

Видео и визуально- языковые модели

Курс “Мультимодальные модели”



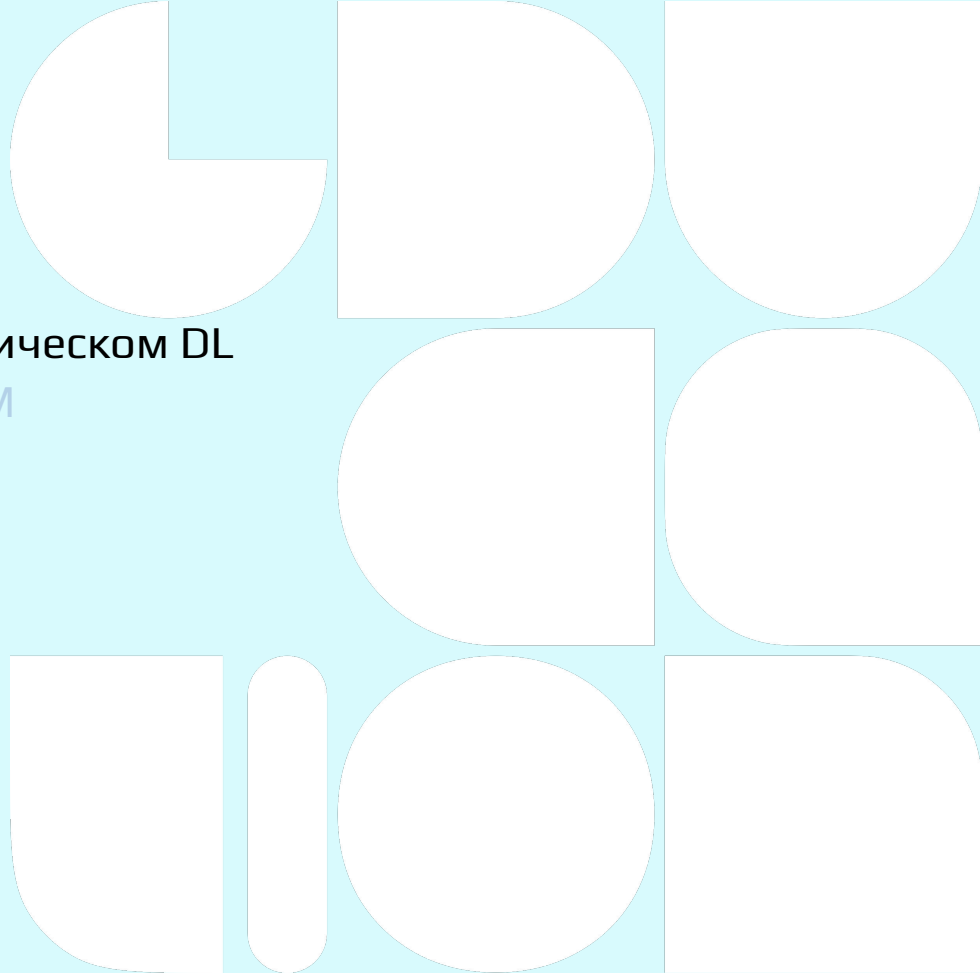
Кирилл Владимирович Каймаков

Кандидат компьютерных наук
Tech Lead группы мультимодальных
языковых моделей VK
Преподаватель департамента политики и
управления НИУ ВШЭ



О чём поговорим?

1. Принципы в работе с видео в классическом DL
2. Принципы в работе с видео в MMLM
3. Задачи в видео
4. Применение видео-MMLM
5. Примеры видео-MMLM

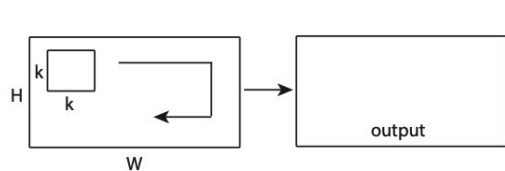
- 
- An abstract geometric pattern composed of various white shapes on a light blue background. The shapes include circles, semi-circles, and rounded rectangles, some of which are partially cut off by the edges of the frame. The arrangement is non-repeating and organic.
1. Принципы в работе с видео в классическом DL
 2. Принципы в работе с видео в MMLM
 3. Задачи в видео
 4. Применение видео-MMLM
 5. Примеры видео-MMLM

Чем видео отличается от картинки?

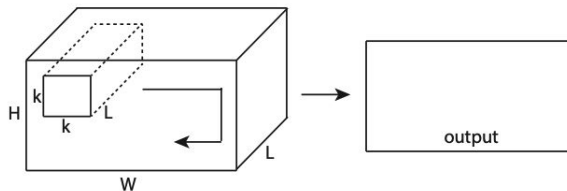
1. Видео нельзя обработать как единый кадр
2. У видео бывает разная частота дискретизации
3. Невизуальная информация о видео (звук, титры, название, описание и тд и тп) гораздо шире
4. Присутствие времени в видео
5. Логическое пространство видео отличается от картиночного
6. Видео не обязательно снимают на одну камеру, а камеры бывают в том числе 360 градусные
7. В видео бывают сцены разной длительности и разной важности

C3D (2014)

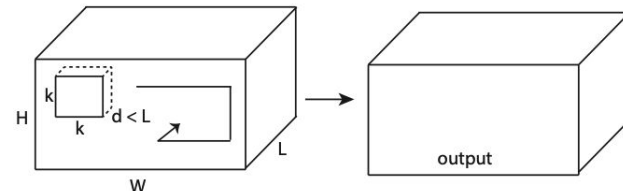
Первые успешные свёрточные сети, расширившие 2D-свёртки для анализа движения во времени.



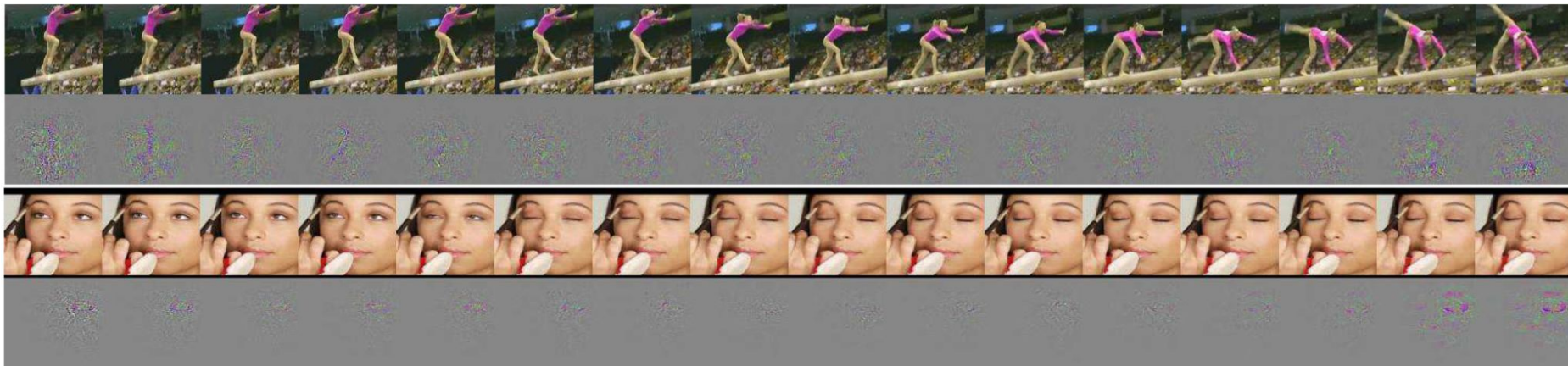
(a) 2D convolution



(b) 2D convolution on multiple frames

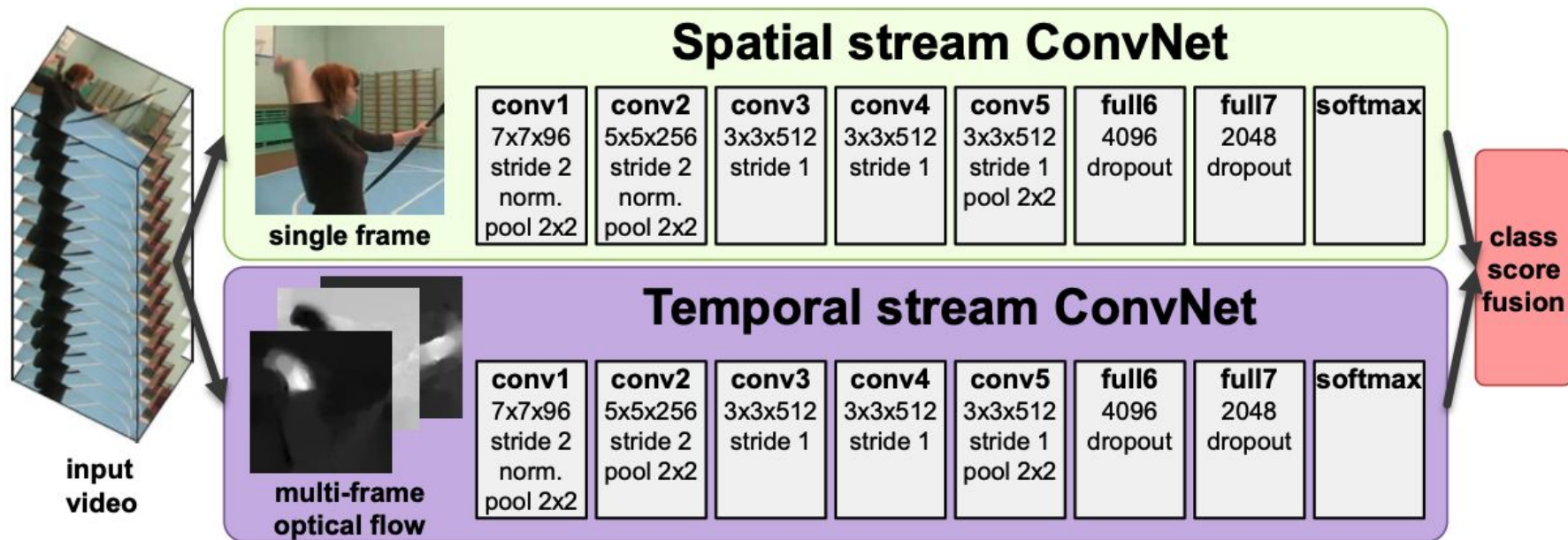


(c) 3D convolution



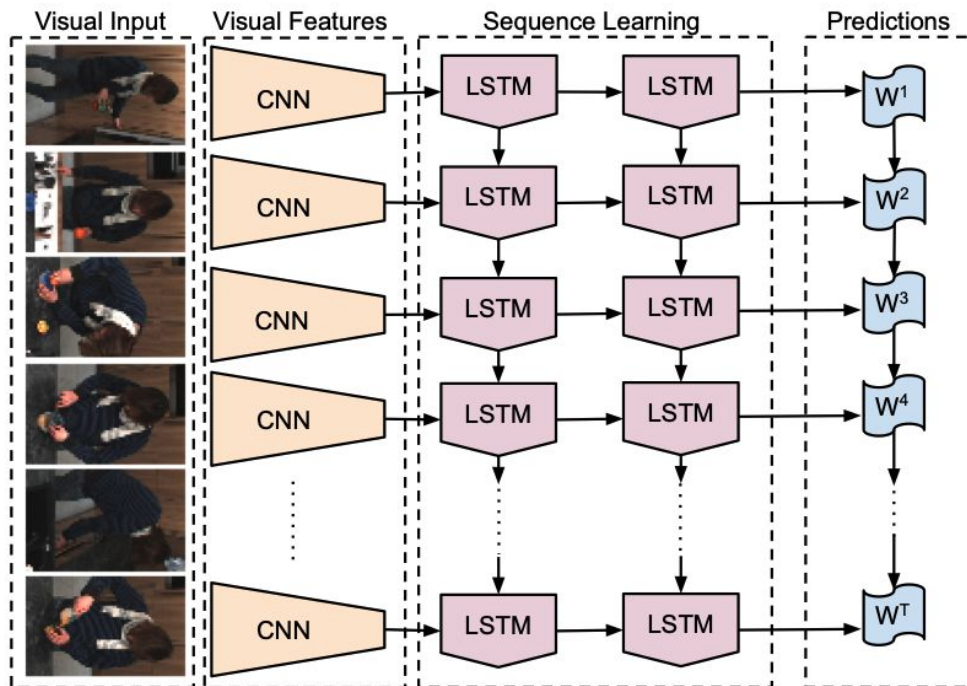
Two-stream network (2014)

У сети две ветви - одна обрабатывает RGB кадры, другая - поток оптического движения

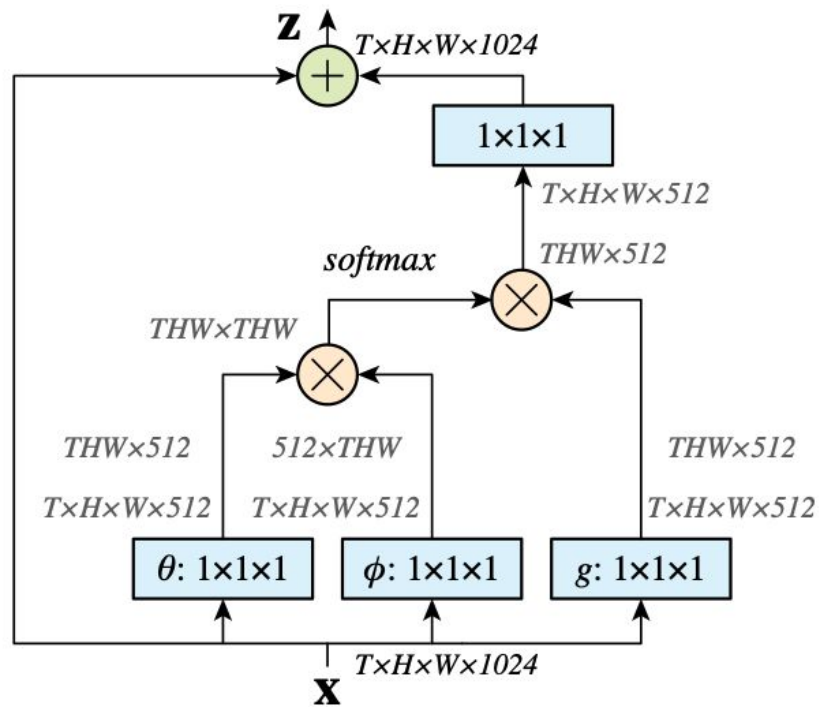


LRCN (2014)

Комбинация CNN + LSTM. Одна из первых рабочих схем для Video Captioning'a.



NLNN (2018)

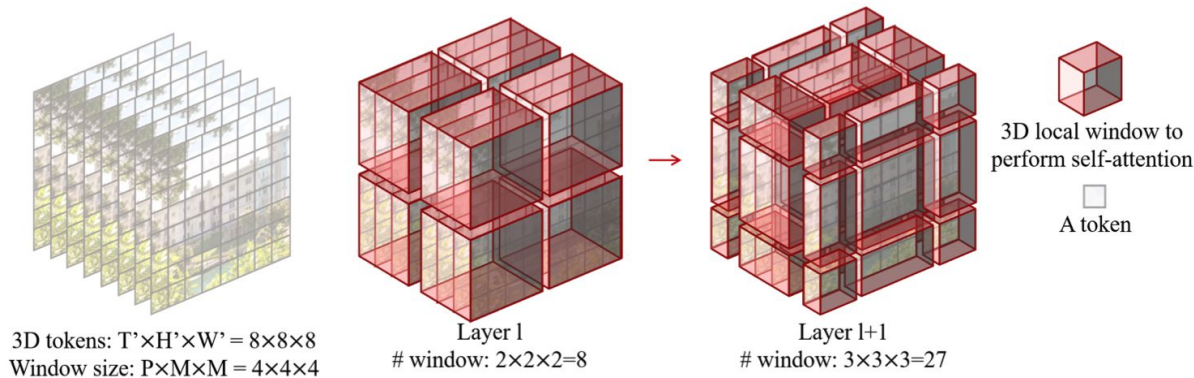
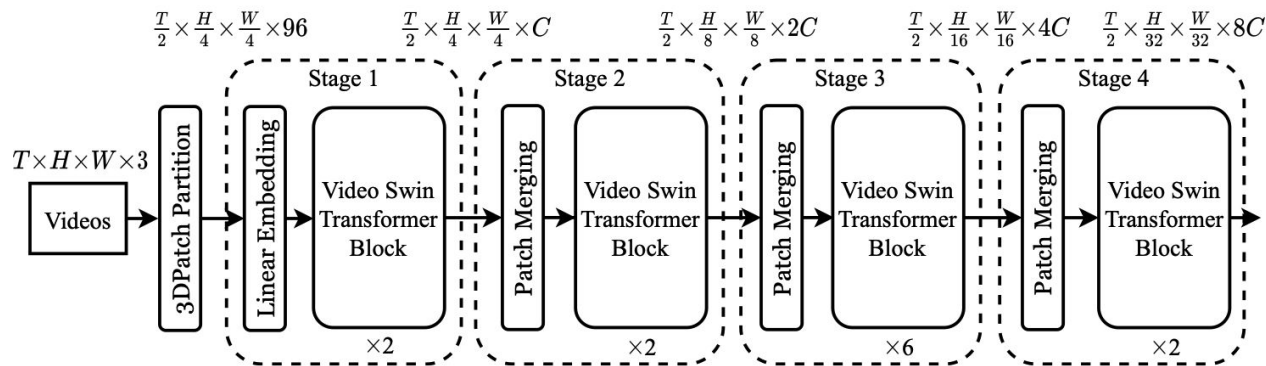


Self-attention для видео, вычисление attention'a между всеми парами позиций по всему видео.




Video SWIN (2021)

Опять 3D свертки. Только теперь с attention'ом.

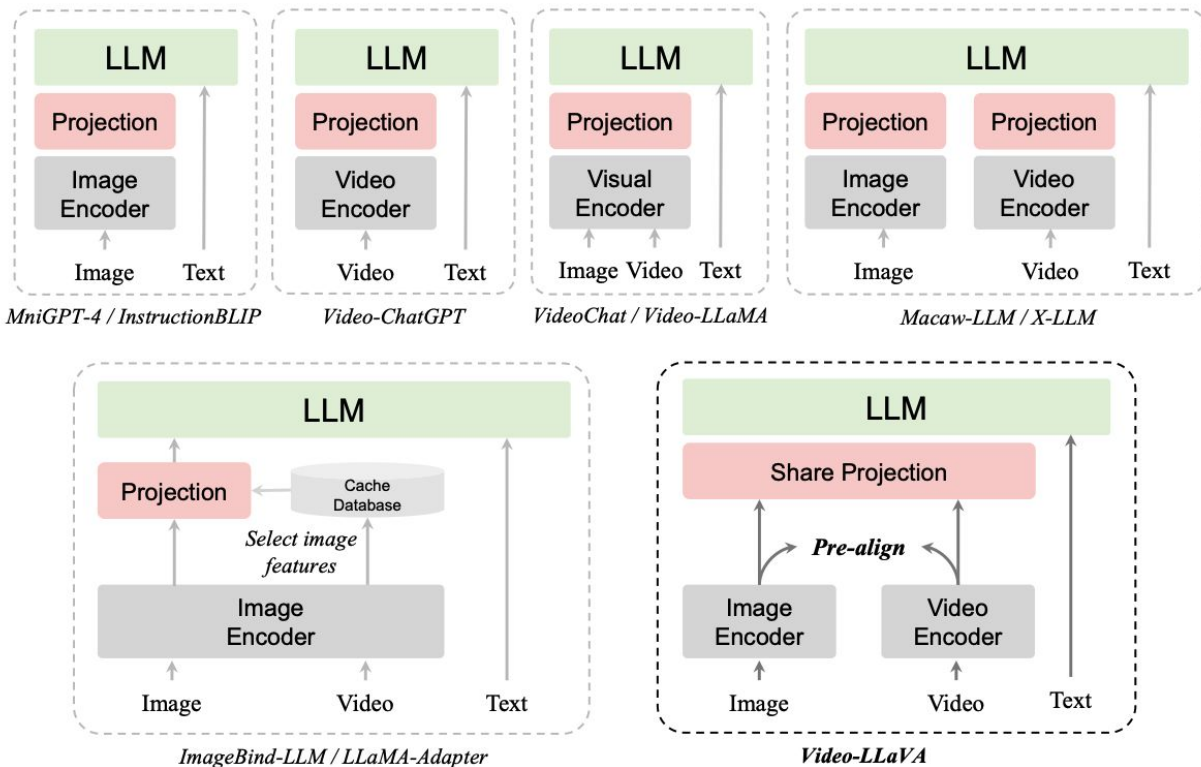


Какие знаковые модели
для видео вы еще
сможете вспомнить?



- 
- An abstract geometric pattern consisting of various white shapes on a light blue background. The shapes include a large circle with a 90-degree wedge removed, a square with a semi-circle on its right side, a semi-circle, a circle with a semi-circle on its right side, a square with a semi-circle on its left side, a circle with a semi-circle on its left side, a square with a semi-circle on its top side, a semi-circle, a circle with a semi-circle on its top side, a square with a semi-circle on its bottom side, a circle with a semi-circle on its bottom side, a square with a semi-circle on its left side, a circle with a semi-circle on its left side, a square with a semi-circle on its right side, and a circle with a semi-circle on its right side.
- 1. Принципы в работе с видео в классическом DL
 - 2. **Принципы в работе с видео в MMLM**
 - 3. Задачи в видео
 - 4. Применение видео-MMLM
 - 5. Примеры видео-MMLM

Варианты работы с энкодерами с видео в VLM



Визуальная модальность

Как подавать кадры в VLM?

1. Image captioning
2. Подача эмбединга видео
3. Подача эмбединга кадров
4. Подача среднего эмбединга кадров
5. Использование Pooling'ов
6. Комбинационные сети

Сколько кадров подавать?

1. 10 или 64?
2. Или раз в секунду?
3. Или подать все кадры из видео?

Звуковая модальность

Как подавать звук в VLM?

1. ASR
2. Эмбединг звуковой сетки

А нужен ли звук вообще?

1. Да?
2. Нет?

Текстовая модальность

Какую информацию о видео полезно подавать в VLM?

1. Название видео
2. Описание
3. Тэги
4. Название опубликовавшего сообщества
5. Поисковые запросы, по которому ищут видео
6. Объяснение информации из RAG'a

Каким образом форматировать время?

1. Фиксированные отрезки да/нет?
2. Час/минута/секунда?
3. Текстом в секундах?
4. Или делать доп энкодер?

Какие еще модальности
бывают?




Способы сбора данных для обучения

Каким образом можно собирать данные для обучения Video-VLM?

1. Существующие датасеты
 - а. Прямое использование
 - б. Генерация инструкций на основе известных данных с помощью LLM
2. Ручная разметка
 - а. Генерация описаний
 - б. Разметка под конкретную задачу
3. Полуручная разметка
 - а. Фильтрация генерации моделей
 - б. Дообучение модели на предыдущих результатах
4. Генерация разметки с помощью LLM моделей

Вопросы?



- 
1. Принципы в работе с видео в классическом DL
 2. Принципы в работе с видео в MMLM
 3. **Задачи в видео**
 4. Применение видео-MMLM
 5. Примеры видео-MMLM

Какие задачи для
обучения Video-VLM вы
сможете придумать?



Задачи для Video-MMLM

Video Caption

1

Генерация текстового описания, отражающего ключевые события или содержание видео

Video Advanced Caption

2

Генерация длинного текстового описания видео, содержащие такую информацию как, к примеру, имена актеров и факты из их биографии

Video QA

3

Ответы на вопросы о содержании видео

Video based-dialogue

4

Ведение диалога на основе информации из видеоролика

Visual Grounding

5

Связывание упомянутых в тексте объектов с их визуальным представлением в видео

Style analysis

6

Анализ визуального и художественного стиля видео

Задачи для Video-MMLM

Video Comparison

1

Сравнение двух и более видео по содержанию, стилю, действиям и характеристикам

Video Reasoning

2

Объяснение причин и следствий для действий, происходящих на видео

Identity Consistency

3

Проверка совпадения персонажа в разных моментах видео

OCR

4

Извлечение и распознавание текста из видео

Assessor

5

Оценка качества видео, адекватности описания, релевантности запросу и тд и тп

Instruction generation

6

Создание пошаговых инструкций на основе действий, показанных в видео

Задачи для Video-MMLM

Temporal changes detection

1

Определение изменений происходящих во времени в видео, к примеру, перемещений объектов

Video emotion and Sentiment Analysis

3

Детектирование и классификация эмоций людей на видео

Video instruction following

5

Выполнение инструкций пользователя по взаимодействию с видео

Action segmentation and classification

2

Детектирование и классификация действий человека или животного, происходящих на видео

Anomaly detection

4

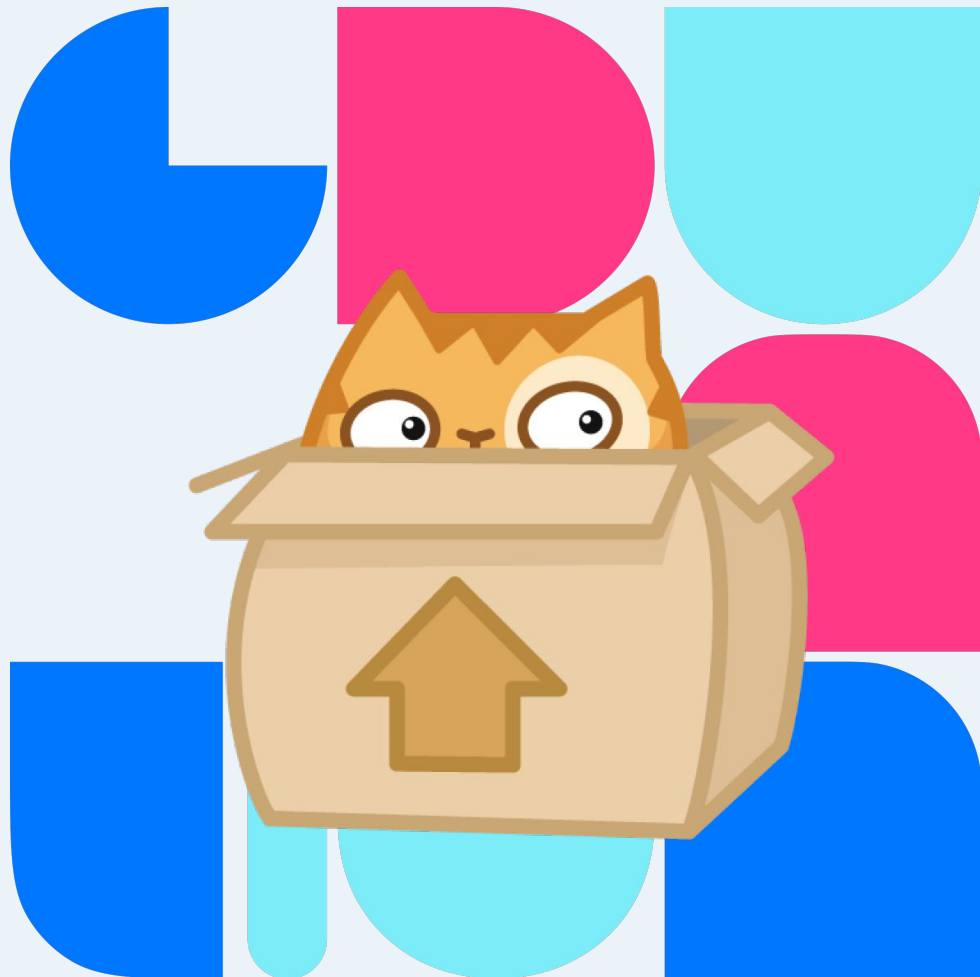
Обнаружение на видео необычных и неожиданных событий и отклонений от нормы

Gesture and body language analysis

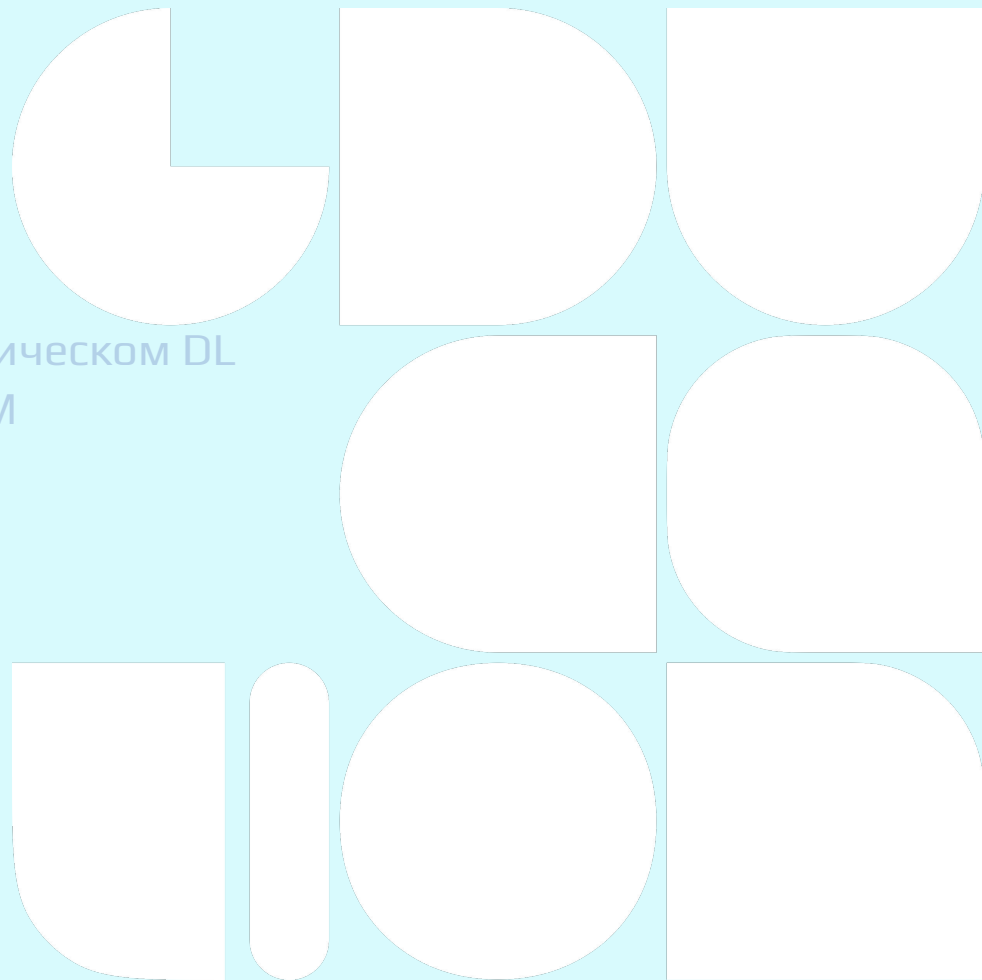
6

Анализ жестов, мимики, языка тела персонажей на видео, распознавание языка жестов

Какие задачи из
перечисленных не
имеют картиночных
аналогов?



1. Принципы в работе с видео в классическом DL
2. Принципы в работе с видео в MMLM
3. Задачи в видео
4. **Применение видео-MMLM**
5. Примеры видео-MMLM



Где можно применить Video-VLM?

Умный поиск по видео

1

Анализ и нахождение нужных пользователю моментов в видео

Модель-ассистент

3

Работа с видео, получаемого с камеры устройства (первый пример - GPT-4o)

Видео-аналитика

5

Анализ событий, происходящих на видео, выявление аномалий

Модель-ассесор

2

Разметка данных для создания более легковесных моделей и уменьшение количества ручного труда

Образование

4

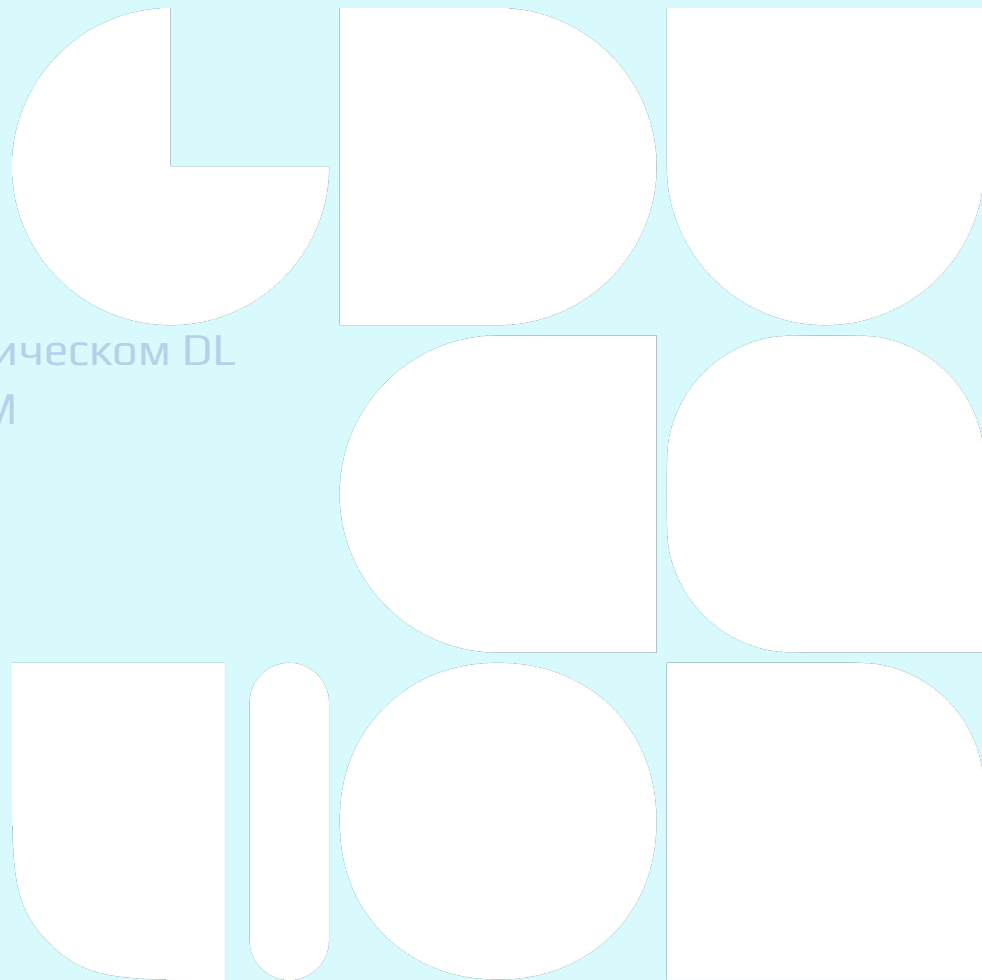
Проверка честности выполнения заданий студентами и генерация индивидуальных вариантов

Робототехника

6

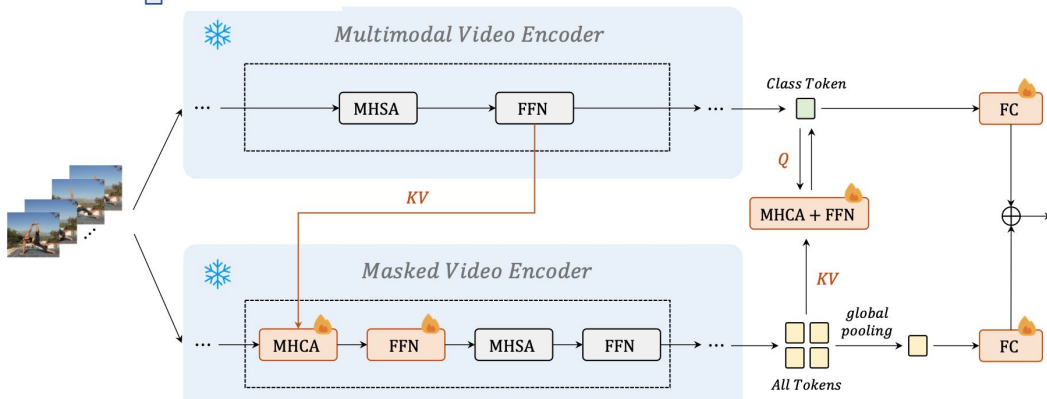
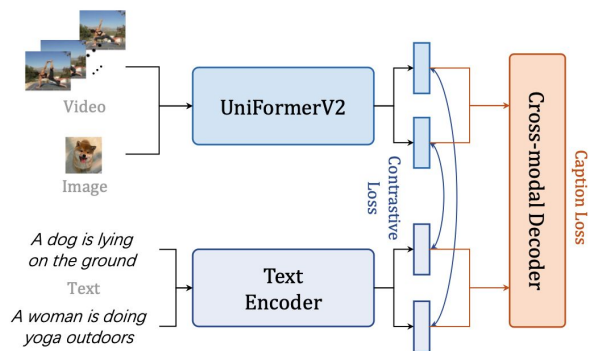
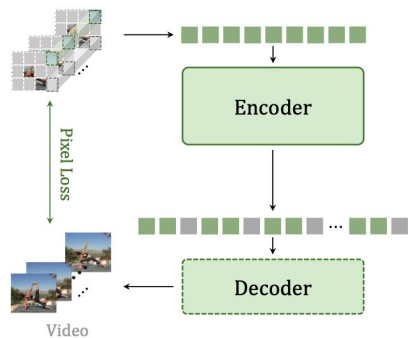
Роботы, понимающие команды человека на естественном языке, военная техника, требующая от человека лишь подтверждения

1. Принципы в работе с видео в классическом DL
2. Принципы в работе с видео в MMLM
3. Задачи в видео
4. Применение видео-MMLM
5. **Примеры видео-MMLM**



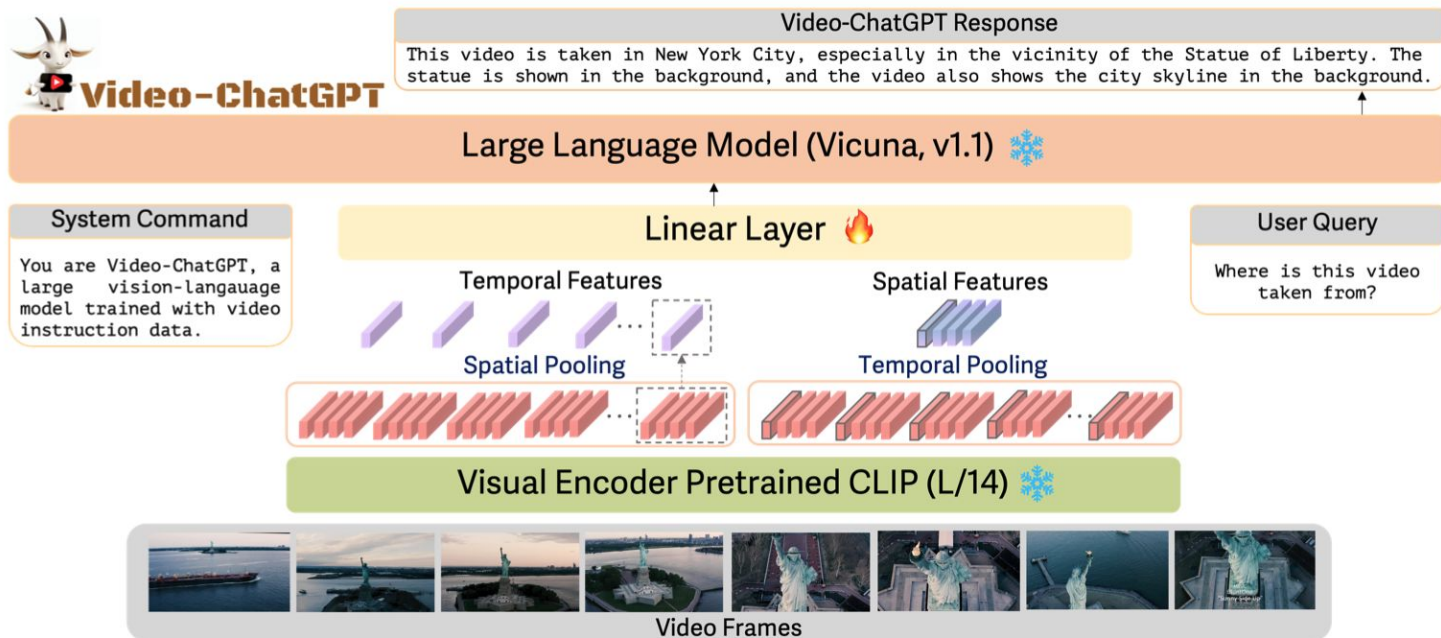
InternVideo

Первая general video модель
Еще скорее backbone



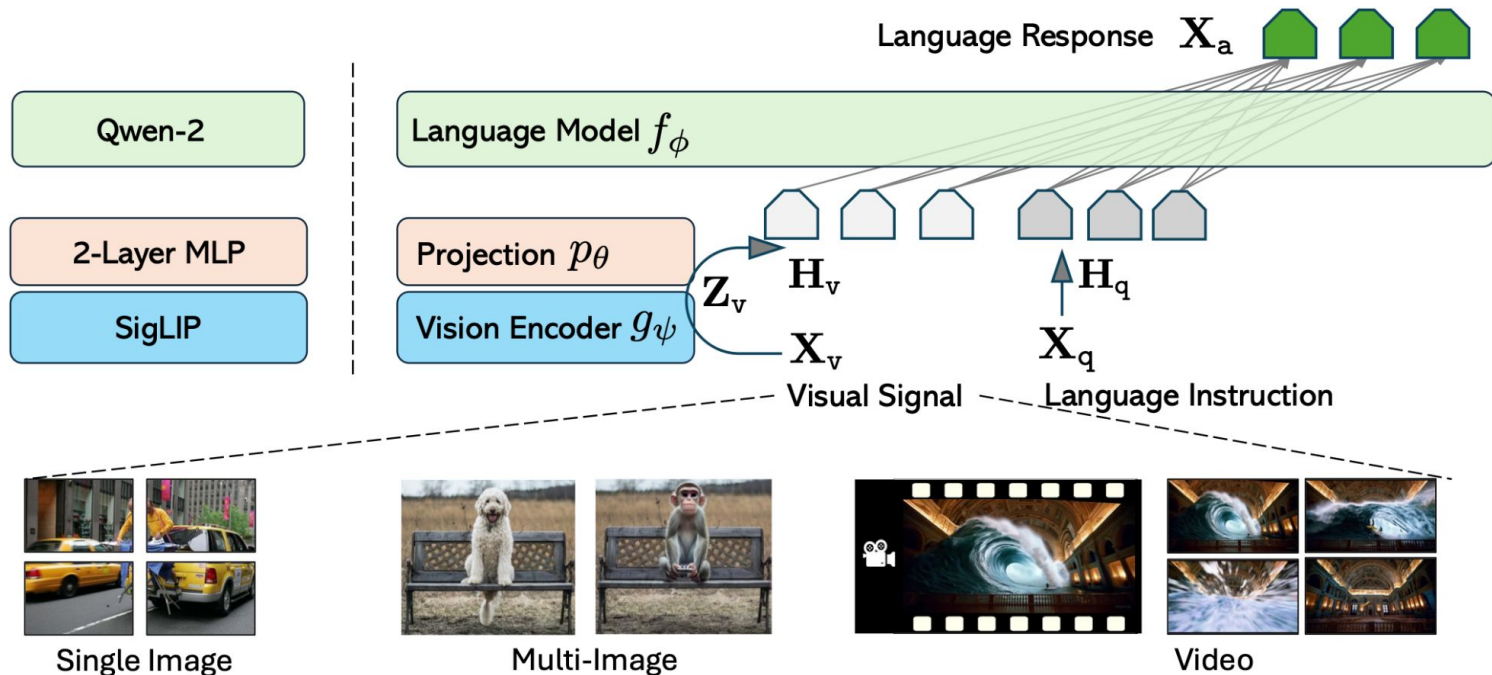
Video-ChatGPT

He OpenAI
Еще старались собрать данные с помощью людей
Одна из первых Video-VLM



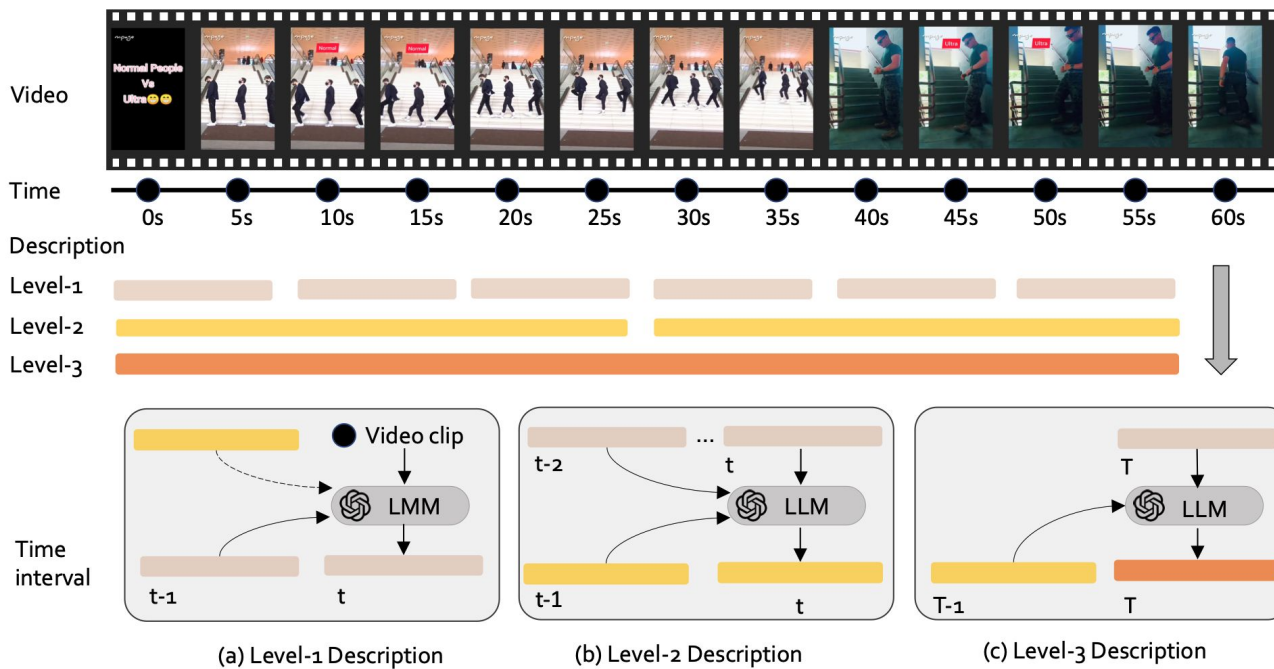
LLaVA-OneVision

Omni-модель LLaVA, работающая с изображениями, видео и текстом в единой архитектуре



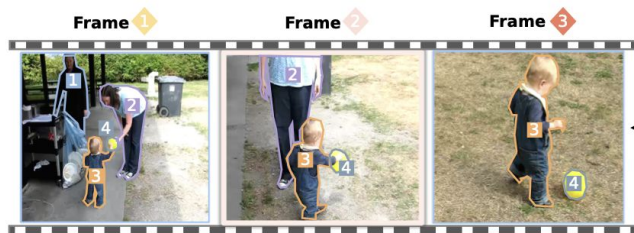
LLaVA-Video

LLaVA-One-Vision, но со сложной многоярусной механикой формирования обучающего датасета. Рассчитана на работу с короткими клипами.



InstIT

Первая известная видео-модель, в обучении которой в инструкциях активно использовались Video-grounding задачи



Instance-level Captions:

- 2: a person wearing a light blue patterned top and dark pants.
- 3: a small child with short hair, wearing a blue denim overall.
- 4: a bright yellow ball being held by 3.

Temporal Change Captions:

- 1 is **no longer visible**.
- 2 has moved **from a bending to an upright position** and....
- 3 has changed direction and appears to have **moved further away from 2**
- 4 is in the grasp of 3, suggesting 3 has **maintained its possession**....

(a) Frame-level annotations at frame 2

Image-level Captions:

- 2 is standing upright, visible from the waist down, wearing dark pants and standing on a concrete and grassy surface.
- 3, a small child in denim overalls, holding 4, a yellow object, in its right hand. The scene appears to be in a partially outdoor.... The angle is elevated, showing the scene from a standing viewpoint, focusing on interaction between 2, 3, and 4....



At 1, 1, a person in a dark outfit, is visible partially under a pavilion. 2, dressed in a light blue patterned top and dark pants, is bending forward towards 3, a small child in blue denim overalls. 2 extends 4, a bright yellow ball, towards 3, who is standing nearby, seemingly reaching for the ball. At 2, 1 has left the frame, and 2 has shifted to an upright position.... 3 has taken possession of 4.... The camera angle has also adjusted, showcasing the gravel path and the surrounding grassy area. At 3, 3 is central in the frame, is facing away from the camera, slightly bent forward, with....

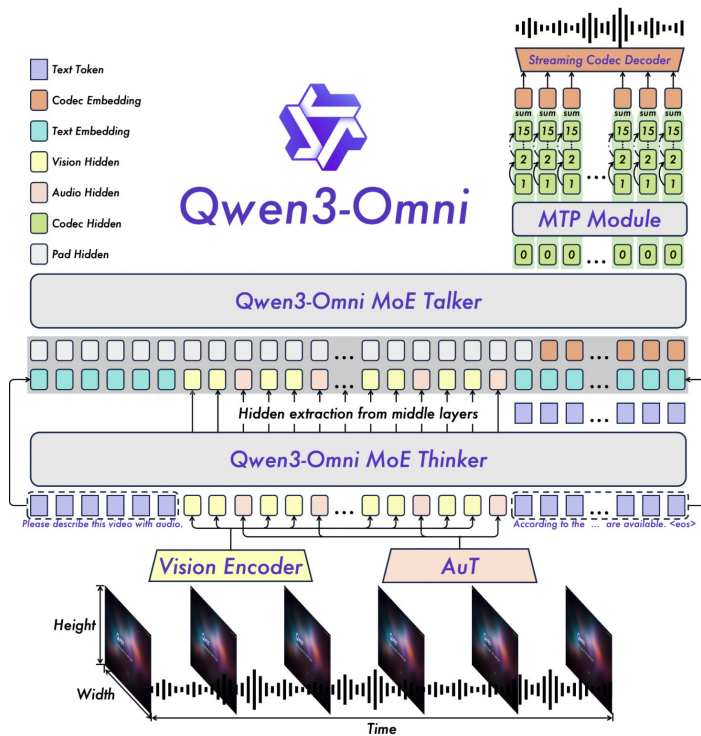
(b) Entire video description

- Q: What object did 2 hold towards 3 in the first frame?
A: 2 held a bright yellow ball 4 towards 3.
- Q: What change occurred to 1 between frames 1 and 2 ?
A: [1] is no longer visible possibly caused by camera movement.
- Q: What happens to 2 from frame 1 to frame 2 ?
A: 2 moved from bending to an upright, and is less visible.
- Q: ...?

(c) Question-answer pairs

Qwen3-Omni

На момент создания презентации самая современная omni-модель: MOE, Thinking, текст, изображения, видео,



Какие особенности
развития видео-
моделей вы заметили?



Что мы обсудили на этой лекции?

Историю методов обработки видео в глубоком обучении

1

Проблемы при работе с видео

2

Задачи, которые может решать в видео MMLM

3

Методы работы с видео в мультимодальных языковых моделях

4

Способы сбора данных для обучения видео MMLM

5

Применение видео MMLM на практике

6



Спасибо
за внимание!

