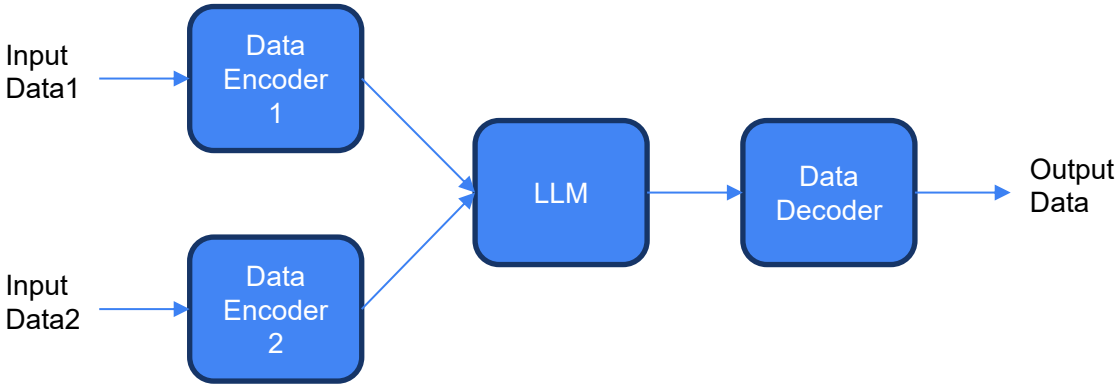
Введение

Что такое модальность?

- Модальность (из психологии) – форма восприятия и представления информации в мышлении и памяти
- Психологи выделяют визуальную, аудиальную и кинестетическую модальности

Что такое модальность?

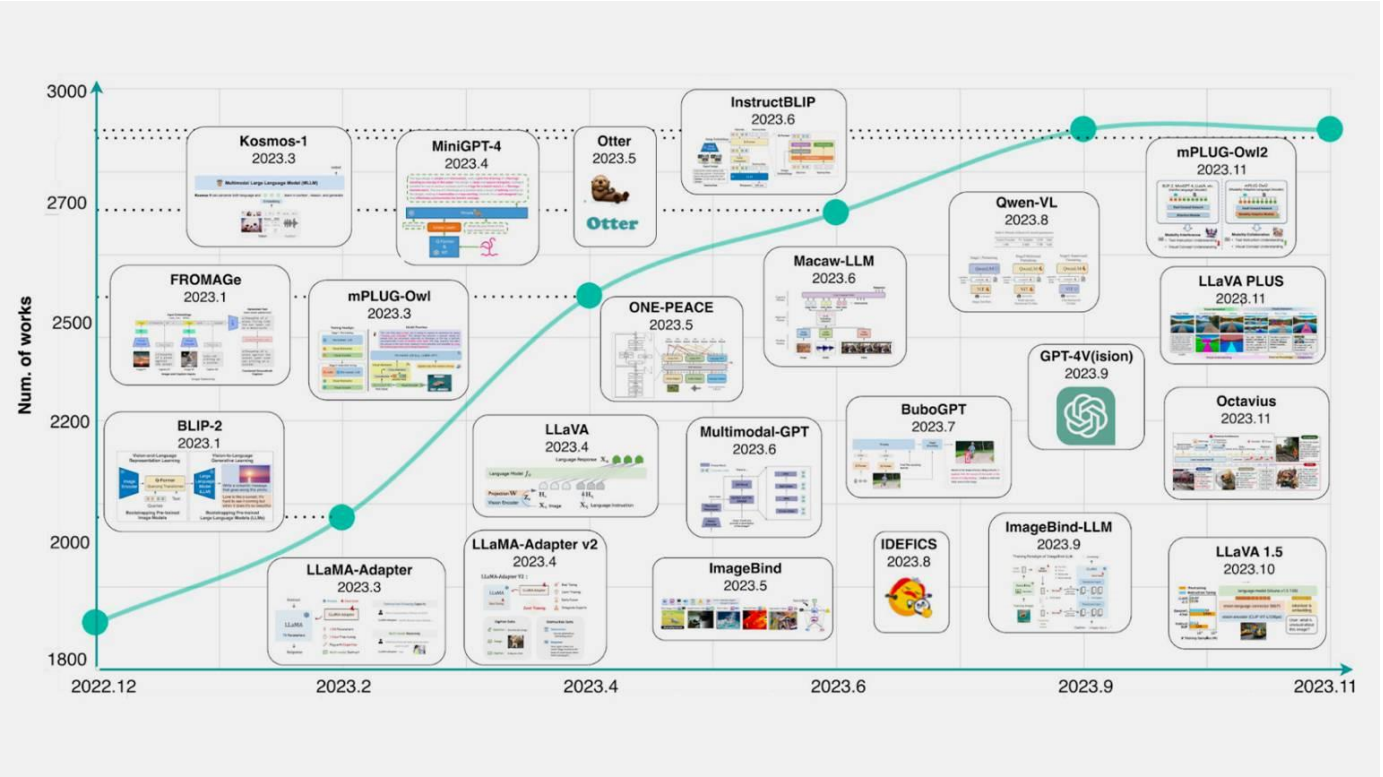
Что такое модальность?



Что такое модальность?

- Модальность – данные, требующие отдельного способа кодирования и декодирования
- Мультимодальная нейросеть – нейросеть, принимающая на вход несколько модальностей

Число публикаций



Почему развивается мультимодальность?

Больше данных ⟹ лучше понимание мира
Генерализация ⟹ zero-shot задачи
Унификация ⟹ меньше зоопарк моделей

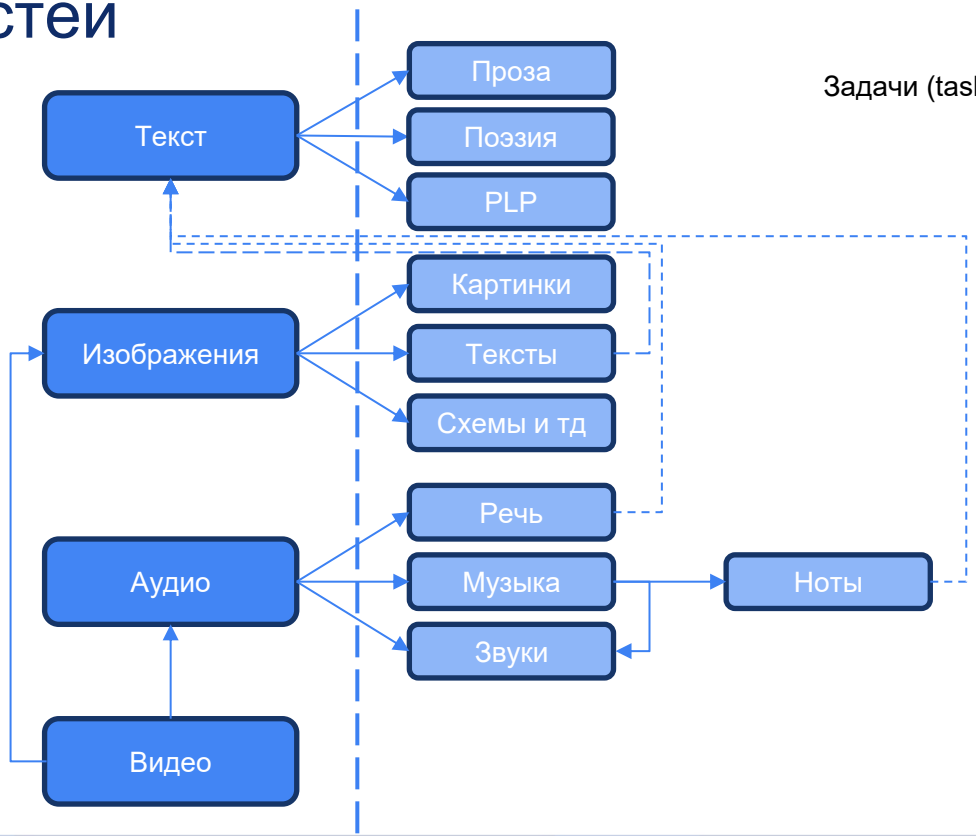
Виды модальностей

- Текст
- Изображения
- Аудио
- Видео

Виды модальностей

Модальности

Задачи (tasks)



Виды модальностей

Графы

```
"graph": {
    "a": ["b", "c"],
    "b": ["c"]
    "c"
},

"nodes": {
    "a": {
        "name": "Adam"
    },
    "b": {
        "name": "Bob"
    },
    "c": {
        "name": "Caillou"
    }
},
```

3D, Cad

```
# OBJ file format with ext .obj
# vertex count = 2503
# face count = 4968
v -3.4101800e-003 1.3031957e-001 2.1754370e-002
v -8.1719160e-002 1.5250145e-001 2.9656090e-002
v -3.0543480e-002 1.2477885e-001 1.0983400e-003
v -2.4901590e-002 1.1211138e-001 3.7560240e-002
...
f 1069 1647 1578
f 1058 909 939
f 421 1176 238
f 1055 1101 1042
f 238 1059 1126
f 1254 30 1261
f 1065 1071 1
f 1037 1130 1120
f 1570 2381 1585
f 2434 2502 2473
f 1632 1654 1646
...
```

Виды модальностей

Графы

```
"graph": {
    "a": ["b", "c"],
    "b": ["c"]
    "c"
},

"nodes": {
    "a": {
        "name": "Adam"
    },
    "b": {
        "name": "Bob"
    },
    "c": {
        "name": "Caillou"
    }
},
```

3D, Cad

```
# OBJ file format with ext .obj
# vertex count = 2503
# face count = 4968
v -3.4101800e-003 1.3031957e-001 2.1754370e-002
v -8.1719160e-002 1.5250145e-001 2.9656090e-002
v -3.0543480e-002 1.2477885e-001 1.0983400e-003
v -2.4901590e-002 1.1211138e-001 3.7560240e-002
...
f 1069 1647 1578
f 1058 909 939
f 421 1176 238
f 1055 1101 1042
f 238 1059 1126
f 1254 30 1261
f 1065 1071 1
f 1037 1130 1120
f 1570 2381 1585
f 2434 2502 2473
f 1632 1654 1646
...
```

Мультимодальные комбинации

- Текст + изображения
- Текст + аудио
- Текст + аудио + изображения
- Омнимодальность (все возможные модальности)

Мультимодальные задачи

Текст + изображения

- Image captioning
- Image generation
- Image editing
- Image QA (VQA, DocVQA)
- Image retrieval

Мультимодальные задачи

Текст + аудио

- ASR
- TTS
- Voice mode
- Emotion recognition
- Sound captioning
- Sound QA
- Speech translation
- Audio retrieval
- etc

Мультимодальные задачи

Текст + аудио + изображения

- Video summarization
- Highlights searching
- etc

Мультимодальные датасеты

- Изображения + текст: Visual Storytelling Dataset, Visual Question Answering Dataset, LAION-5B Dataset.
- Аудио + изображения: VGG-Sound Dataset, RAVDESS Dataset, Audio-Visual Identity Database (AVID).
- Изображения + аудио + текст: RECOLA Database, IEMOCAP Dataset.

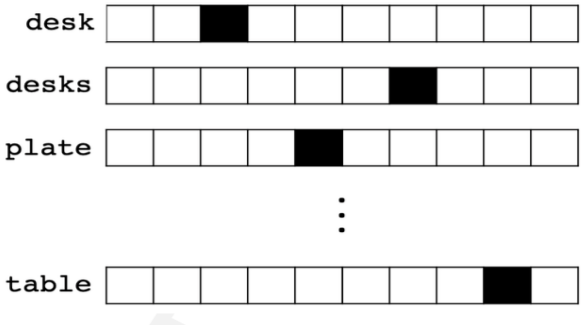
Ограничения при обучении мультимодальных моделей

- Объем размеченных данных
- Длинна контекста трансформера
- Квадратичная сложность при вычислении аттеншена

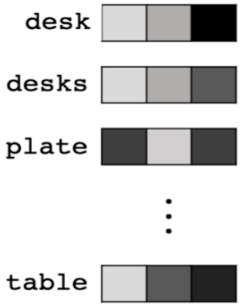
Текст

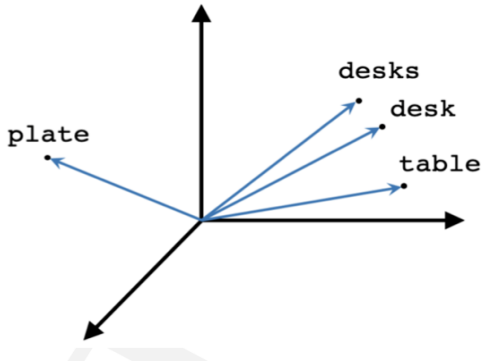
BOW

```
{basket:1,
 fork:2,
 desk:3,
 cloud:4,
 plate:5,
 rabbit:6,
 desks:7,
 tree:8,
 table:9,
 lion:10}
```

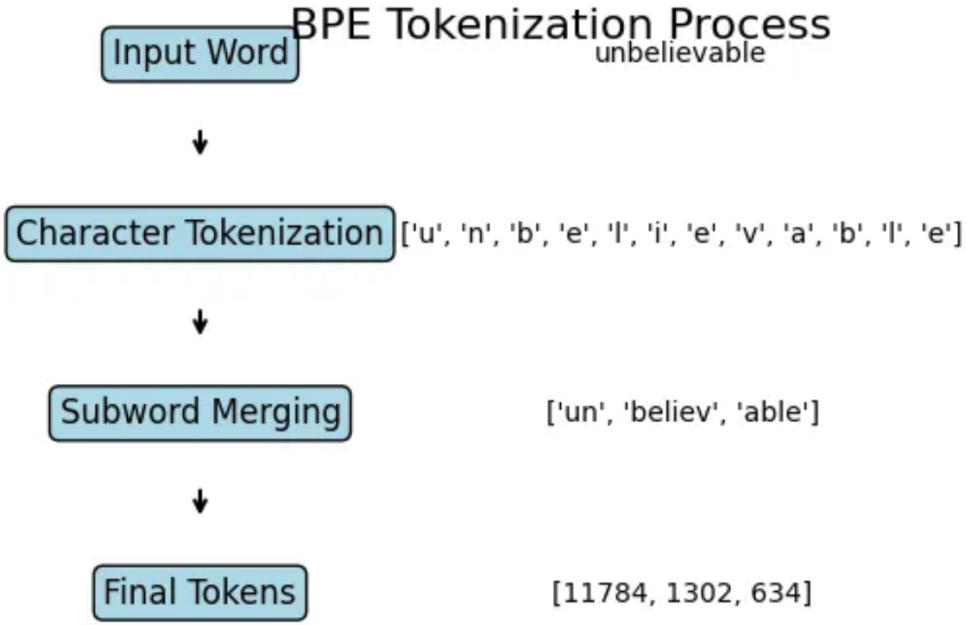


word
embeddings

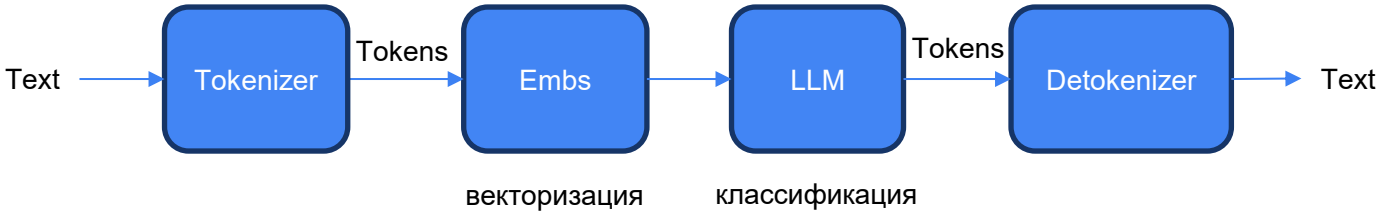




Текст


BPE Tokenization Process
Input Word
unbelievable
Character Tokenization
['u', 'n', 'b', 'e', 'l', 'i', 'e', 'v', 'a', 'b', 'l', 'e']
Subword Merging
['un', 'believ', 'able']
Final Tokens
[11784, 1302, 634]

Текст



Задача – предсказание следующего токена

Особенности токенизации

PLP

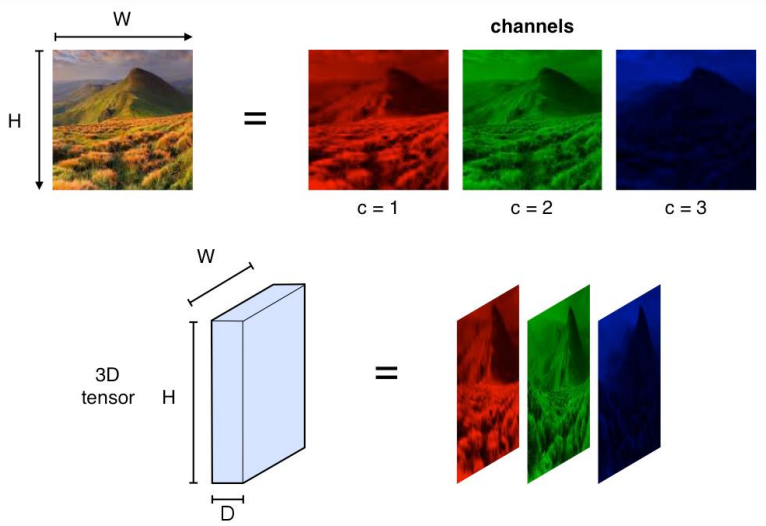
- другое распределение токенов
- структура языка

Поэзия

- нужна ритмика языка

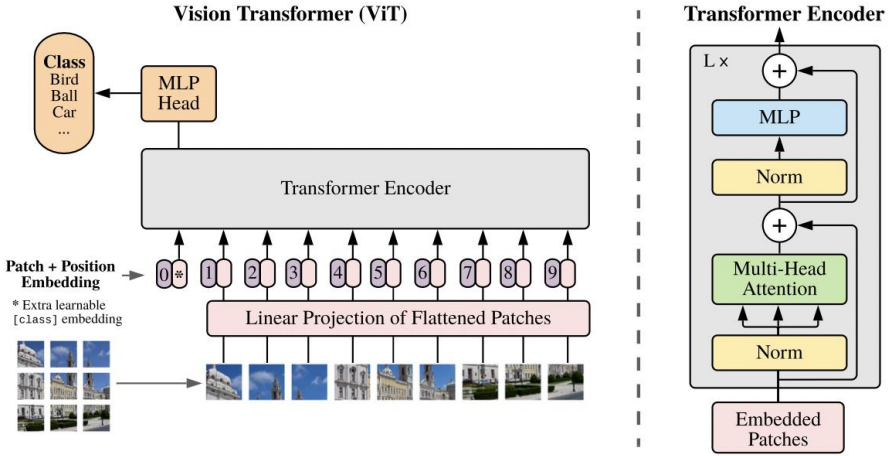
Изображения

- Дискретные данные (пиксели)
- Пространственная структура (матрица пикслей, слои)
- Детали могут быть где угодно



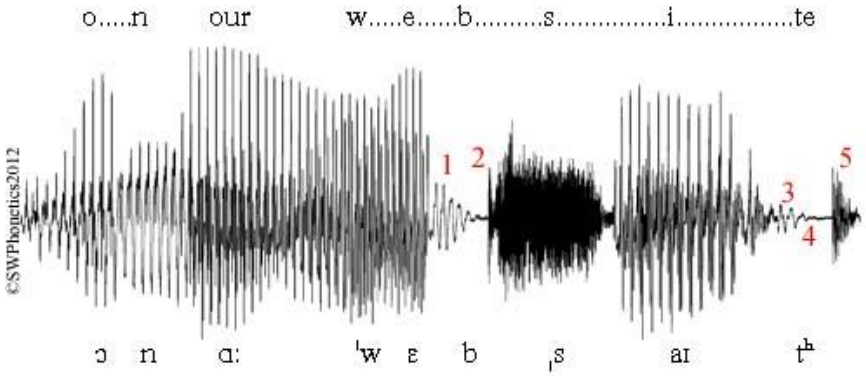
ViT. Изображение в последовательность

- Нарезка изображения на сегменты
- Преобразование сегмента в вектор
- Обучение трансформера



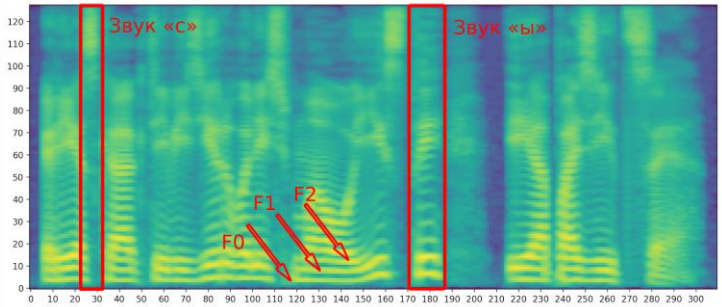
Аудио

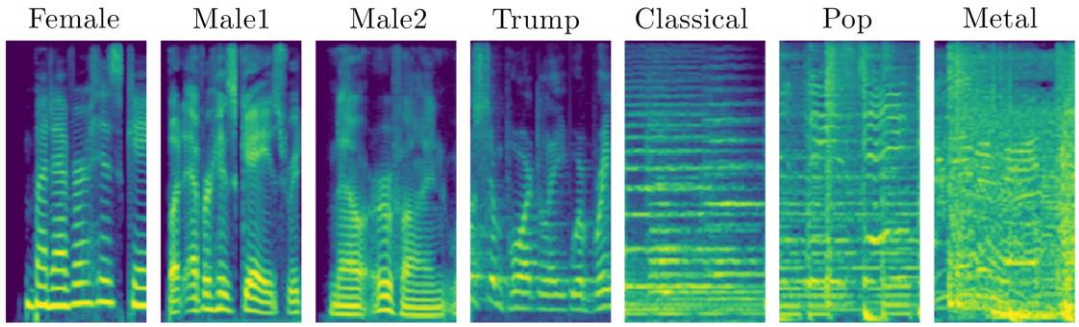
- Непрерывный сигнал
- Последовательность кадров звука



Аудио. Спектрограмма

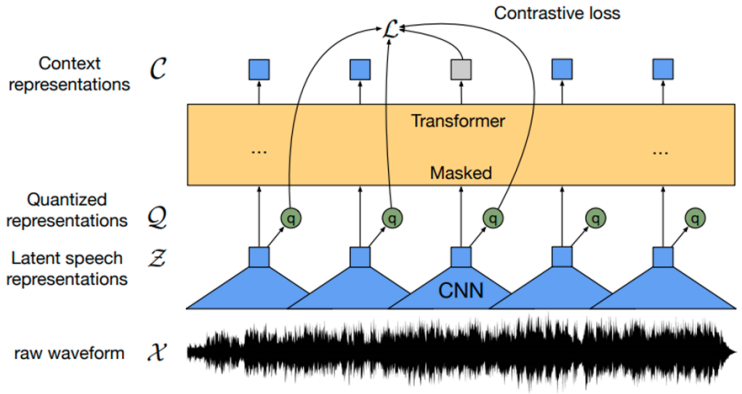
- Дискретное преобразование Фурье
- Разложение сигнала на амплитуду и фазу



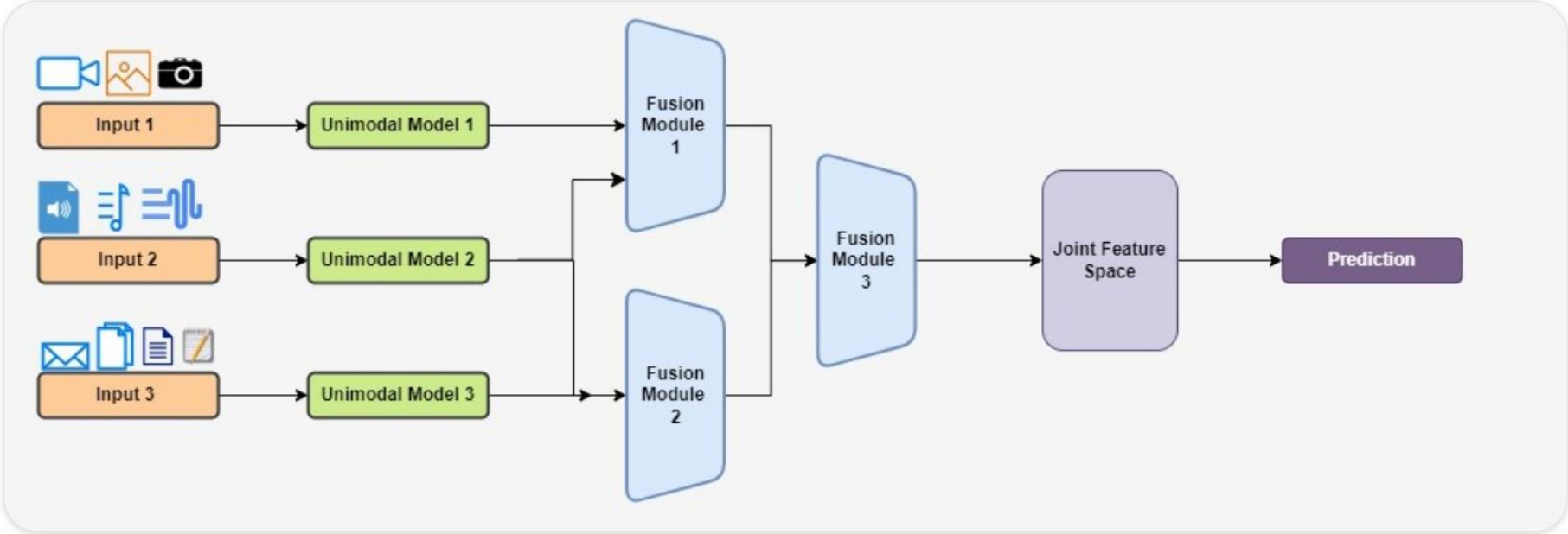


Аудио. Векторизация и токенизация

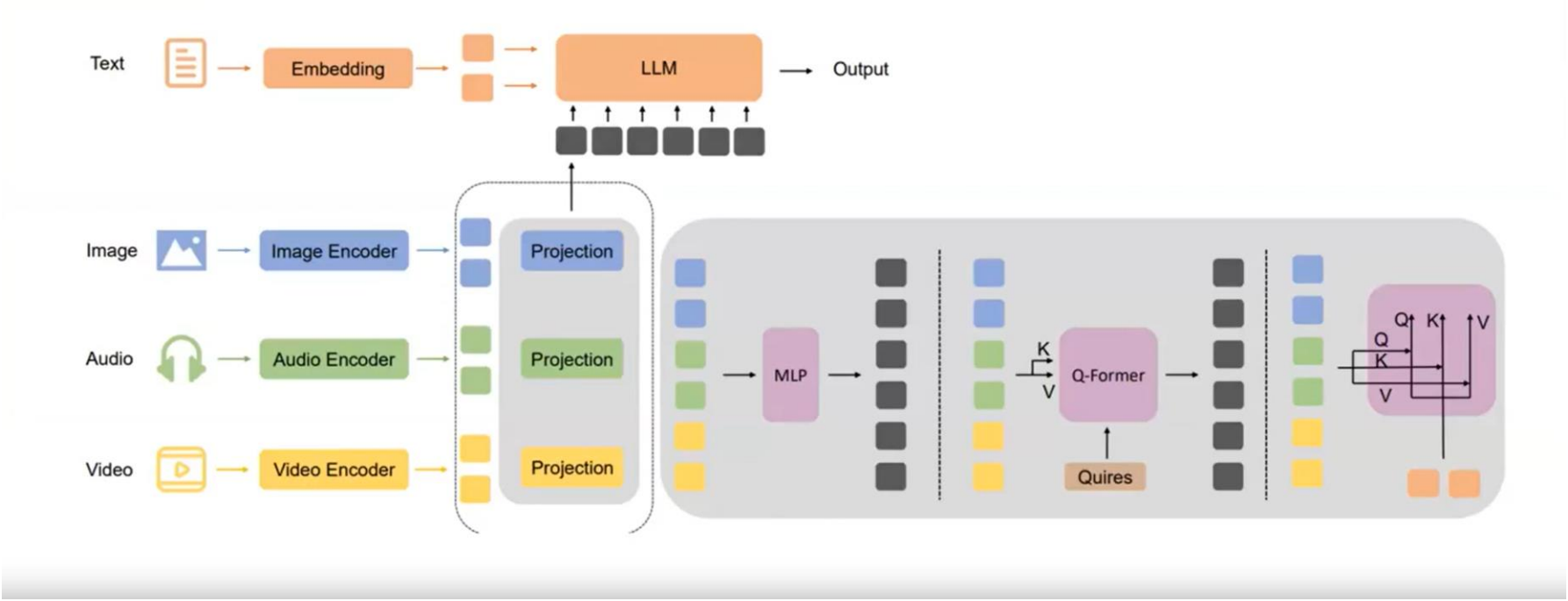
- Wav2vec
- HuBERT
- RVQ (Encodec, Descript, etc)



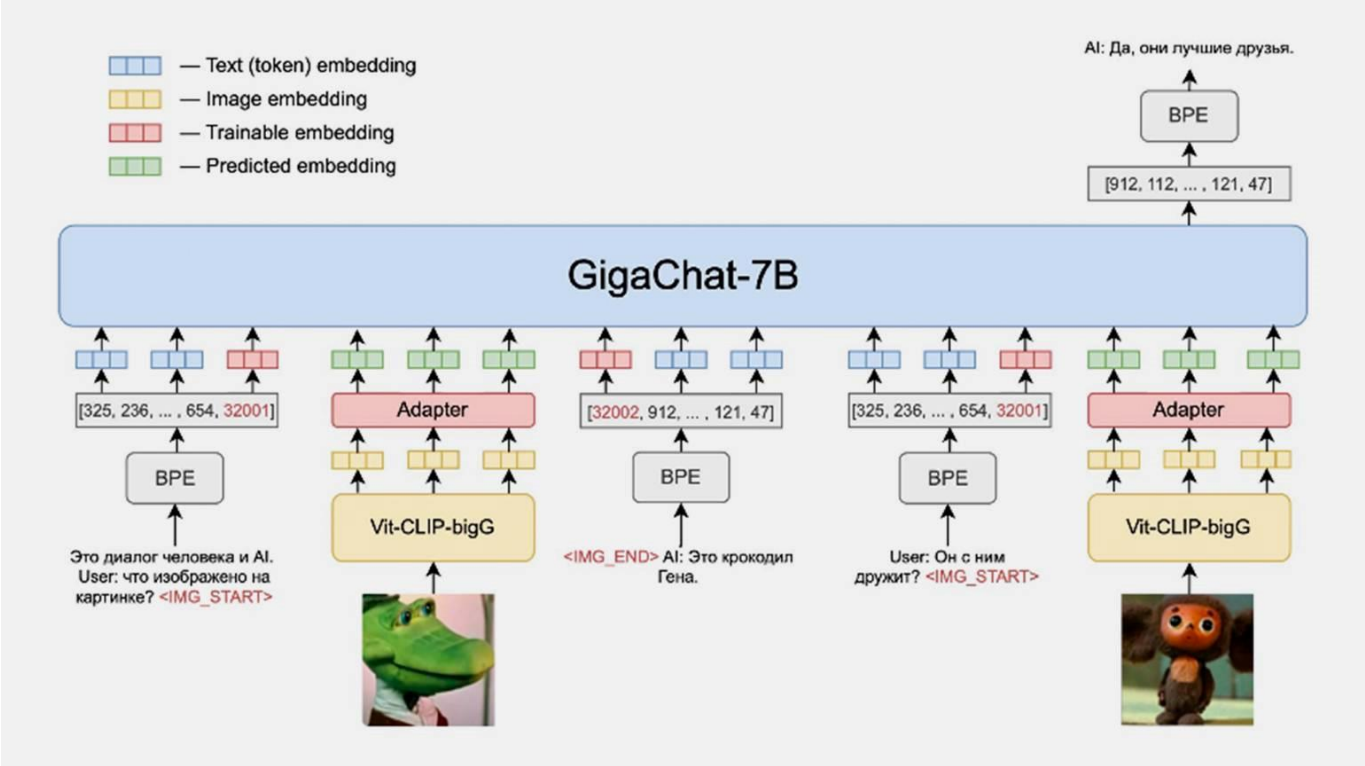
Объединение модальностей. Фьюжн



Объединение модальностей. Примеры адаптеров



Объединение модальностей. Примеры адаптеров



Видео

- Высокая частоста кадров
- Синхронизация между модальностями

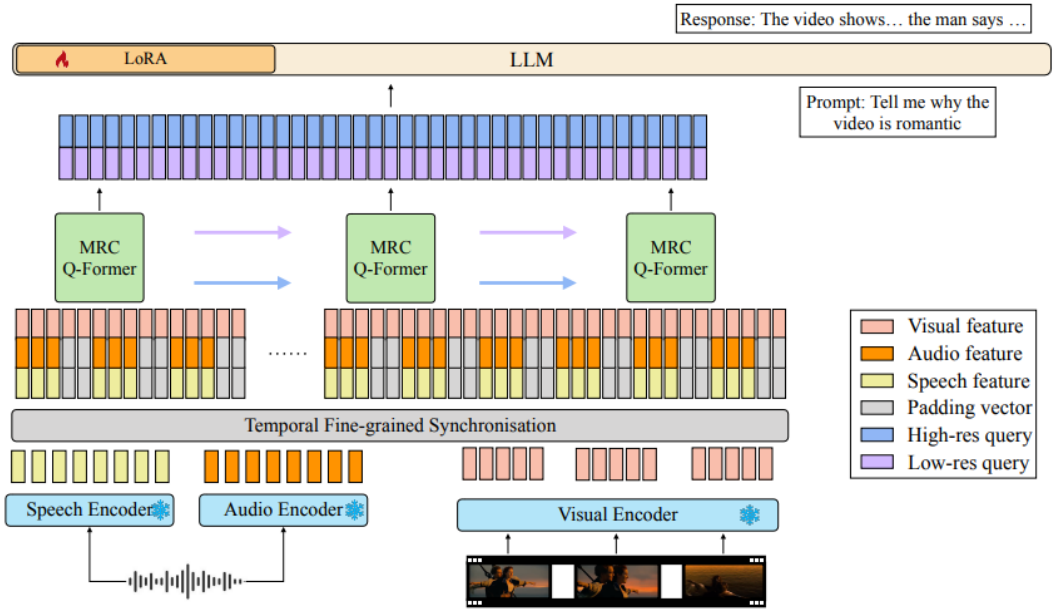


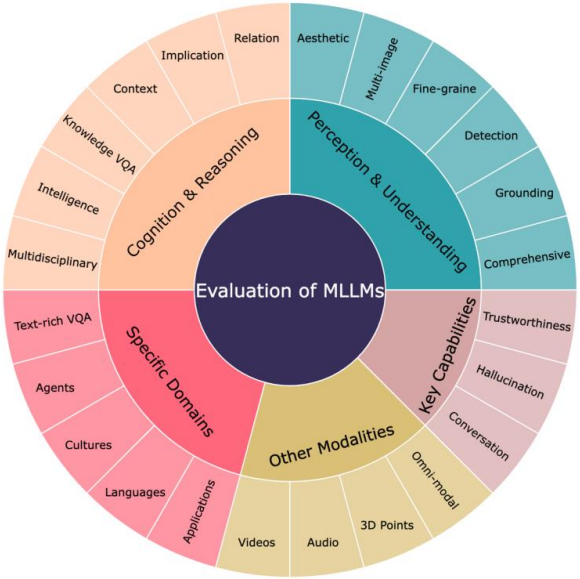
Table 4: Summary of LLM-based multimodal models with diverse modalities.

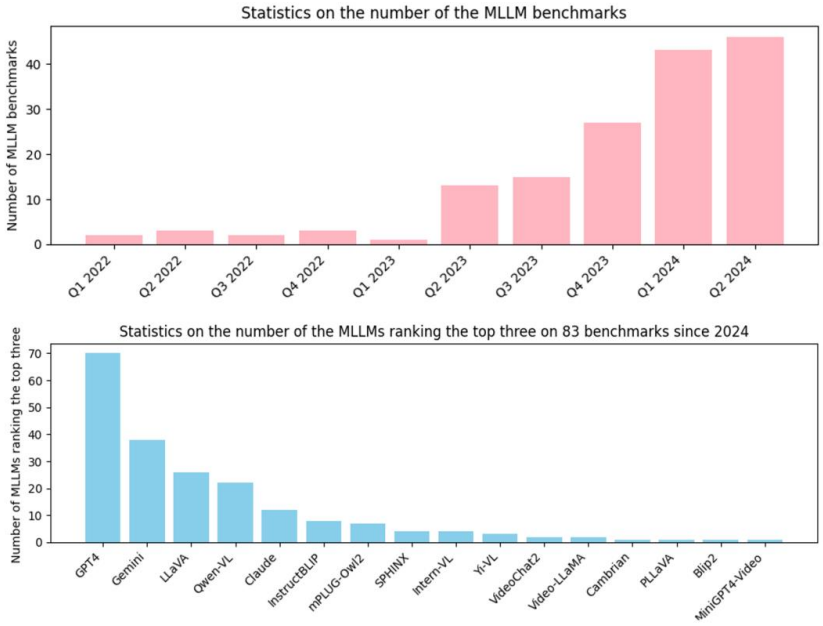
Model	Year	Modality	Used LLM	Training Modules
MiniGPT-4 [259]	2023	Text, Image	Vicuna	Two-stage training: Stage 1: Freeze visual feature extractor, train projection layer to align visual features with Vicuna; Stage 2: Instruction finetuning on dialogue data
Qwen-VL [14]	2023	Text, Image	Qwen-7B	Stage 1: Image caption generation; Stage 2: Multi-task pretraining; Stage 3: Supervised finetuning
BLIP-2 [95]	2023	Text, Image	OPT, FlanT5	Stage 1: Vision-Language Representation Learning; Stage 2: Vision-to-Language Generation Learning
LLaVA [113]	2023	Text, Image	GPT-3, GPT-3.5, LLaMA	Visual Instruction Tuning
LaVIN [118]	2023	Text, Image	LLaMA	Fine-tuning with MoE adapter
MiniGPT-v2 [32]	2023	Text, Image	Vicuna (7B/13B)	Multitask learning
InstructBLIP [40]	2023	Text, Image	Vicuna (7B/13B)	Visual Instruction Tuning
InternLM-XComposer [242]	2023	Text, Image	InternLM-Chat-7B	Pre-training, Multi-task Training, Instruction Fine-tuning
Macaw-LLM [119]	2023	Text, Image, Audio, 3D	LLaMA	Multimodal language modeling with unified representation
3D-MMLM [66]	2024	Text, Image, 3D	LLaMA-3	3D understanding, point cloud processing, cross-modal alignment
Qwen2-VL [201]	2024	Text, Image	Qwen-2	Visual Instruction Tuning
Moshi [52]	2024	Text, Audio	LLaMA-2	Text-to-speech, speech-to-text, audio understanding
MM-LLMs [237]	2024	Text, Image, Video, Audio, 3D	Various	Handling different modalities where X can be image, video, audio, 3D, etc.
AudioPaLM [154]	2023	Text, Audio	PaLM-2	Speech recognition, speech synthesis, multilingual audio understanding
Qwen-Audio [38]	2024	Text, Audio	Qwen-1.5	Audio instruction tuning, audio-text alignment
Yi-VL [226]	2024	Text, Image	Yi-Chat	Three-stage training: 1. Train ViT and projection module; 2. Increase image resolution and retrain; 3. Train entire model
InternLM-XComposer-2.5 [243]	2024	Text, Image	InternLM2-7B	Pre-training, Multi-task Training, Instruction Fine-tuning
CogVLM [203]	2024	Text, Image	LLaMA-2	Pre-training + Supervised Fine-tuning on vision-language tasks
ViLA [205]	2024	Text, Image	Supports Frozen and Finetuned (LoRA) usage of LLM	Distillation loss; Visual Question Answering loss
LLaVA-Video [251]	2025	Text, Image, Video	LLaMA-2	Video instruction tuning, video question answering
Qwen2.5-Omni [217]	2025	Text, Image, Audio, Video, 3D	Qwen-2.5	Video and audio branches with cross-attention for multimodal understanding

Мультимодальная генерация

- Токенизация всех модальностей
- Вызов отдельных моделей через функции

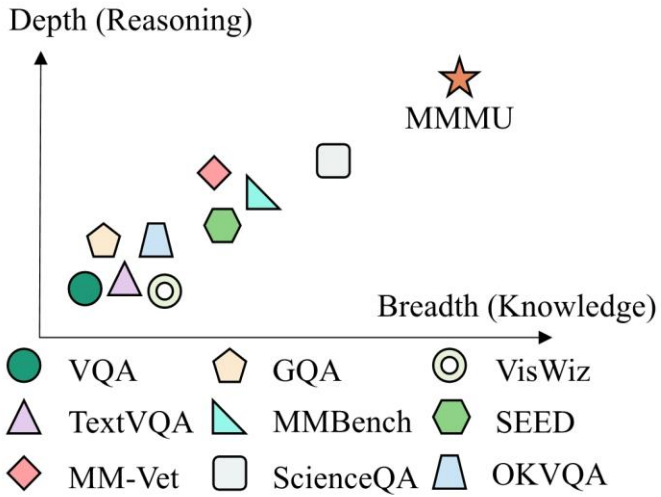
Мультимодальные бенчмарки





https://arxiv.org/pdf/2408.08632

Мультимодальные бенчмарки. MMMU



Dataset	Size	Images	Format	Source	Answer
VQA	> 1M	V	I+T	Annotated	Open
GQA	> 1M	V	I+T	Synthesized	Open
VisWiz	32K	V	I+T	Annotated	Open
TextVQA	45K	OC	I+T	Annotated	MC
OKVQA	14K	V+OC	I+T	Annotated	Open
SEED	19K	V+OC	I+T	Annotated	MC
MMBench	3K	V+OC	I+T	Repurposed	MC
MM-Vet	0.2K	V+OC	I+T	Repurposed	MC
ScienceQA	6K	5 Types	I+T	Textbooks	MC
MMMU	11.5K	32 Types	Interleaved	Textbooks, Internet, Annotated	Open / MC

https://mmmu-benchmark.github.io