savarese, S. Hoi // International Conference on Machine Learning (ICML). – 2023. – P. 19730–19742.

7. Touvron H. LLaMA: Open and Efficient Foundation Language Models / H. Touvron, T. Lavril, G. Izacard [et al.] // arXiv preprint. – 2023. – URL: https://arxiv.org/abs/2302.13971 (дата обращения: 21.06.2026).

8. Song E. MovieChat: From Dense Token to Sparse Memory in Long Video Understanding / E. Song, W. Chai, G. Wang [et al.] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – 2024. – P. 18233–18243.