

Audio-LLM

Павел Андреев

pavel.andreev@heygen.com

План

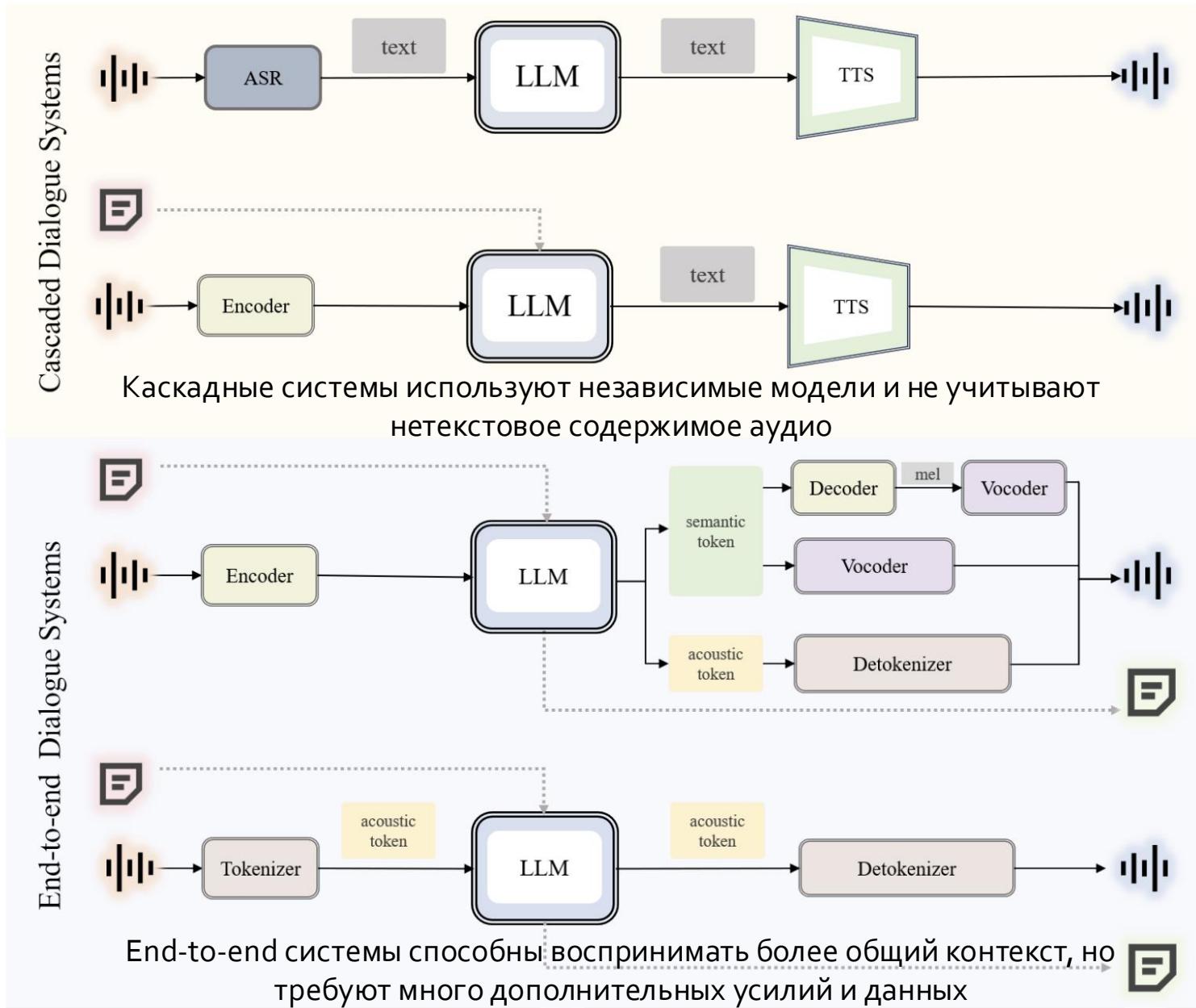
- Верхнеуровневое ревью подходов
- Представления речи
- Этапы обучения
- Дуплексные системы, интерактивность
- Ревью статьи: Moshi
- Ревью статьи: Qwen2.5-omni

Верхнеуров- невое ревью подходов

Чего мы хотим от
audio-LLM?

1. Мы хотим чтобы Audio-LLM могла делать все что делает обычная LLM
2. Мы хотим чтобы она могла принимать и анализировать аудио сигнал на вход
3. Мы хотим чтобы она могла генерировать аудио сигнал в ответ, имея ввиду аудио сигнал на входе, смысл и контекст ситуации
4. Иногда мы хотим чтобы она могла вести интерактивный диалог, то есть моделировать структуру диалога: понимать когда нужно начать говорить, когда заканчивать, а когда нет

Верхнеуров- невое ревью подходов

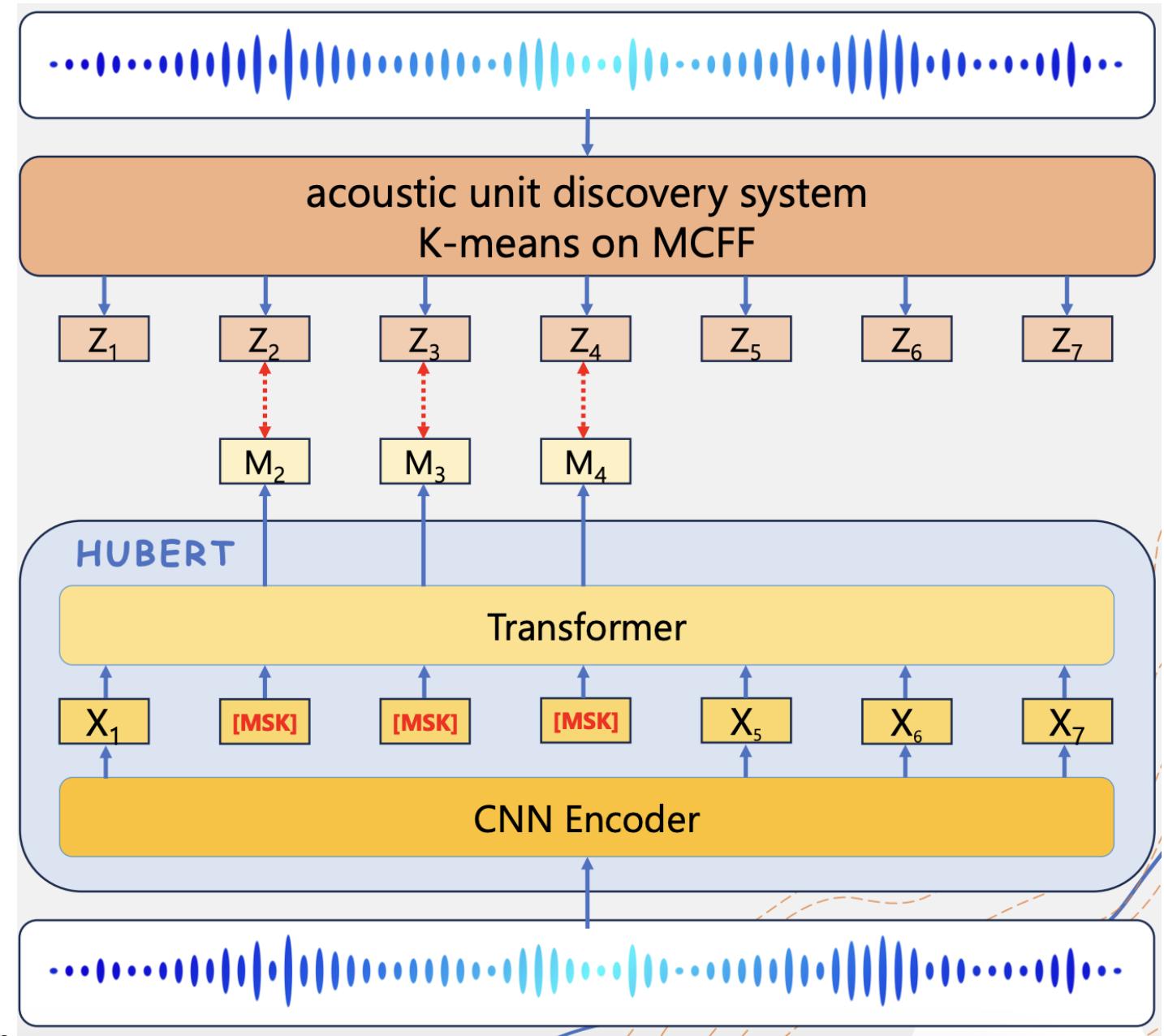


Представле- ния речи

- **Семантические** -> цель: иметь представления речи как можно более близкие к речи; обычно генерируются из self-supervised либо ASR моделями
- **Акустические** -> цель: иметь представления, из которых исходное аудио восстанавливается наиболее точно
- **Смешанные** -> цель: объединить акустические и семантические
- Чтобы LLM могла генерировать речь, ее представления должны быть **токенизуемыми** (дискретными)

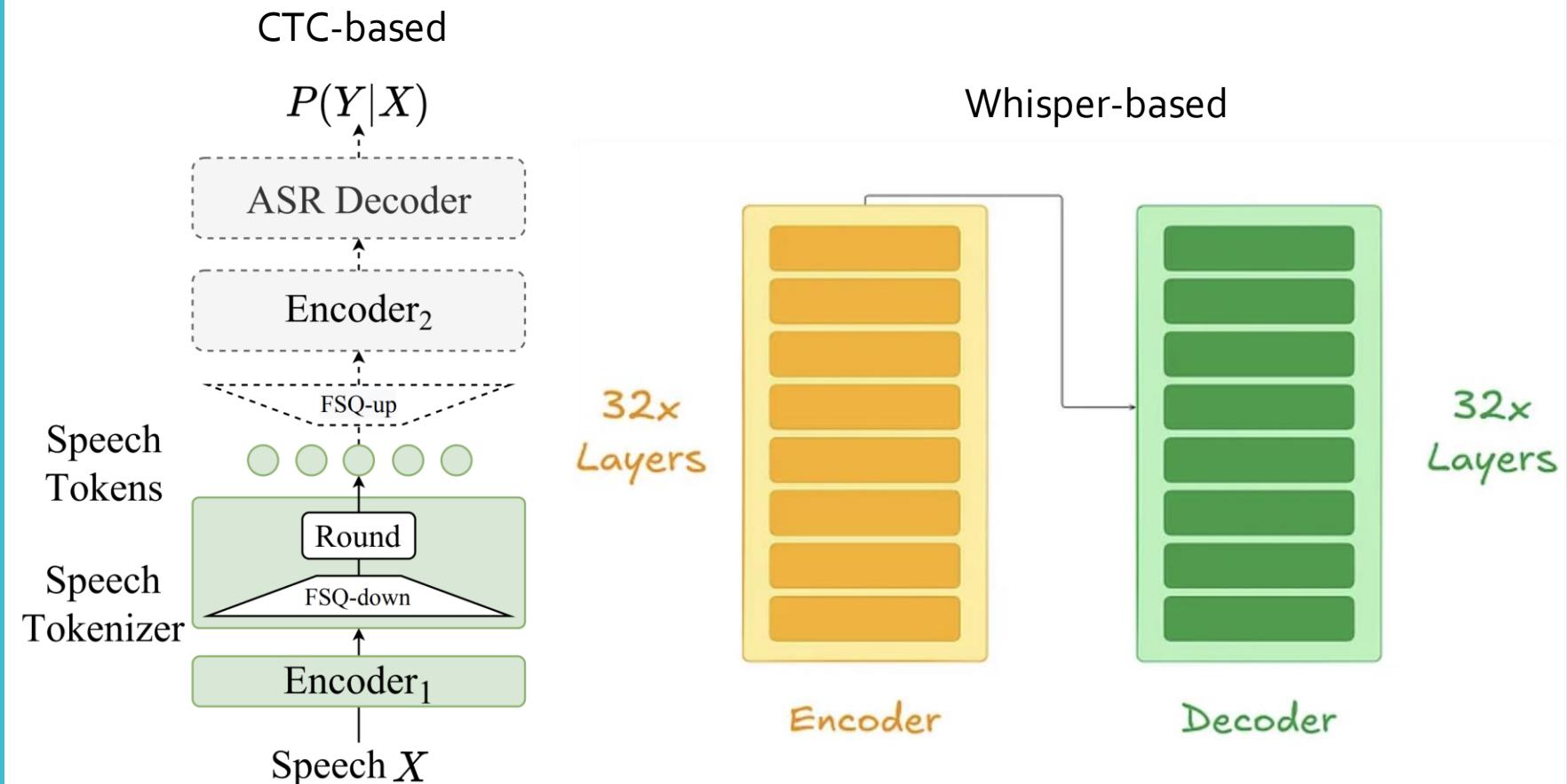
Представле- ния речи

Семантические self-
supervised
представления



Представле- ния речи

Семантические ASR представления



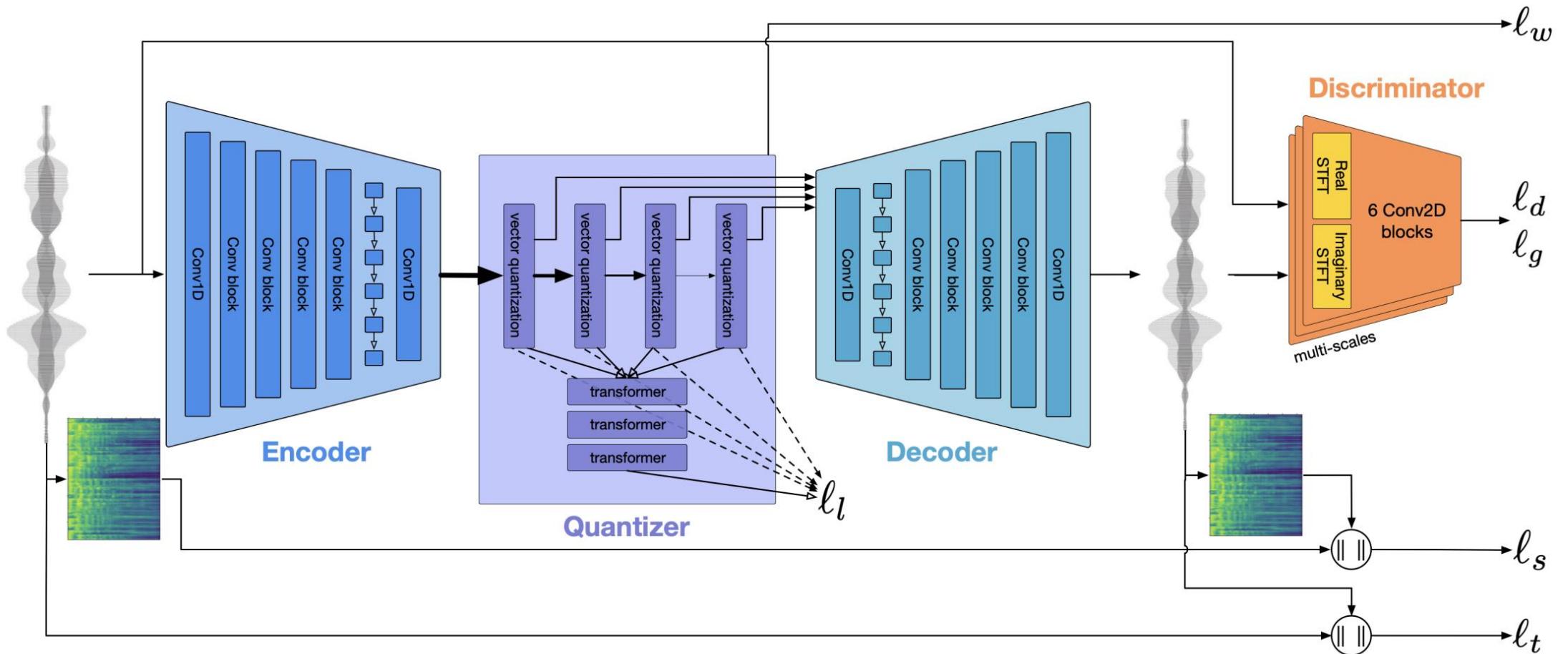
Hsu, Wei-Ning, et al. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units." *IEEE/ACM transactions on audio, speech, and language processing* 29 (2021): 3451-3460.

Cao, Nan, et al. "Whisper: Tracing the spatiotemporal process of information diffusion in real time." *IEEE transactions on visualization and computer graphics* 18.12 (2012): 2649-2658.

Представления речи

Акустические представления

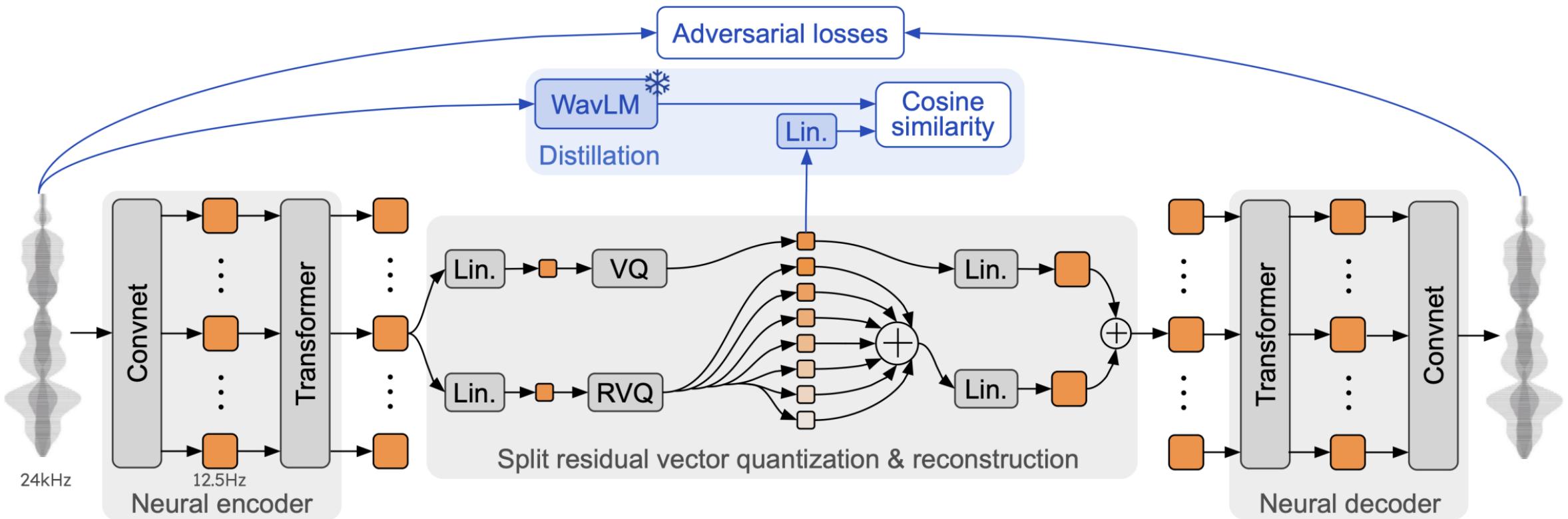
Défossez, Alexandre, et al. "High Fidelity Neural Audio Compression." *Transactions on Machine Learning Research*.



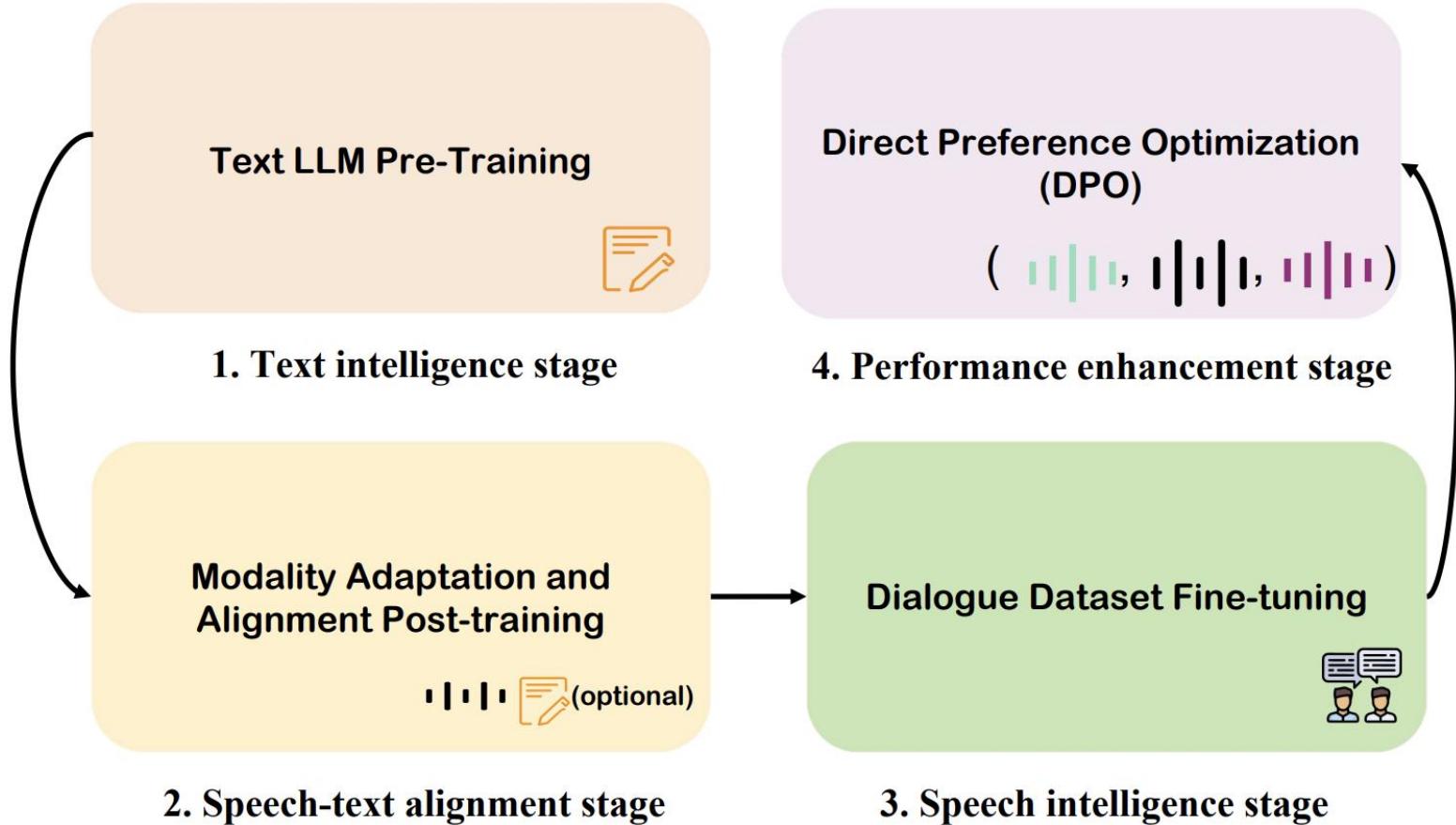
Представления речи

Смешанные представления

Défossez, Alexandre, et al. "Moshi: a speech-text foundation model for real-time dialogue." *arXiv preprint arXiv:2410.00037* (2024).

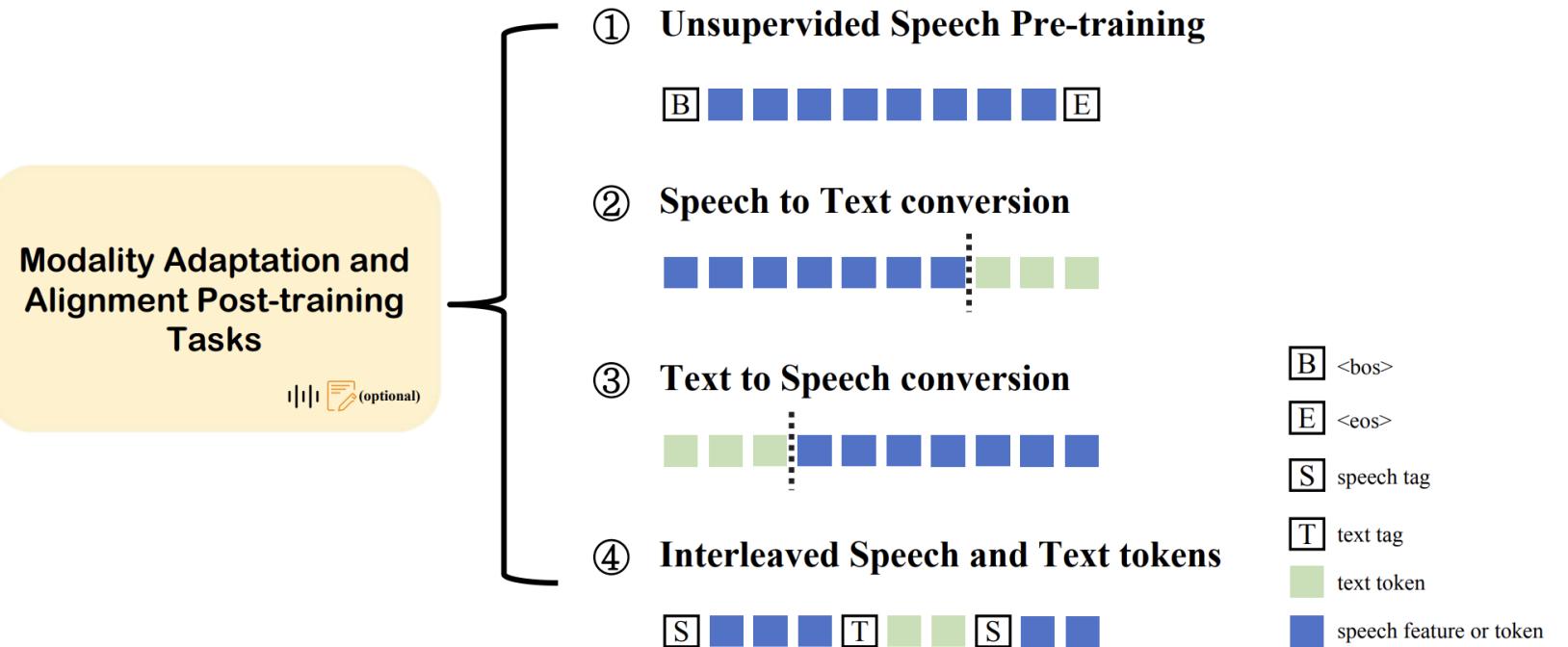


Этапы обучения



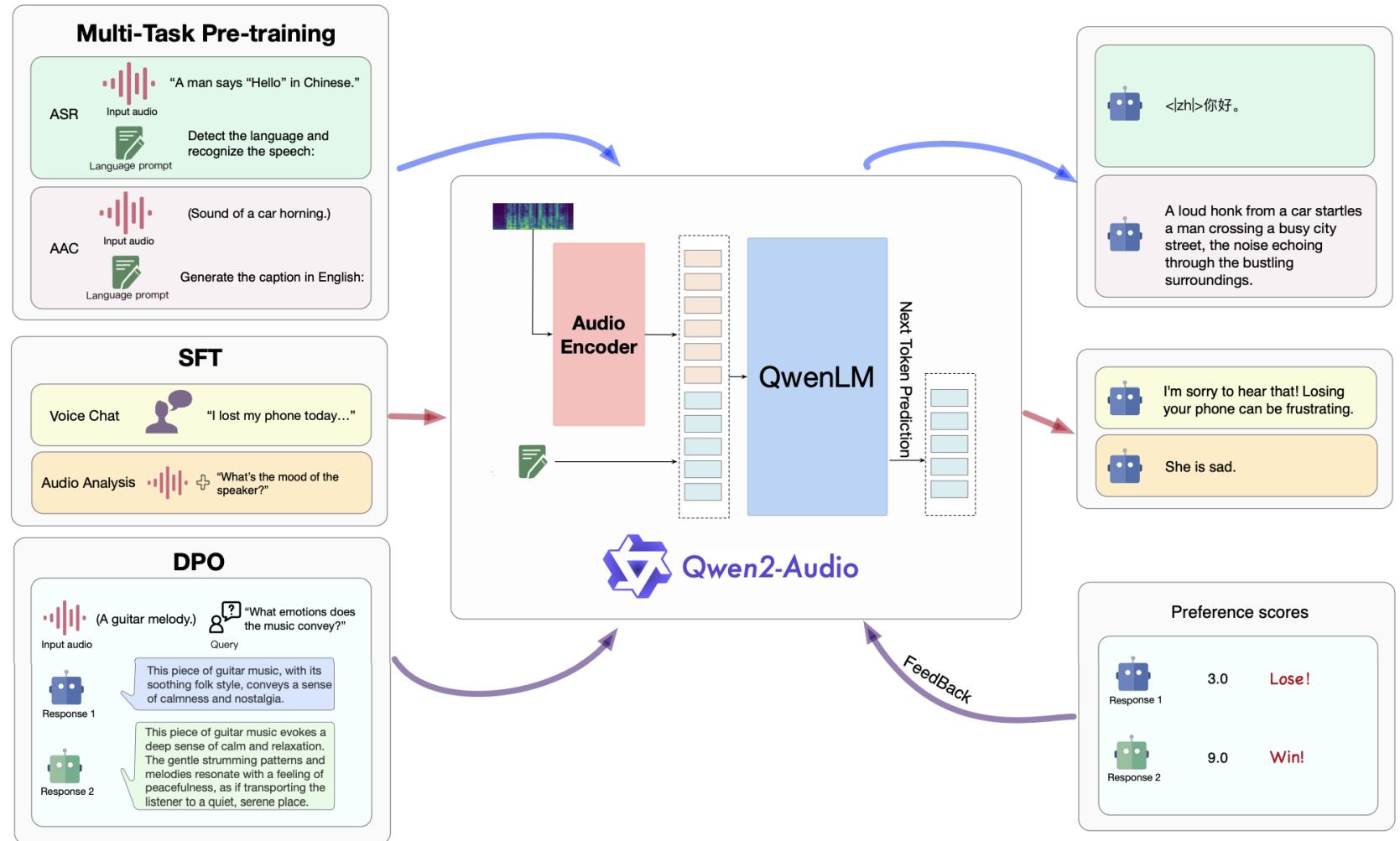
Этапы обучения

speech-text alignment



Этапы обучения

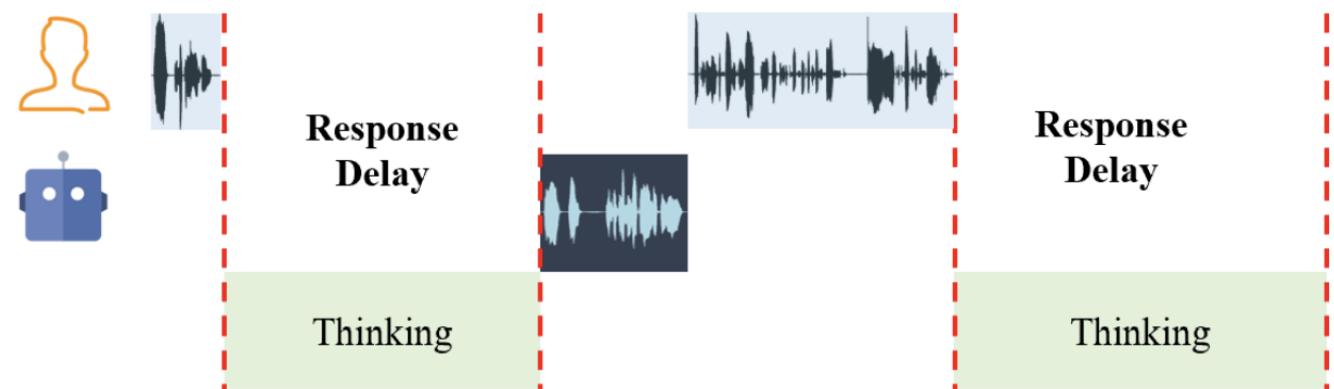
speech-text alignment
и DPO



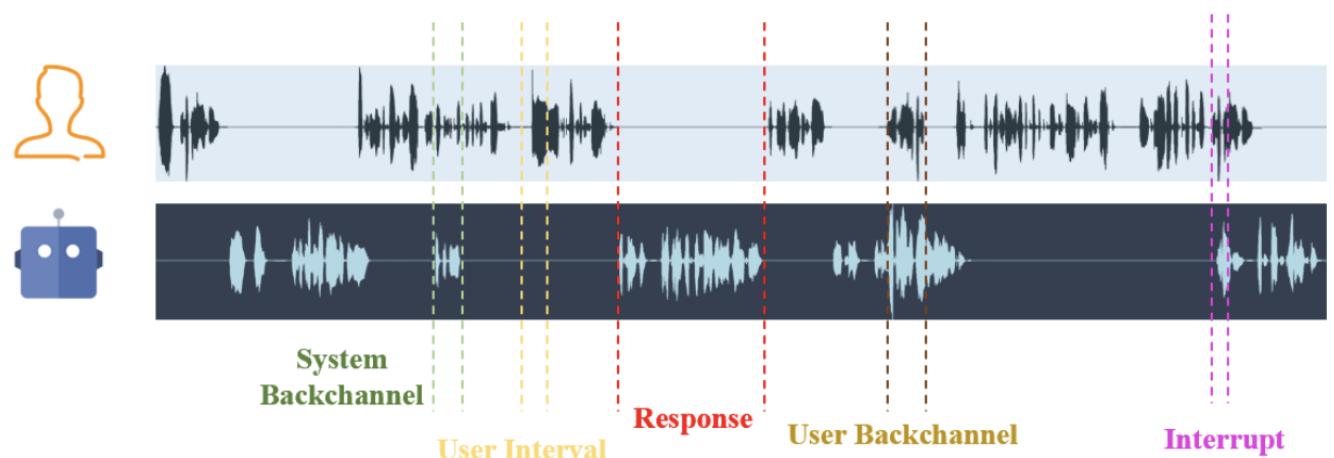
Дуплексные системы, интерактив- ность



(a) Simplex: One-way communication, and the direction is fixed.



(b) Half-duplex: Two-way communication, but not simultaneously.

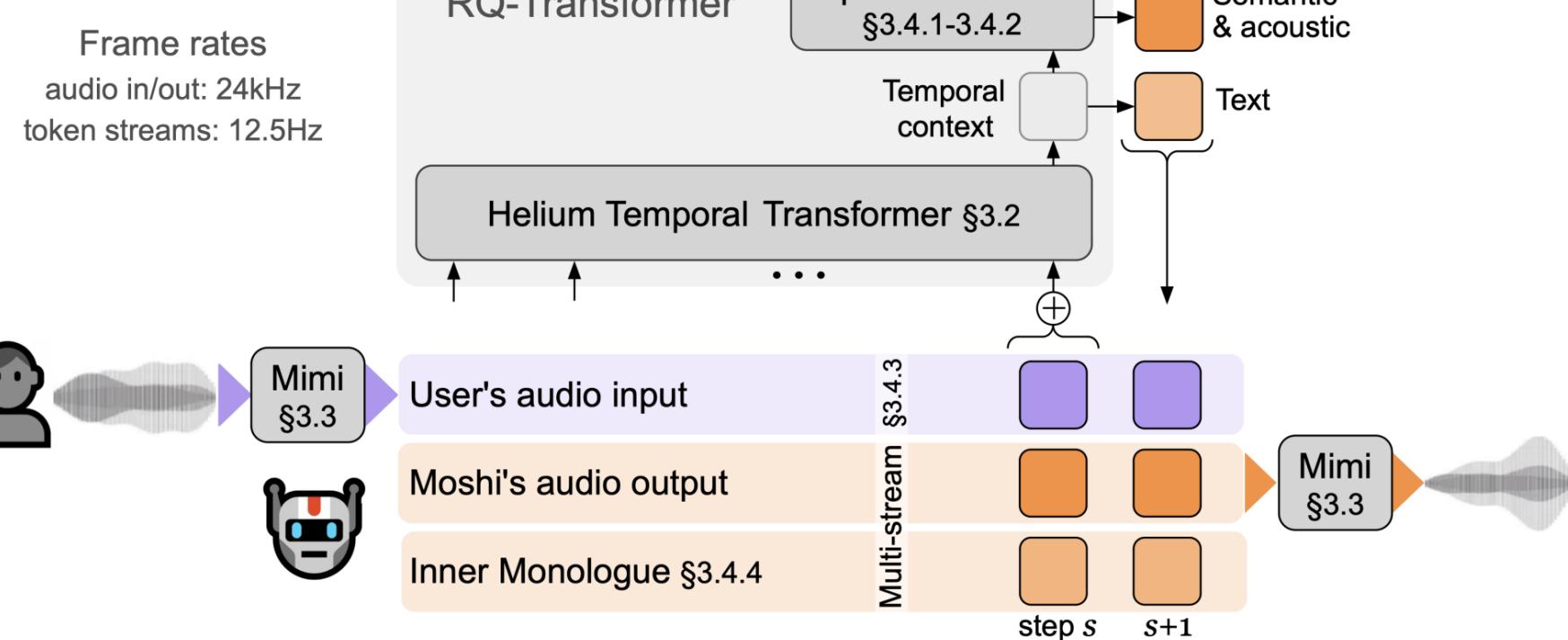


(c) Full-duplex: Two-way communication, simultaneously.

Дуплексные системы, интерактивность: Moshi

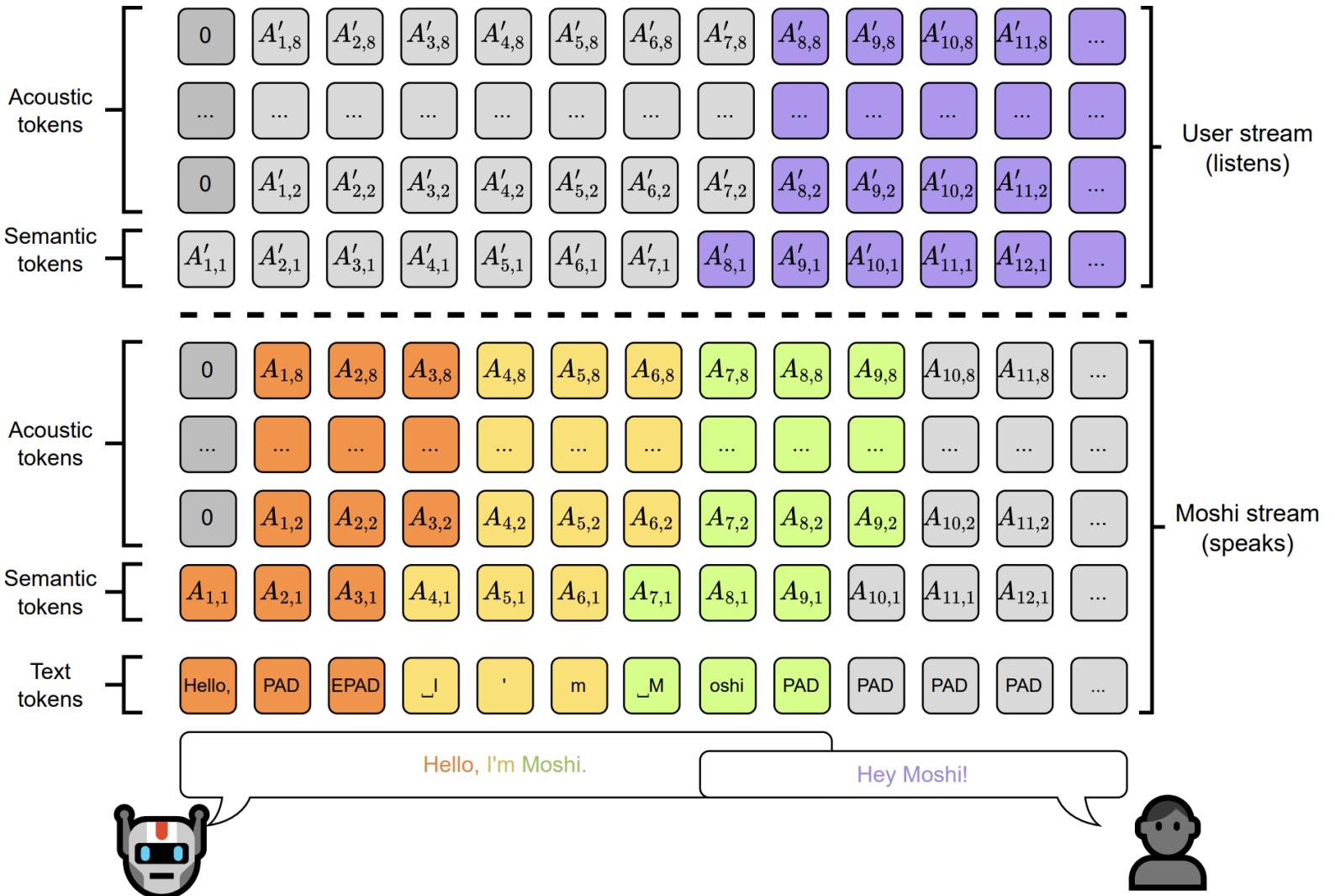
Défossez, Alexandre, et al. "Moshi: a speech-text foundation model for real-time dialogue." *arXiv preprint arXiv:2410.00037* (2024).

1. End-to-end дуплексная audio-llm
2. Разработали Mimi
3. Сделали иерархическую структуру предсказания аудио токенов, моделирующую речь обоих спикеров
4. Предложили идею "внутреннего диалога"
5. Обучают на текст -> raw audio -> аудио диалоги

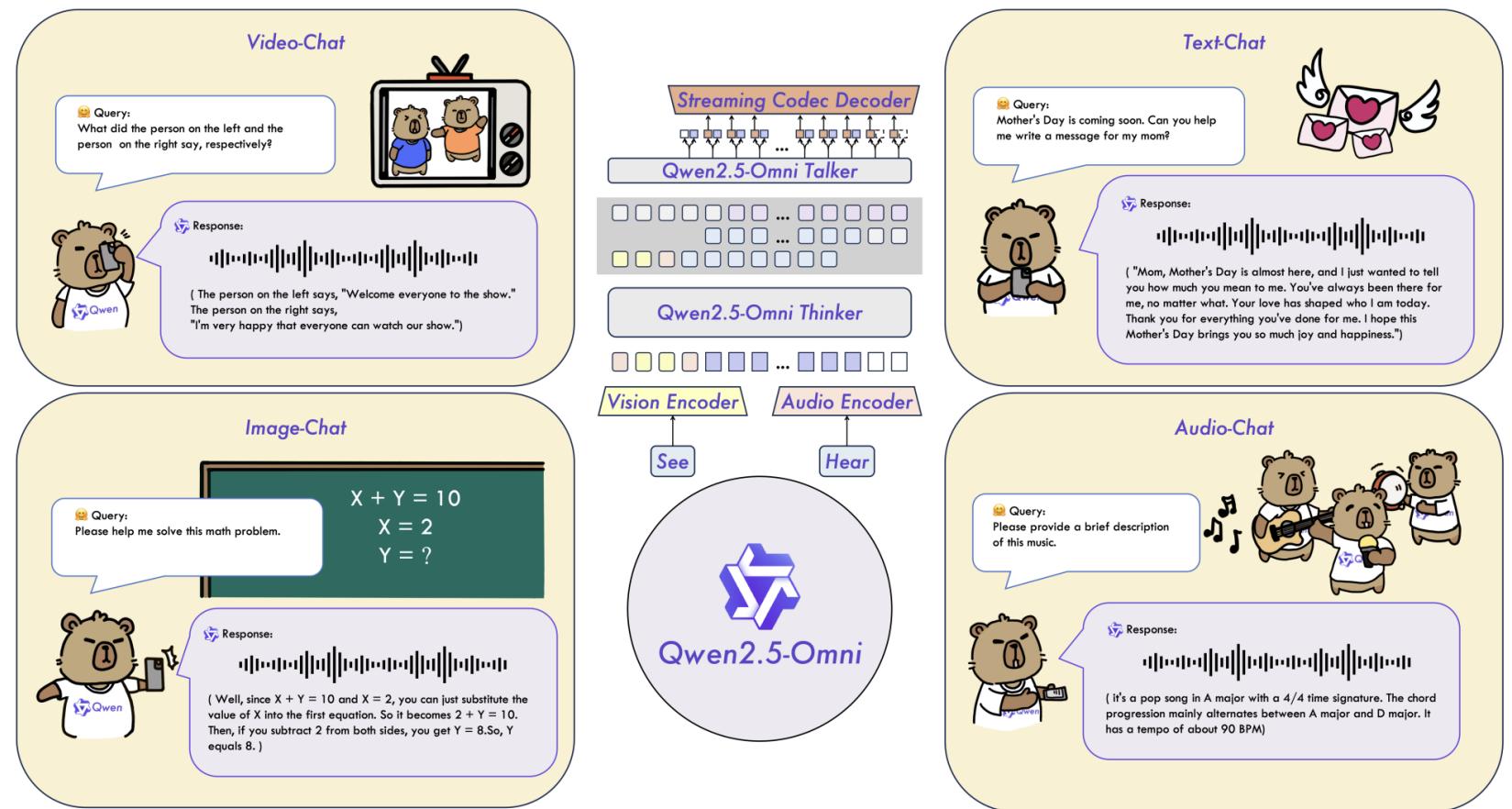


Дуплексные системы, интерактив- ность Moshi

Используют задержку акустических токенов относительно семантических



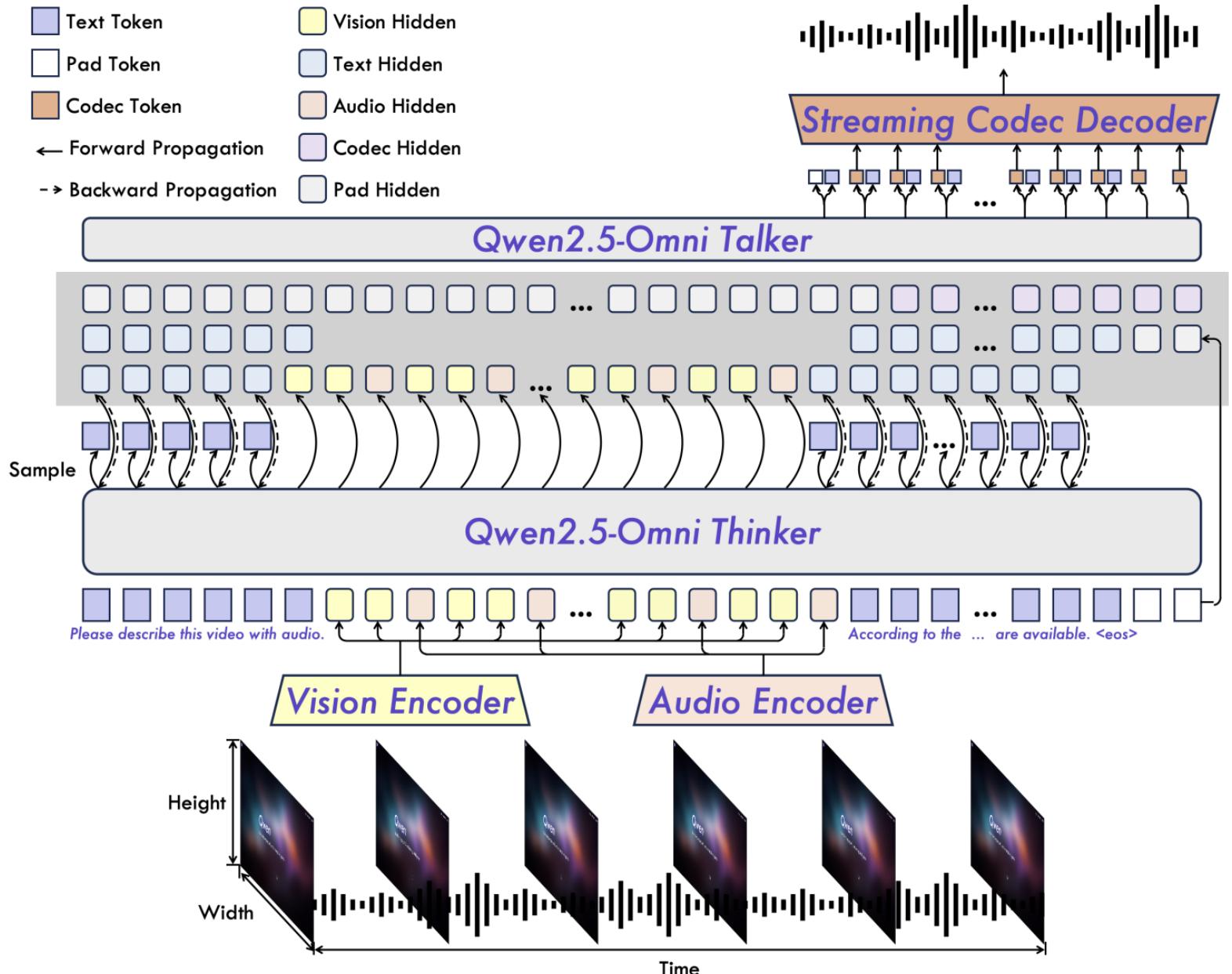
SoTA: Qwen2.5-omni



1. Текст, видео, аудио на вход; текст, аудио на выход
2. Thinker-Talker архитектура
3. Берут текстовую LLM, добавляют энкодер визуальной информации и аудио энкодер (whisper)
4. Сначала замораживают параметры LLM и энкодеров и тренируют адаптеры, потом энкодеры, и finally - всю модель
5. Pretrain на сырых мультимодальных данных, дальше instruction following и DPO

SoTA: Qwen2.5-omni

Похоже на RQ-transformer как у Moshi, но talker берет весь контекст



SoTA: Qwen2.5-omni

Table 1: Text → Text performance of 7B+ pure text models and Qwen2.5-Omni

Datasets	Gemma2-9B	Llama3.1-8B	Qwen2-7B	Qwen2.5-7B	Qwen2.5-Omni-7B
<i>General Tasks</i>					
MMLU-Pro	52.1	48.3	44.1	56.3	47.0
MMLU-redux	72.8	67.2	67.3	75.4	71.0
LiveBench ₀₈₃₁	30.6	26.7	29.2	35.9	29.6
<i>Mathematics & Science Tasks</i>					
GPQA	32.8	32.8	34.3	36.4	30.8
MATH	44.3	51.9	52.9	75.5	71.5
GSM8K	76.7	84.5	85.7	91.6	88.7
<i>Coding Tasks</i>					
HumanEval	68.9	72.6	79.9	84.8	78.7
MBPP	74.9	69.6	67.2	79.2	73.2
MultiPL-E	53.4	50.7	59.1	70.4	65.8
LiveCodeBench ₂₃₀₅₋₂₄₀₉	18.9	8.3	23.9	28.7	24.6

SoTA: Qwen2.5-omni

Table 3: **Audio → text performance of State-of-the-art and Qwen2.5-Omni**

Datasets	Model	Performance
SER		
Meld	WavLM-large (Chen et al., 2022)	0.542
	MiniCPM-o (Yao et al., 2024)	0.524
	Qwen-Audio (Chu et al., 2023a)	0.557
	Qwen2-Audio (Chu et al., 2024a)	0.553
	Qwen2.5-Omni-7B	0.570
VSC		
VocalSound	CLAP (Elizalde et al., 2022)	0.495
	Pengi (Deshmukh et al., 2023)	0.604
	Qwen-Audio (Chu et al., 2023a)	0.929
	Qwen2-Audio (Chu et al., 2024a)	0.939
	Qwen2.5-Omni-7B	0.939
<i>Music</i>		
GiantSteps <i>Tempo</i>	LLark-7B (Gardner et al., 2023)	0.86
	Qwen2.5-Omni-7B	0.88
MusicCaps	LP-MusicCaps (Doh et al., 2023)	0.291 0.149 0.089 0.061 0.129 0.130
	Qwen2.5-Omni-7B	0.328 0.162 0.090 0.055 0.127 0.225

— — — — —