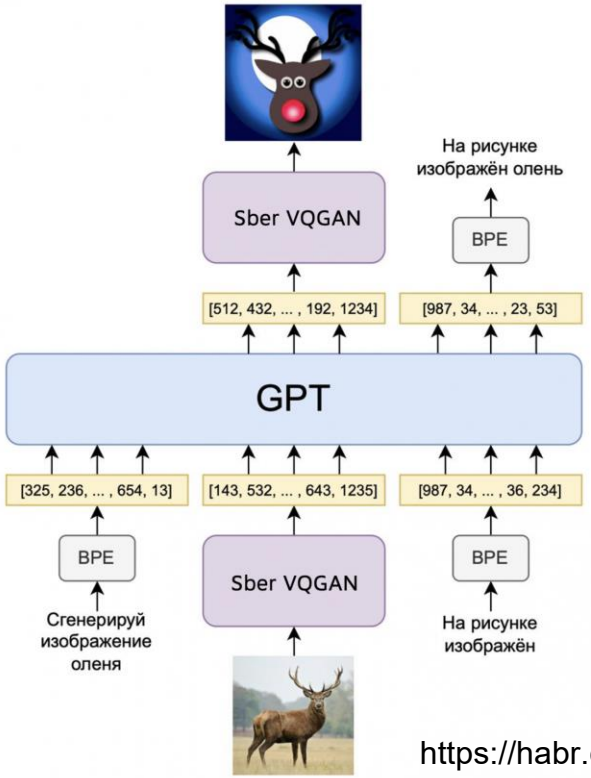
Multimodal LLMs (MLLM)

Мультимодальные LLM. Задачи

- Понимание данных разных модальностей
- Вопрос-ответные системы
- Инструктивное редактирование и генерация
- Продвинутые агенты

RUDOLPH



https://habr.com/ru/companies/sberbank/articles/733470/

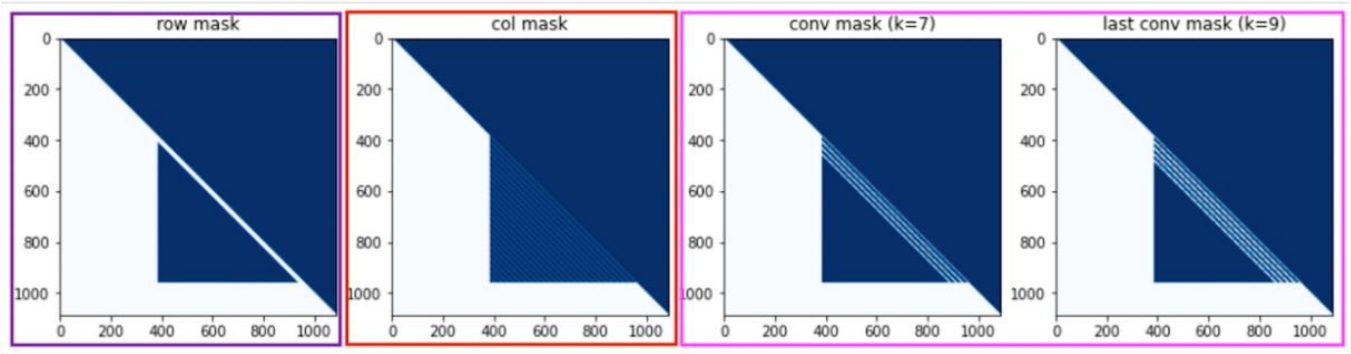
RUDOLPH

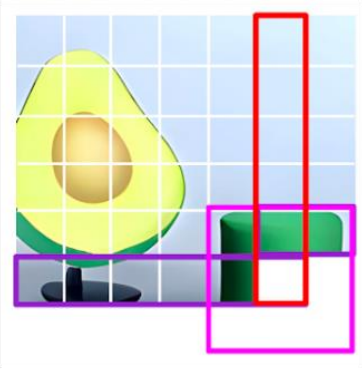

Кодировщик изображения:
Enc Sber VQGAN
24x24
384
128
Левый текст
Правый текст
Изображение (последовательность токенов)
Декодировщик изображения:
Dec Sber VQGAN / Image Diffusion

RUDOLPH

Количество параметров	**350M**	**1.3B**	**2.7B**
Параметры последовательности токенов			
Число левых токенов	64	128	384
Число токенов изображения	256	1024	576
Число правых токенов	64	128	128
Параметры архитектуры			
Количество скрытых слоев	24	24	32
Размер скрытого слоя	1024	2048	2560
Количество голов внимания	16	16	32

RUDOLPH


row mask
col mask
conv mask (k=7)
last conv mask (k=9)
0
200
400
600
800
1000



RUDOLPH

<LT_T2I> — *для задачи генерации изображения по тексту.*
<LT_I2T> — *для задачи генерации текстового описания изображения.*
<LT_T2T> — *для задачи языкового моделирования (LM).*
<RT_I2T> — *для задачи генерации текстового описания по изображению.*

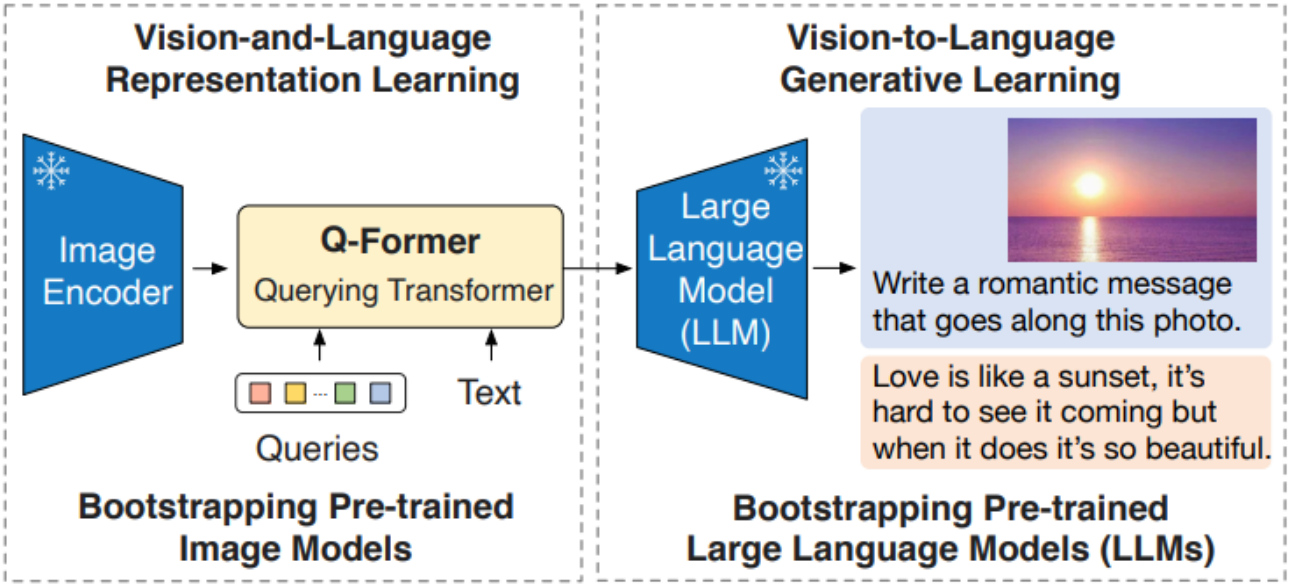
RUDOLPH. Задачи

- описание изображения (image captioning *i2t*)
- генерация изображения по тексту (image generation *t2i*)
- генерация текста в левых токенах (language modeling *t2t*)

RUDOLPH. Finetuning

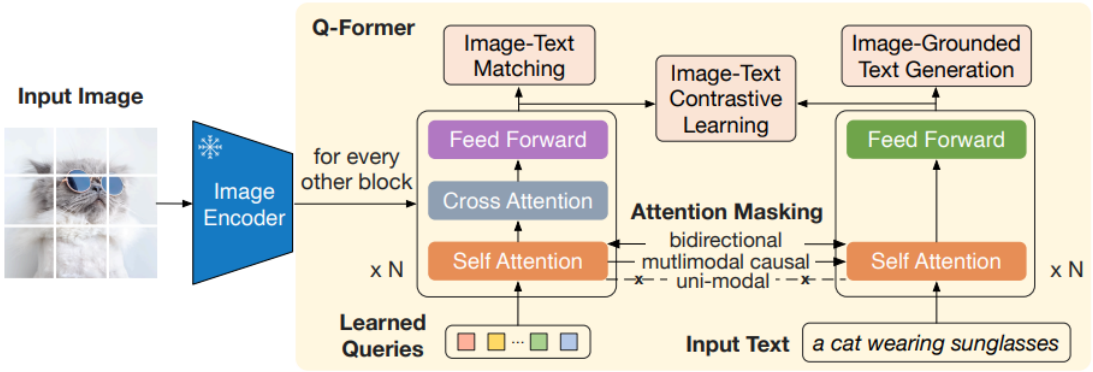
1. Понимание прочитанного текста (Text QA)
2. Решение математических задач (Mathematical QA)
3. Генерация изображения по тексту (Image Generation)
4. Описание изображения (Image Captioning)
5. Ответ на вопросы по изображению (Visual QA)
6. Распознавание текста на изображении (Text Recognition in the Wild)

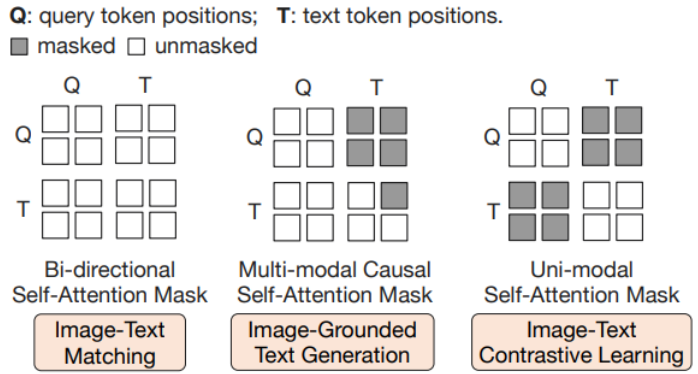
BLIP-2



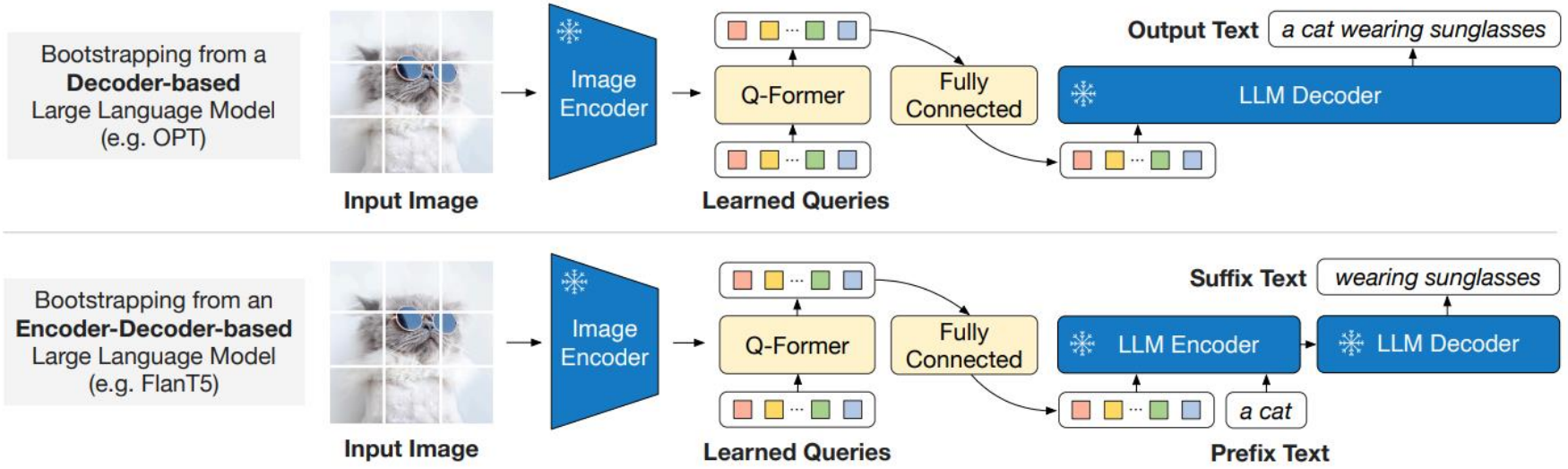
https://arxiv.org/pdf/2301.12597

BLIP-2


Input Image
Image Encoder
Q-Former
for every other block
x N
Image-Text Matching
Image-Text Contrastive Learning
Image-Grounded Text Generation
Feed Forward
Cross Attention
Self Attention
Attention Masking
bidirectional
mutlimodal causal
uni-modal
Feed Forward
Self Attention
x N
Learned Queries
Input Text
a cat wearing sunglasses


Q: query token positions; T: text token positions.
masked
unmasked
Q
T
Bi-directional Self-Attention Mask
Image-Text Matching
Multi-modal Causal Self-Attention Mask
Image-Grounded Text Generation
Uni-modal Self-Attention Mask
Image-Text Contrastive Learning

BLIP-2


Bootstrapping from a
Decoder-based
Large Language Model
(e.g. OPT)
Input Image
Image Encoder
Learned Queries
Q-Former
Fully Connected
LLM Decoder
Output Text
a cat wearing sunglasses
Bootstrapping from an
Encoder-Decoder-based
Large Language Model
(e.g. FlanT5)
Input Image
Image Encoder
Learned Queries
Q-Former
Fully Connected
LLM Encoder
LLM Decoder
Suffix Text
wearing sunglasses
a cat
Prefix Text

BLIP-2

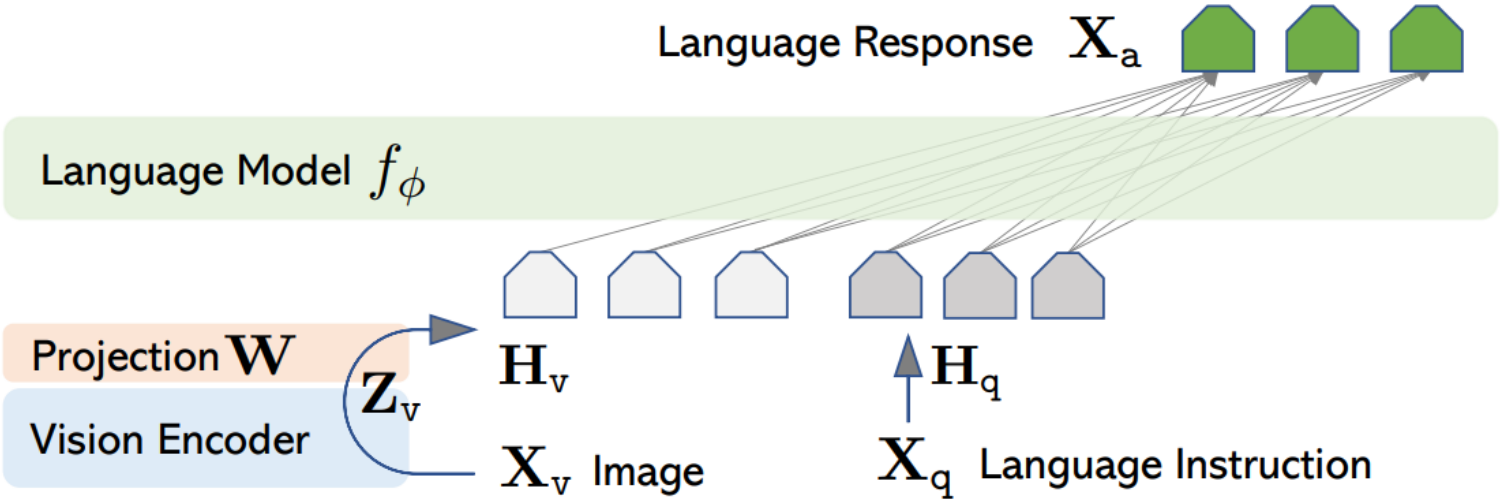
Models	#Trainable Params	#Total Params	VQAv2 val	VQAv2 test-dev	OK-VQA test	GQA test-dev
VL-T5$_{\text{no-vqa}}$	224M	269M	13.5	-	5.8	6.3
FewVLM (Jin et al., 2022)	740M	785M	47.7	-	16.5	29.3
Frozen (Tsimpoukelli et al., 2021)	40M	7.1B	29.6	-	5.9	-
VLKD (Dai et al., 2022)	406M	832M	42.6	44.5	13.3	-
Flamingo3B (Alayrac et al., 2022)	1.4B	3.2B	-	49.2	41.2	-
Flamingo9B (Alayrac et al., 2022)	1.8B	9.3B	-	51.8	44.7	-
Flamingo80B (Alayrac et al., 2022)	10.2B	80B	-	56.3	**50.6**	-
BLIP-2 ViT-L OPT$_{2.7B}$	104M	3.1B	50.1	49.7	30.2	33.9
BLIP-2 ViT-g OPT$_{2.7B}$	107M	3.8B	53.5	52.3	31.7	34.6
BLIP-2 ViT-g OPT$_{6.7B}$	108M	7.8B	54.3	52.6	36.4	36.4
BLIP-2 ViT-L FlanT5$_{XL}$	103M	3.4B	62.6	62.3	39.4	44.4
BLIP-2 ViT-g FlanT5$_{XL}$	107M	4.1B	63.1	63.0	40.7	44.2
BLIP-2 ViT-g FlanT5$_{XXL}$	108M	12.1B	**65.2**	**65.0**	45.9	**44.7**

BLIP-2

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

Model	#Trainable Params	Flickr30K Zero-shot (1K test set)						COCO Fine-tuned (5K test set)					
		Image → Text			Text → Image			Image → Text			Text → Image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Dual-encoder models													
CLIP (Radford et al., 2021)	428M	88.0	98.7	99.4	68.7	90.6	95.2	-	-	-	-	-	-
ALIGN (Jia et al., 2021)	820M	88.6	98.7	99.7	75.7	93.8	96.8	77.0	93.5	96.9	59.9	83.3	89.8
FILIP (Yao et al., 2022)	417M	89.8	99.2	99.8	75.0	93.4	96.3	78.9	94.4	97.4	61.2	84.3	90.6
Florence (Yuan et al., 2021)	893M	90.9	99.1	-	76.7	93.6	-	81.8	95.2	-	63.2	85.7	-
BEIT-3(Wang et al., 2022b)	1.9B	94.9	99.9	**100.0**	81.5	95.6	97.8	84.8	96.5	98.3	67.2	**87.7**	**92.8**
Fusion-encoder models													
UNITER (Chen et al., 2020)	303M	83.6	95.7	97.7	68.7	89.2	93.9	65.7	88.6	93.8	52.9	79.9	88.0
OSCAR (Li et al., 2020)	345M	-	-	-	-	-	-	70.0	91.1	95.5	54.0	80.8	88.5
VinVL (Zhang et al., 2021)	345M	-	-	-	-	-	-	75.4	92.9	96.2	58.8	83.5	90.3
Dual encoder + Fusion encoder reranking													
ALBEF (Li et al., 2021)	233M	94.1	99.5	99.7	82.8	96.3	98.1	77.6	94.3	97.2	60.7	84.3	90.5
BLIP (Li et al., 2022)	446M	96.7	**100.0**	**100.0**	86.7	97.3	98.7	82.4	95.4	97.9	65.1	86.3	91.8
BLIP-2 ViT-L	474M	96.9	**100.0**	**100.0**	88.6	97.6	**98.9**	83.5	96.0	98.0	66.3	86.5	91.8
BLIP-2 ViT-g	1.2B	**97.6**	**100.0**	**100.0**	**89.7**	**98.1**	**98.9**	**85.4**	**97.0**	**98.5**	**68.3**	**87.7**	92.6

LLaVa



$$\mathbf{H}_v = \mathbf{W} \cdot \mathbf{Z}_v, \text{ with } \mathbf{Z}_v = g(\mathbf{X}_v)$$

https://arxiv.org/pdf/2304.08485

LLaVa

$$\mathbf{X}_{\texttt{instruct}}^{t} = \begin{cases} \text{Randomly choose } [\mathbf{X}_{\texttt{q}}^{1}, \mathbf{X}_{\texttt{v}}] \text{ or } [\mathbf{X}_{\texttt{v}}, \mathbf{X}_{\texttt{q}}^{1}], & \text{the first turn } t = 1 \\ \mathbf{X}_{\texttt{q}}^{t}, & \text{the remaining turns } t > 1 \end{cases}$$

$$p(\mathbf{X}_{\texttt{a}} | \mathbf{X}_{\texttt{v}}, \mathbf{X}_{\texttt{instruct}}) = \prod_{i=1}^{L} p_{\boldsymbol{\theta}}(\boldsymbol{x}_i | \mathbf{X}_{\texttt{v}}, \mathbf{X}_{\texttt{instruct},<i}, \mathbf{X}_{\texttt{a},<i}),$$

$\mathbf{X}_{\texttt{system-message}}$ <STOP>
Human : $\mathbf{X}_{\texttt{instruct}}^{1}$ <STOP> Assistant: $\mathbf{X}_{\texttt{a}}^{1}$ <STOP>
Human : $\mathbf{X}_{\texttt{instruct}}^{2}$ <STOP> Assistant: $\mathbf{X}_{\texttt{a}}^{2}$ <STOP> ...

LLaVa

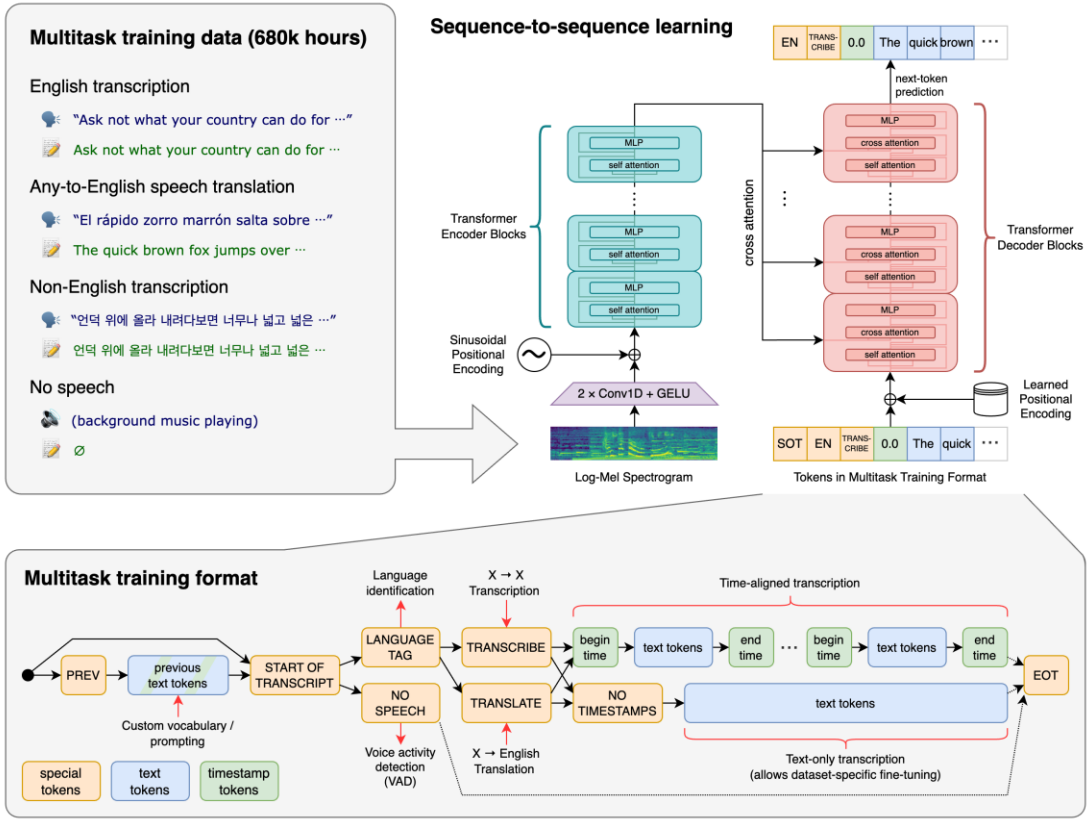
Претрейн: обучение проекции на CC3M
FT E2E: разморозка LLM и проекций на 2х сценариях Чатбот и Science QA

Особенности: уменьшение LR на второй стадии с 2e-2 до 2e-5

LLaVa

Visual input example, Extreme Ironing:	
	
	Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg
User	What is unusual about this image?
LLaVA	The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.
	[*Start a new conversation, and clear the history*]
User	What's happening in the scene?
LLaVA	The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.
User	What is unusual about this image?
GPT-4 [36]	The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.
User	What is unusual about this image?
BLIP-2	a man is sitting on the back of a yellow cab
User	What is unusual about this image?
OpenFlamingo	The man is drying his clothes on the hood of his car.

Whisper


Multitask training data (680k hours)
English transcription
"Ask not what your country can do for ···"
Ask not what your country can do for ···
Any-to-English speech translation
"El rápido zorro marrón salta sobre ···"
The quick brown fox jumps over ···
Non-English transcription
"언덕 위에 올라 내려다보면 너무나 넓고 넓은 ···"
언덕 위에 올라 내려다보면 너무나 넓고 넓은 ···
No speech
(background music playing)
Ø
Sequence-to-sequence learning
Transformer Encoder Blocks
MLP
self attention
Sinusoidal Positional Encoding
2 × Conv1D + GELU
Log-Mel Spectrogram
cross attention
Transformer Decoder Blocks
next-token prediction
EN
TRANS-CRIBE
0.0
The
quick
brown
SOT
Learned Positional Encoding
Tokens in Multitask Training Format
Multitask training format
PREV
previous text tokens
Custom vocabulary / prompting
START OF TRANSCRIPT
LANGUAGE TAG
Language identification
NO SPEECH
Voice activity detection (VAD)
TRANSCRIBE
X → X Transcription
TRANSLATE
X → English Translation
Time-aligned transcription
begin time
text tokens
end time
NO TIMESTAMPS
EOT
Text-only transcription (allows dataset-specific fine-tuning)
special tokens
text tokens
timestamp tokens

AudioLLaMA

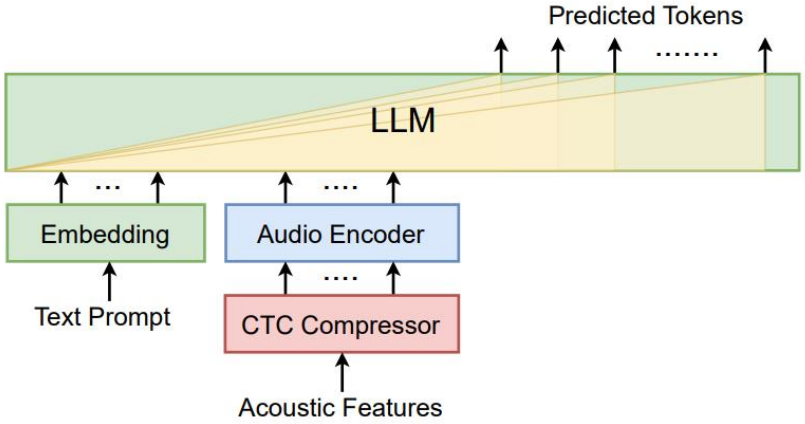


Fig. 1. High-level architecture of our proposed approach with LLM. The green blocks indicate the part of the LLM. In this work, we only learn parameters in the "Audio Encoder", keeping everything else frozen.

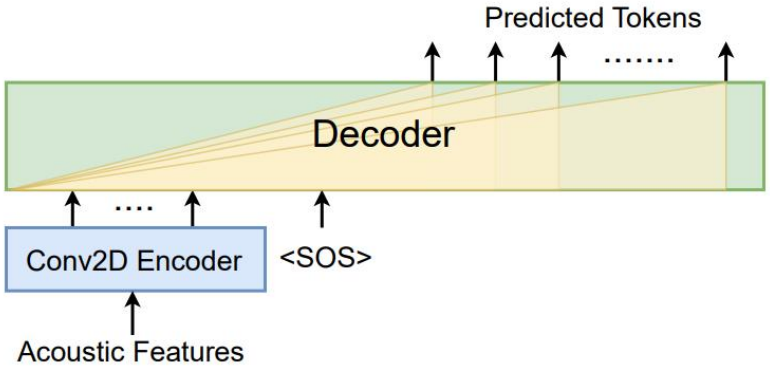
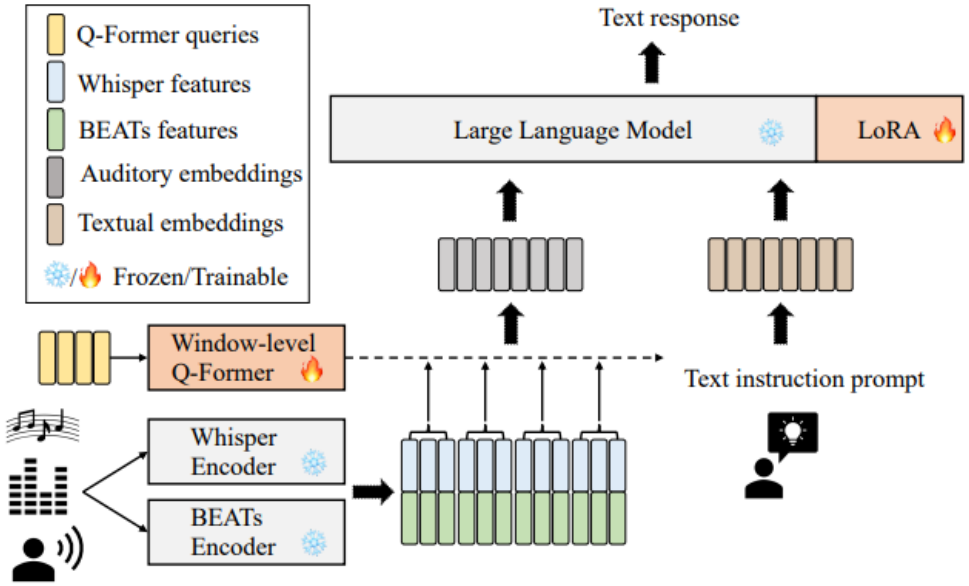


Fig. 2. The architecture of the decoder-only model for the from-scratch training. We use ⟨SOS⟩ token to indicate the starting of the text generation.

https://arxiv.org/pdf/2307.03917

SALMONN



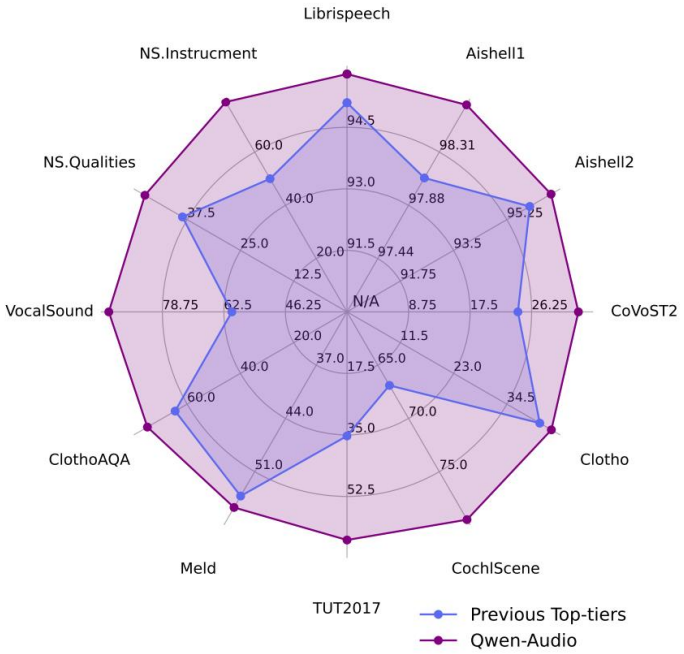
https://arxiv.org/pdf/2310.13289

SALMONN

Претрейн: обучение проекции при замороженных энкодерах
Инструктивный FT: текстовые промпты + аудиоданные
Activation Tuning: настройка коэффициента масштабирования LoRA

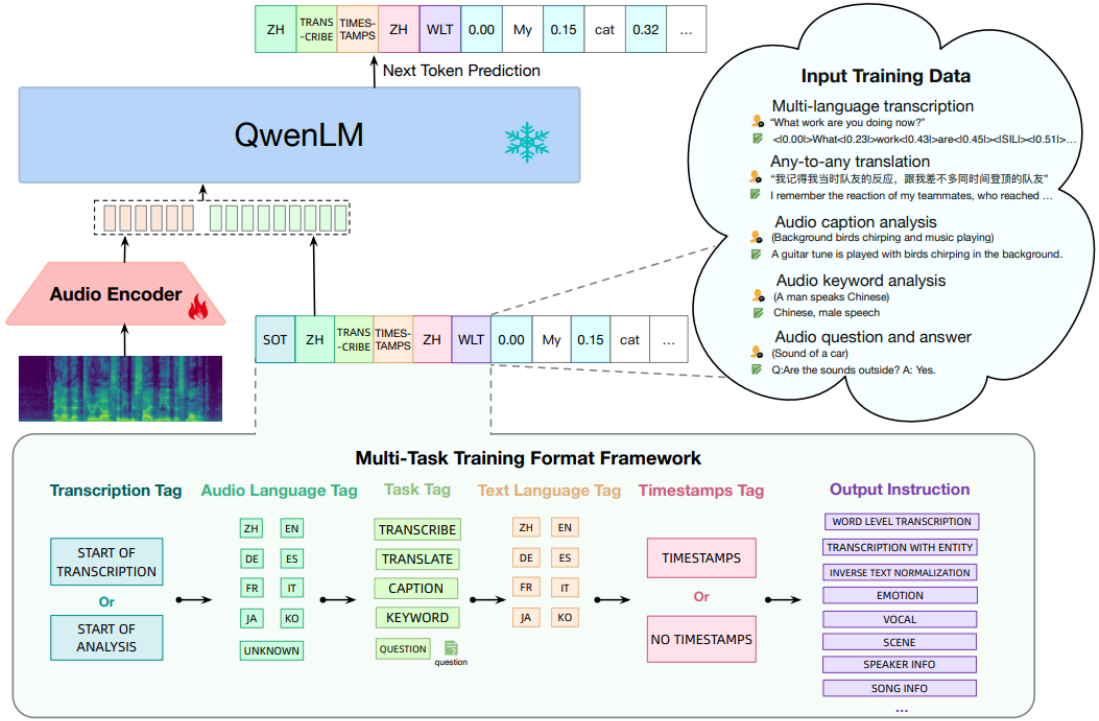
Особенности: Использование 2х энкодеров для извлечения речевых фичей и общих аудиофичей

QWEN-Audio



https://arxiv.org/pdf/2311.07919

QWEN-Audio


ZH
TRANS-CRIBE
TIMES-TAMPS
ZH
WLT
0.00
My
0.15
cat
0.32
...
Next Token Prediction
QwenLM
Audio Encoder
SOT
ZH
TRANS-CRIBE
TIMES-TAMPS
ZH
WLT
0.00
My
0.15
cat
...
Input Training Data
Multi-language transcription
"What work are you doing now?"
<|0.00|>What<|0.23|>work<|0.43|>are<|0.45|><|SILI|><|0.51|>...
Any-to-any translation
"我记得我当时队友的反应，跟我差不多同时间登顶的队友"
I remember the reaction of my teammates, who reached ...
Audio caption analysis
(Background birds chirping and music playing)
A guitar tune is played with birds chirping in the background.
Audio keyword analysis
(A man speaks Chinese)
Chinese, male speech
Audio question and answer
(Sound of a car)
Q:Are the sounds outside? A: Yes.
Multi-Task Training Format Framework
Transcription Tag
Audio Language Tag
Task Tag
Text Language Tag
Timestamps Tag
Output Instruction
START OF TRANSCRIPTION
Or
START OF ANALYSIS
ZH
EN
DE
ES
FR
IT
JA
KO
UNKNOWN
TRANSCRIBE
TRANSLATE
CAPTION
KEYWORD
QUESTION
question
ZH
EN
DE
ES
FR
IT
JA
KO
TIMESTAMPS
Or
NO TIMESTAMPS
WORD LEVEL TRANSCRIPTION
TRANSCRIPTION WITH ENTITY
INVERSE TEXT NORMALIZATION
EMOTION
VOCAL
SCENE
SPEAKER INFO
SONG INFO
...

QWEN-Audio

Types	Task	Description	Hours
Speech	ASR	Automatic speech recognition (multiple languages)	30k
	S2TT	Speech-to-text translation	3.7k
	OSR	Overlapped speech recognition	<1k
	Dialect ASR	Automatic dialect speech recognition	2k
	SRWT	English speech recognition with word-level timestamps	10k
		Mandarin speech recognition with word-level timestamps	11k
	DID	Dialect identification	2k
	LID	Spoken language identification	11.7k
	SGC	Speaker gender recognition (biologically)	4.8k
	ER	Emotion recognition	<1k
	SV	Speaker verification	1.2k
	SD	Speaker diarization	<1k
	SER	Speech entity recognition	<1k
	KS	Keyword spotting	<1k
	IC	Intent classification	<1k
	SF	Slot filling	<1k
	SAP	Speaker age prediction	4.8k
	VSC	Vocal sound classification	<1k
Sound	AAC	Automatic audio caption	8.4k
	SEC	Sound event classification	5.4k
	ASC	Acoustic scene classification	<1k
	SED	Sound event detection with timestamps	<1k
	AQA	Audio question answering	<1k
Music&Song	SID	Singer identification	<1k
	SMER	Singer and music emotion recognition	<1k
	MC	Music caption	25k
	MIC	Music instruments classification	<1k
	MNA	Music note analysis such as pitch, velocity	<1k
	MGR	Music genre recognition	9.5k
	MR	Music recognition	<1k
	MQA	Music question answering	<1k

QWEN-Audio

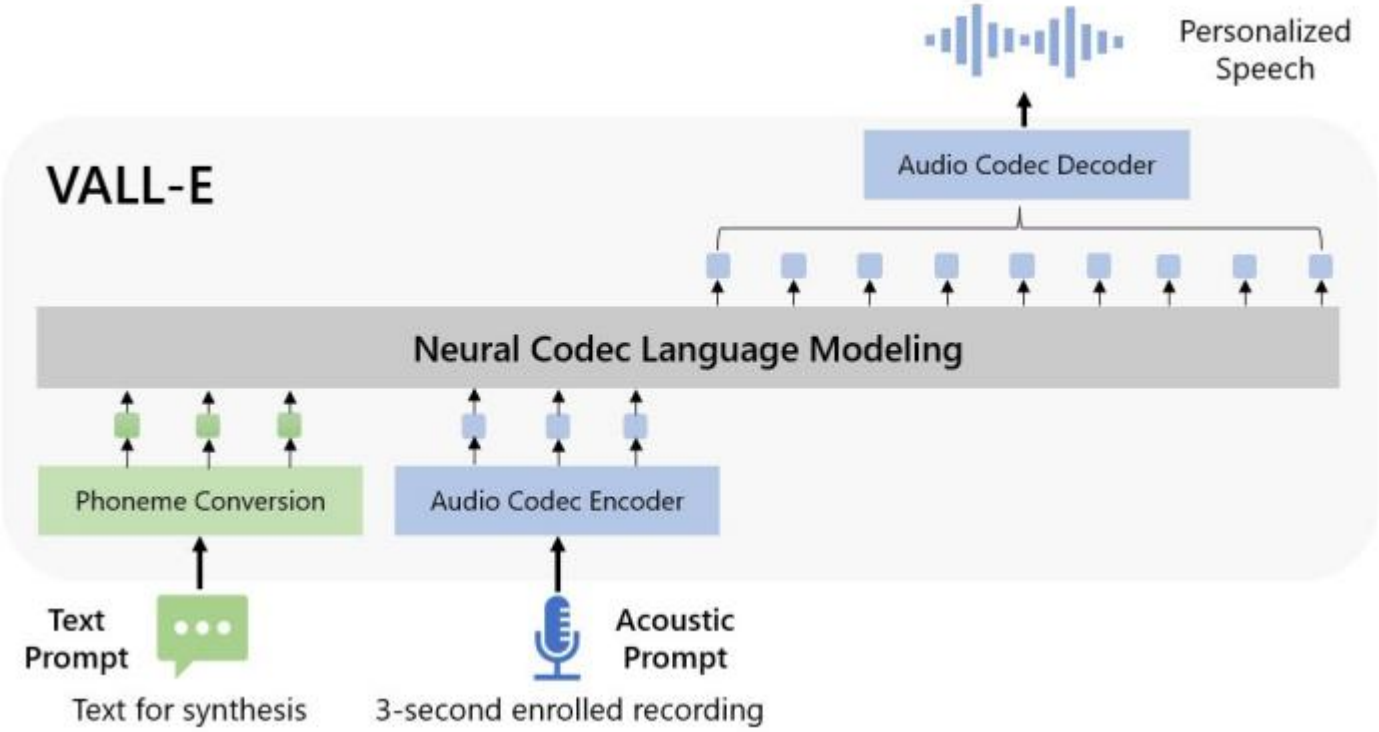
Претрейн: обучается только аудио энкодер
Инструктивный FT: дообучается LLM при замороженном аудиоэнкодере

QWEN-Audio

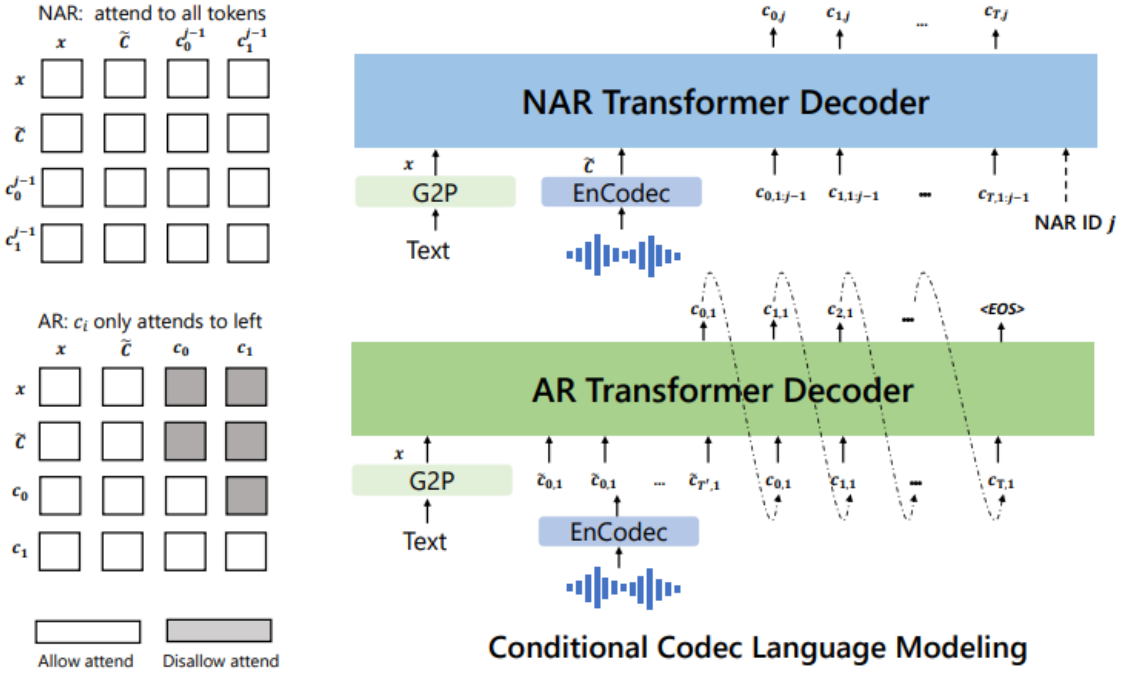
The Data Format Example of Supervised Fine-Tuning.

<im_start>user
Audio 1: <audio>emov-db/141-168-0155.wav</audio>what does the speaker say?<im_end>
<im_start>assistant
The speaker says in English, "Won't you draw up, gentlemen.".<im_end>
<im_start>user
What's the mood of the person?<im_end>
<im_start>assistant
Based on the voice, the mood of the person is disgusted.<im_end>

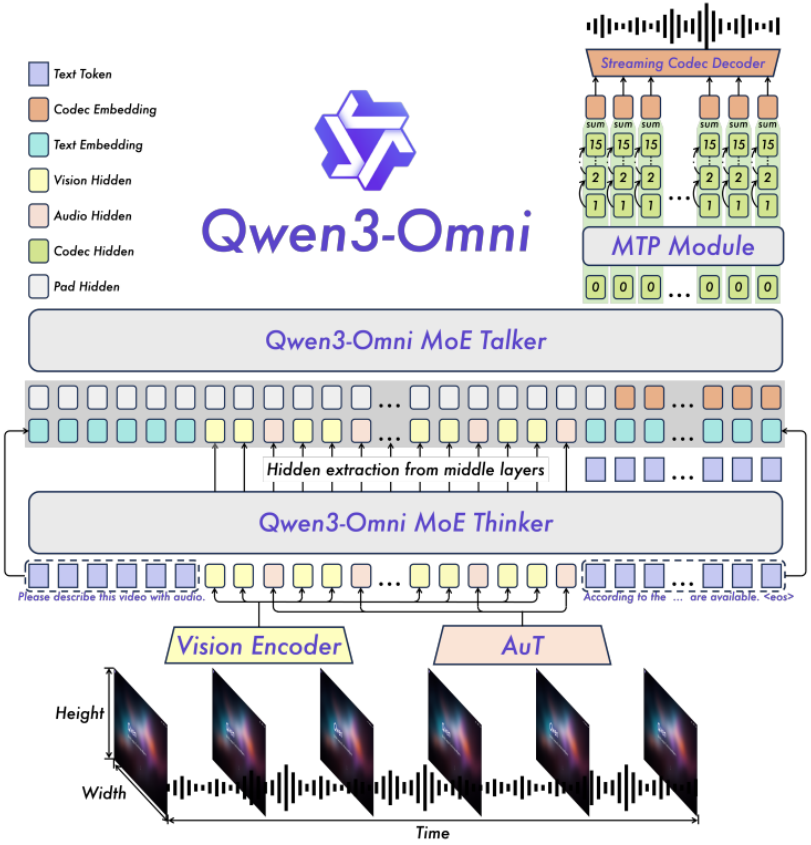
Генерация аудио. Vall-e



Генерация аудио. Vall-e



QWEN3-omni


Text Token
Codec Embedding
Text Embedding
Vision Hidden
Audio Hidden
Codec Hidden
Pad Hidden
Qwen3-Omni
Streaming Codec Decoder
MTP Module
Qwen3-Omni MoE Talker
Hidden extraction from middle layers
Qwen3-Omni MoE Thinker
Please describe this video with audio.
According to the ... are available. <eos>
Vision Encoder
AuT
Height
Width
Time

QWEN3-omni

- Разнообразные промпты под каждую задачу
- Многостадийное обучение под даунстрим задачи

Токены vs проекции

Проекции

+ Меньше потери информации
+ Низкая частота кадров
- Не подходит для генерации

Токены

+ Подходят и для инпута, и для аутпута
- Высокая частота кадров
- Потеря информации