Multimodal LLMs: From Speech LLMs to Multimodal Agents

Shujie Liu, Microsoft Research Asia

Text LLMs Set the Foundation for Multimodal Agents

Scaling

Capacity + knowledge

More parameters and data expand what the model can learn.

Alignment

Useful interaction

Instructions, interaction, and tool use turn capability into practical intelligence.

Next question

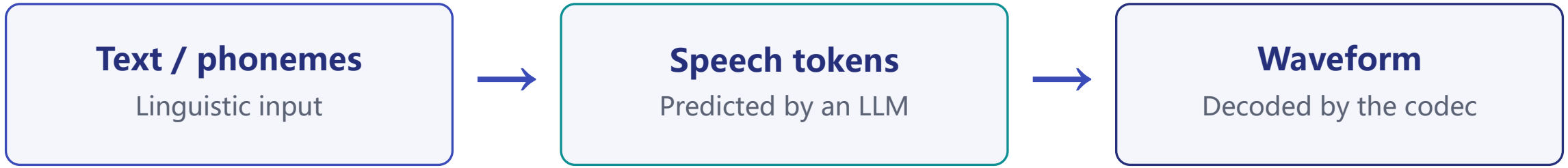
Beyond text

Represent and align speech, vision, and other continuous signals for direct reasoning.

Can we represent and align speech, vision, and other signals so an LLM can reason over them directly?

Give Speech a Language-Like Form

Core idea: Represent speech with discrete codec tokens, then model TTS as next-token prediction.



Discrete tokens bridge continuous speech and LLM-style training.

In-context capabilities

Voice prompting

A short prompt specifies speaker, style, and emotion.

Voice-preserving editing

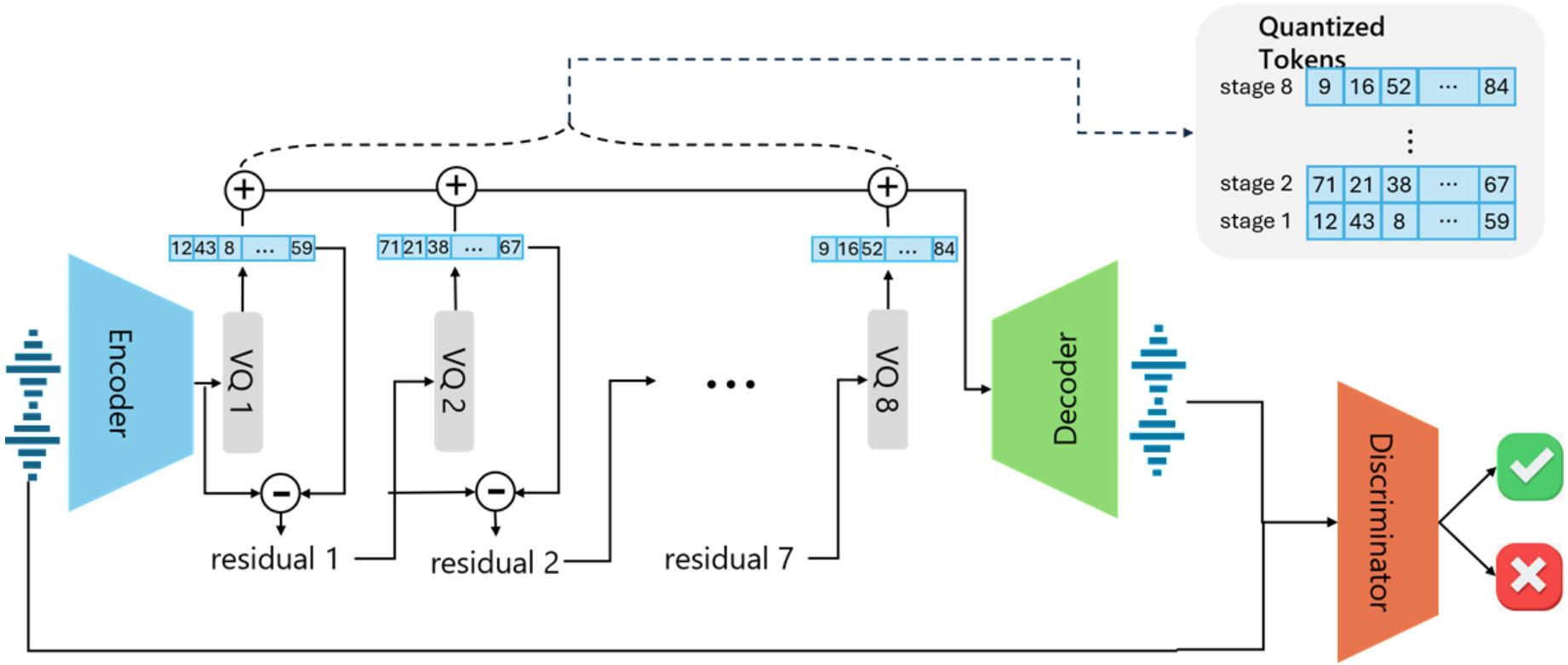
Change the content while retaining voice and style.

Expressive generation

Generate natural, varied, and expressive speech.

Tokenize Speech: SoundStream/Encodec

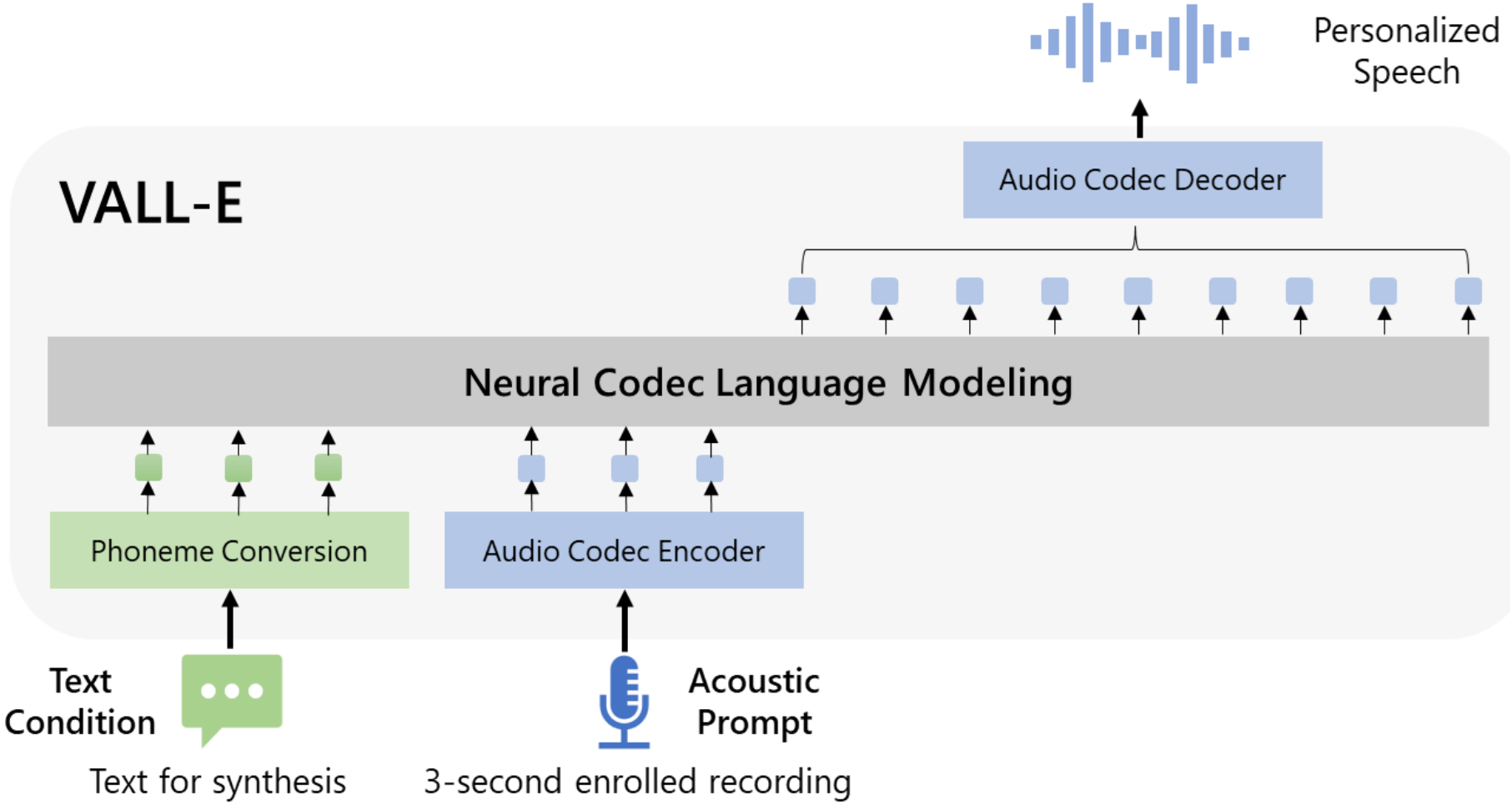
Core idea: Residual vector quantization converts speech into stacked discrete token streams.



Stacked codec tokens provide the discrete speech representation used by VALL-E.

Zeghidour et al., **SoundStream: An End-to-End Neural Audio Codec**. 2021
Defossez et al., **High Fidelity Neural Audio Compression**. 2022

Modeling Speech as Language: VALL-E



Wang et al., **Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers**. 2023

Extending VALL-E: VALL-E X and VALL-E 2

VALL-E

Neural codec language models

Discrete token conditional language model for zero-shot TTS.

VALL-E X

Cross-lingual

Cross-language zero-shot TTS and maintaining the speaker identity—speaking a foreign language with your own voice.

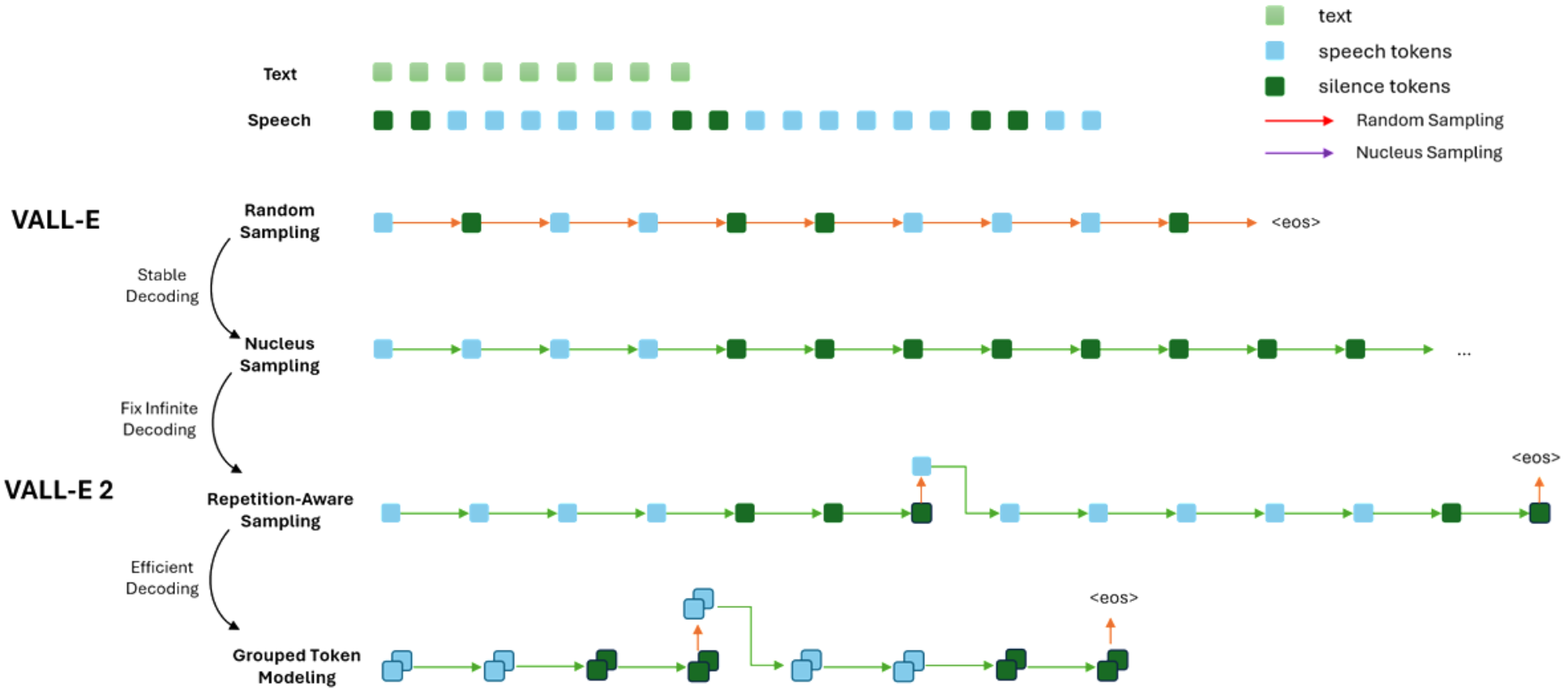
VALL-E 2

Reaching human levels

Pushing zero-shot TTS to human levels through repeated perceptual sampling (RAS) and grouped code modeling.

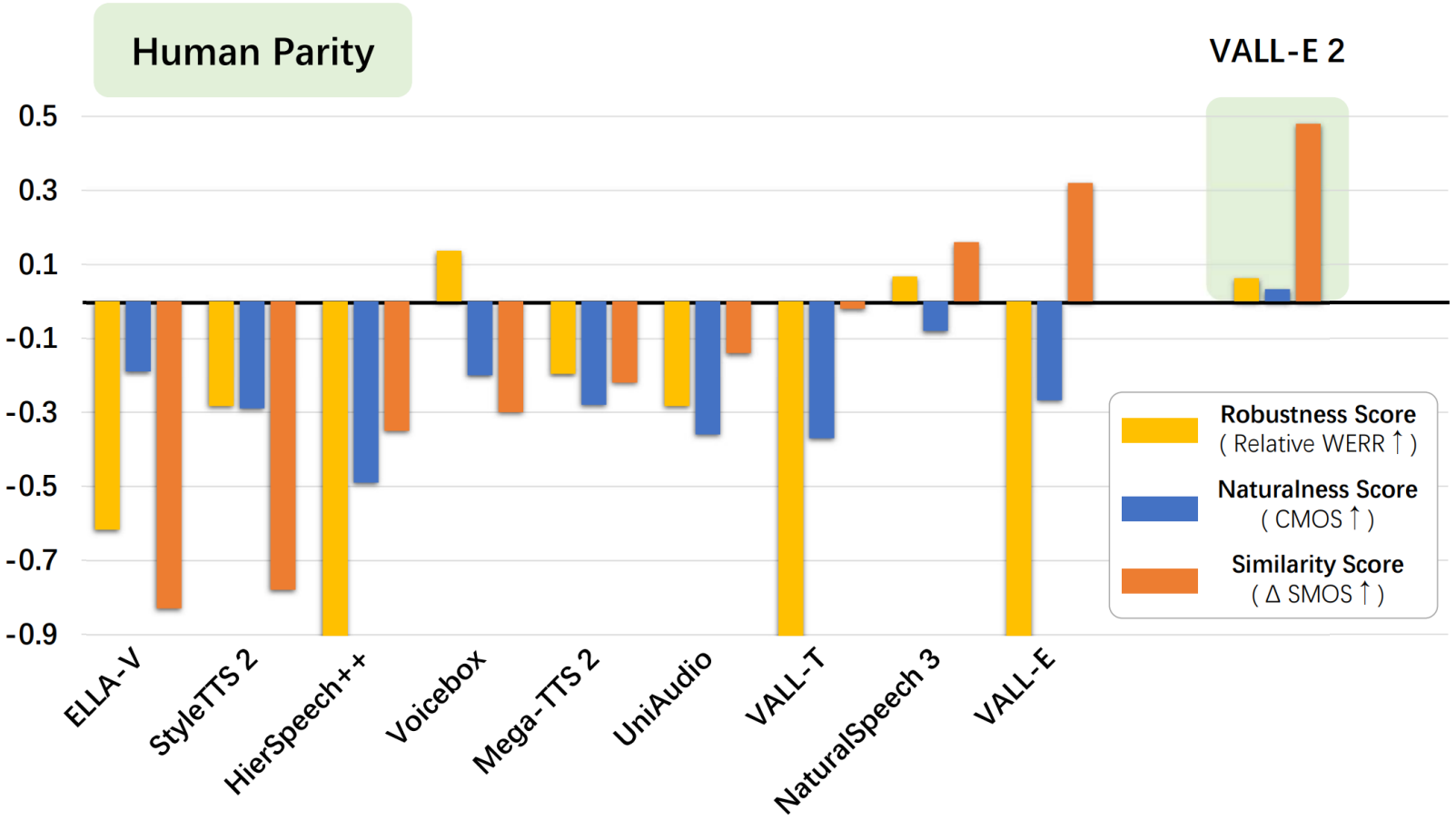
Once speech is discrete, it is possible to use text large language model technology to naturally extend neural coding language models to multilingual environments, balancing efficiency and robustness.

VALL-E 2: Robustness and Efficiency



Chen et al., **VALL-E 2: Neural Codec Language Models are Human Parity Zero-Shot Text to Speech Synthesizers**. 2024

Human Parity Zero-Shot TTS System



Numbers calculated based on the results reported in the original papers, irrespective of differences in model architecture and training data. This conclusion is drawn solely from experimental results on the LibriSpeech and VCTK datasets.

Chen et al., **VALL-E 2: Neural Codec Language Models are Human Parity Zero-Shot Text to Speech Synthesizers**. 2024

VALL-E vs Traditional TTS Models

	Traditional Models	LM-based Models
Representations	Mel Spectrogram	Discrete Tokens
Training Objective	Regression Loss	Cross Entropy Loss
Training Data	High Quality Data Collected in Studio	Diverse Data Collected from Web
Data Size	≤600h	>60Kh
In-Context Learning Capability	✗	✓
Diversity	Poor	Rich
Integrate with LLMs	Hard	Easy

Zero-Shot TTS Demos

Emotion Keeping Examples		
	Prompt	**VALL-E**
Anger		
Sleepy		

Hard Examples			
Text	**Speaker Prompt**	**VALL-E**	**VALL-E 2**
F one F two F four F **eight H** sixteen H thirty two H sixty four			
Curious koalas curiously climbed curious **curious** climbers			

Zero-Shot Cross-Lingual TTS Demos

Emotion	Source Language	Target Language	Original speech	VALL-E X
Anger	English	Chinese		
	Chinese	English		
Happy	English	Chinese		
	Chinese	English		

English Prompt	Chinese Speech with English ID	Chinese Speech with Chinese ID

Foreign Accent Problem

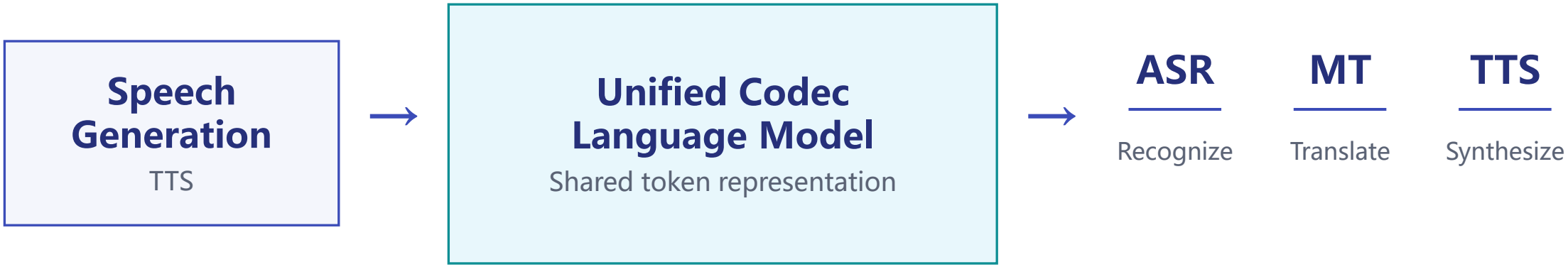
Naturalness of Zero-Shot TTS

Transcription	MELLE
它, 它喜欢在那个毛拖鞋上, 我也是, 我也是醉了, 而且不, 我们不知道怎么指导它到那里上厕所, 是很尴尬的事情, 养了半年多长得,长得长得贼肥, 然后, 后来实在养不了, 然后就别人也不想养, 我们就是送那个宿管大叔了, 不知道是吃了还是怎么样.。	
要说别人各种各种好,其实这种话呢,不传在自己的耳朵里的话, 就觉得无所谓,但是你真正有人给你讲了之后,你可能就心里还是多多少少不舒服的你知道吗?	

The text for the speech prompt is in blue, and the model generates the speech for the rest.

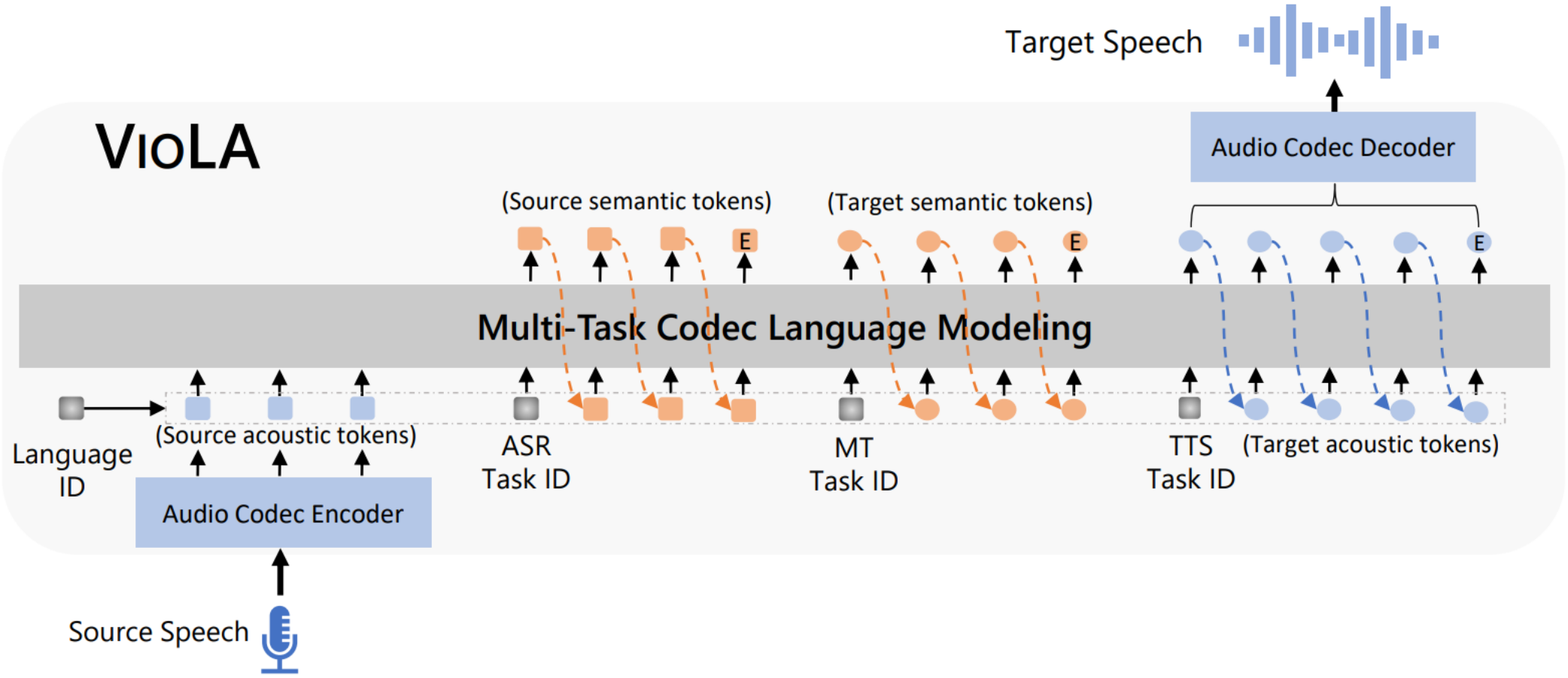
From TTS to Other Speech Processing Tasks

EXPANDING THE PARADIGM



One token-based framework expands from synthesis to recognition, translation and other speech processing tasks.

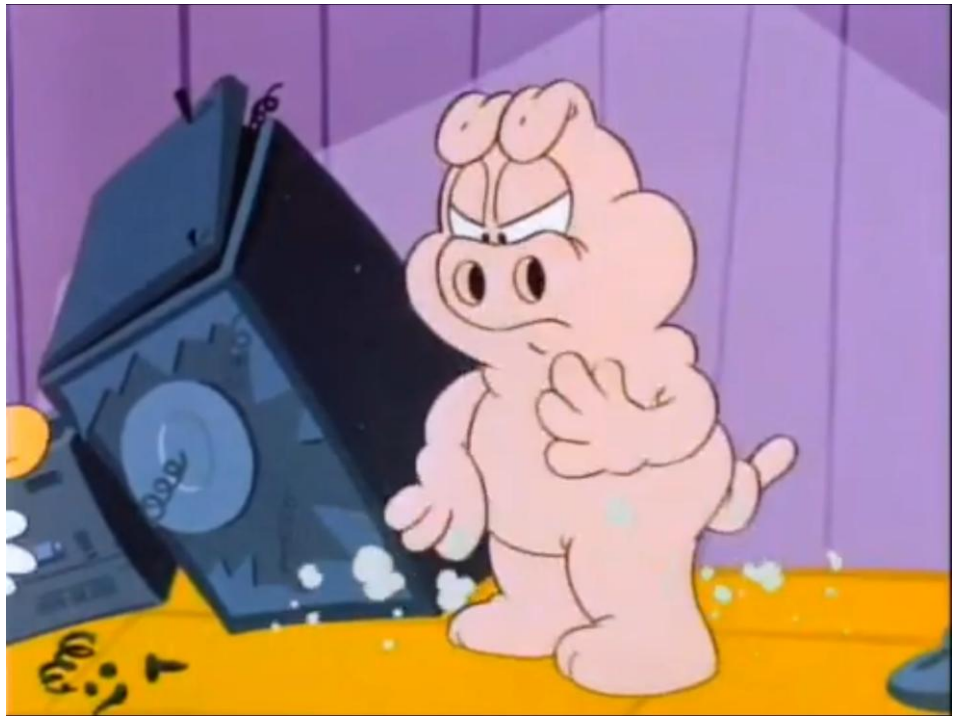
VIOLA: Decoder Only LLMs for ASR, MT and TTS



Wang et al., **VioLA: Unified Codec Language Models for Speech Recognition, Synthesis, and Translation**. 2023

VIOLA: Decoder Only LLMs for ASR, MT and TTS

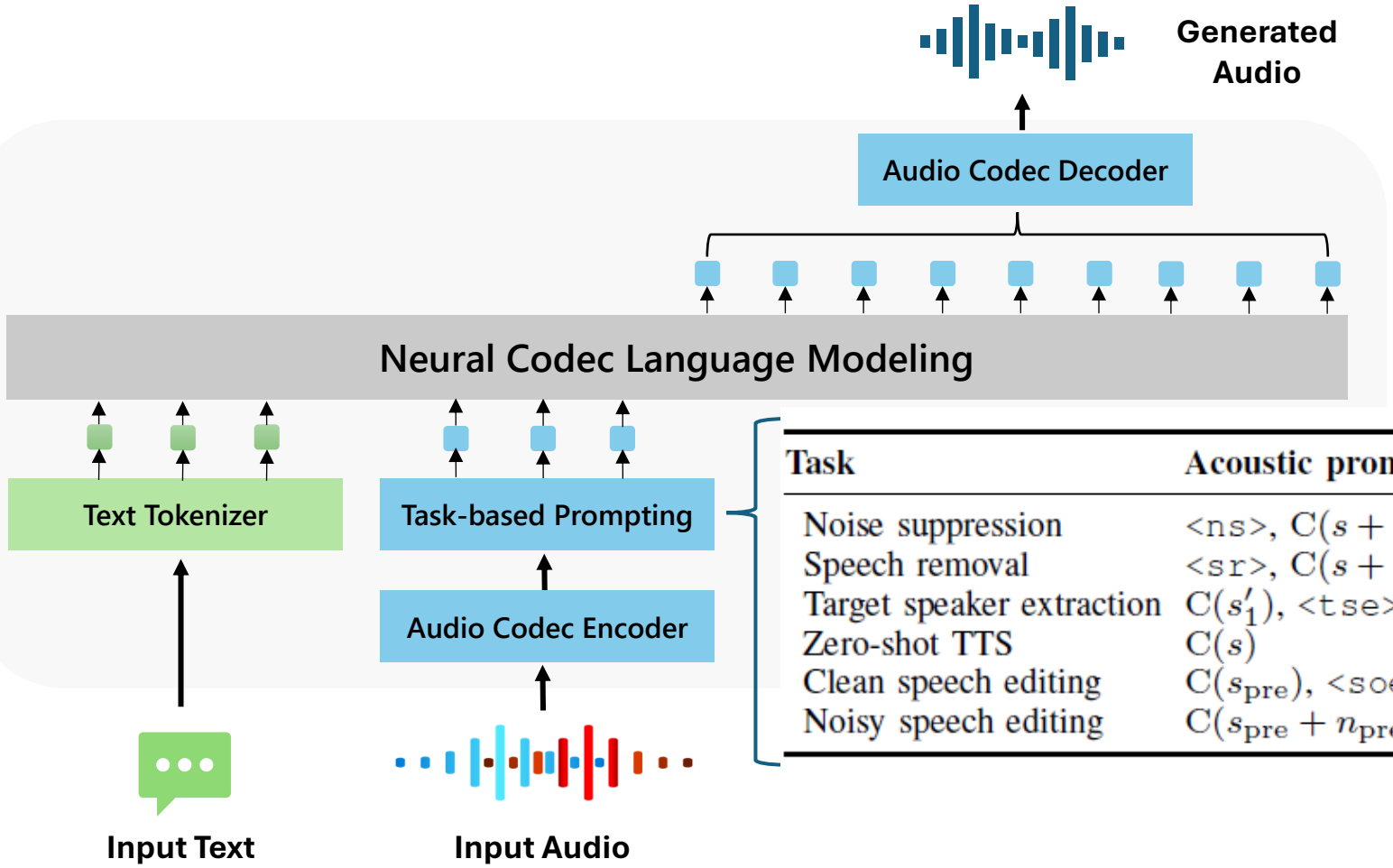
Source Video



VioLA

Wang et al., **VioLA: Unified Codec Language Models for Speech Recognition, Synthesis, and Translation**. 2023

SpeechX: Versatile Speech Transformer



Task	Acoustic prompt $\mathcal{A}$
Noise suppression	<ns>, $\mathrm{C}(s+n)$
Speech removal	<sr>, $\mathrm{C}(s+n)$
Target speaker extraction	$\mathrm{C}(s_1')$, <tse>, $\mathrm{C}(s_1+s_2)$
Zero-shot TTS	$\mathrm{C}(s)$
Clean speech editing	$\mathrm{C}(s_{\text{pre}})$, <soe>, <mask>, <eoe>, $\mathrm{C}(s_{\text{post}})$
Noisy speech editing	$\mathrm{C}(s_{\text{pre}}+n_{\text{pre}})$, <soe>, $\mathrm{C}(s_{\text{mid}}+n_{\text{mid}})$, <eoe>, C(

Wang et al., **SpeechX: Neural Codec Language Model as a Versatile Speech Transformer**. 2023

SpeechX: Versatile Speech Transformer

Background-preserving Spoken Content Editing			
Original Text	**Edited Text**	**Original speech**	**SpeechX**
we will go out together to the bower there is a **way** down to the **court from my window**	we will go out together to the bower there is a **pathway** down to the **courtyard**		

Noise Suppression			
Text	**Noisy Speech**	**SpeechX**	**Ground truth**
the salient features of this development of domestic service have already been indicated			

Target Speaker Extraction			
Text	**Speaker Prompt**	**Mixed Speech**	**SpeechX**
at that moment the gentleman entered bearing a huge object concealed by a piece of green felt			

From Speech Processing to Speech LLM

Speech is the natural interface for large models: listening, thinking, and speaking simultaneously.

Listen → **Thinking** → **Speaking**

WavLLM

Directly use voice as input

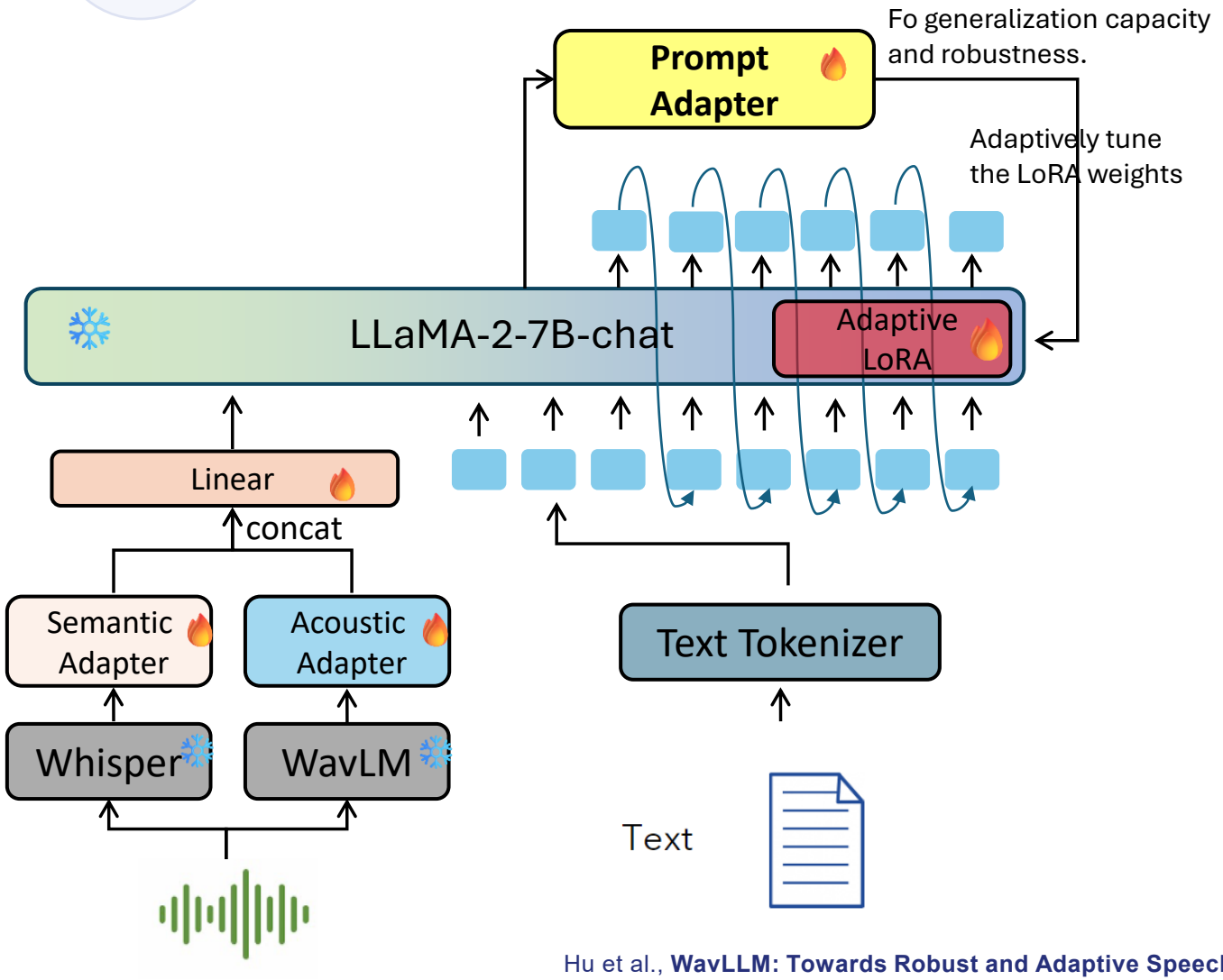
Treat speech as the model's native input, rather than a "front-end ASR + back-end LLM" pipeline.

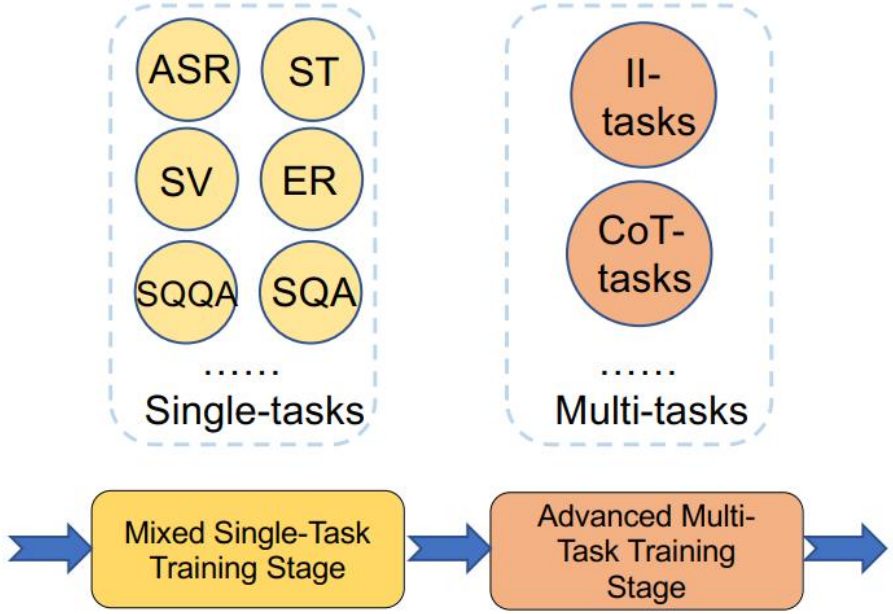
SLAM-Omni

End-to-end voice interaction

End-to-end, controllable voice interaction, speaking while reasoning.

WavLLM: Speech to Text LLM





II-tasks: Independent instruction combines multiple independent prompts in a single instruction.
CoT-tasks: Sequantically progressive instruction decomposes a complex task into multiple sub-tasks.

Hu et al., **WavLLM: Towards Robust and Adaptive Speech Large Language Model**. 2024

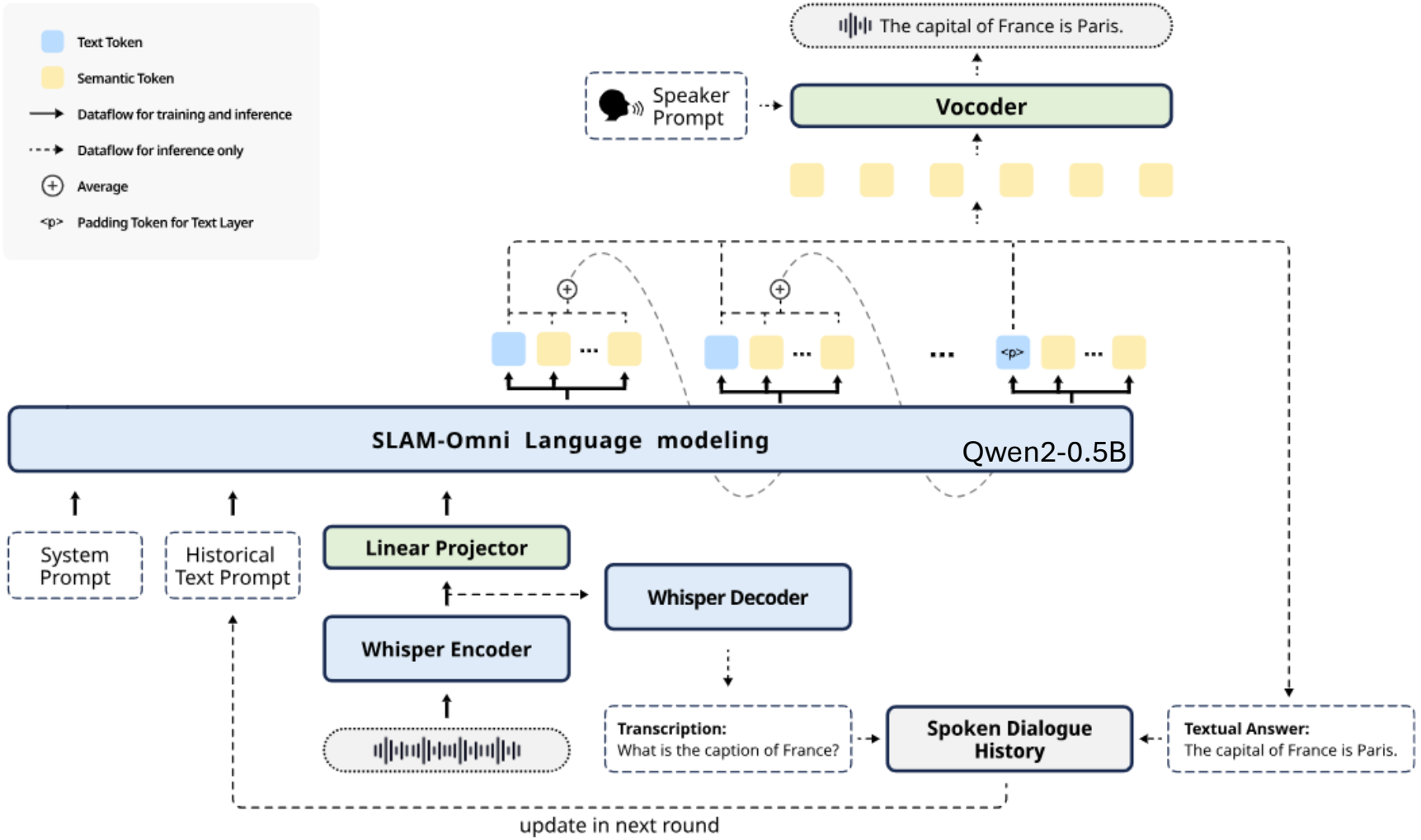
WavLLM: Speech to Text LLM

Models	ASR		ST (En2De)		SV	ER	SQQA	SQA
	test-clean	test-other	CoVoST2	MUSTC				
	WER↓		BLEU↑		Acc.↑	Acc.↑	Acc.↑	Acc.↑
Whisper + LLM	2.7*	5.2*	18.2	11.5	-	-	0.78	59.30% (63.50%)
SALMONN-7B	2.4	5.4	17.1	12.5	0.86	-	-	39.95% (40.00%)
SALMONN-13B	2.1*	4.9*	18.6*	19.5	0.94*	0.69*	0.41*	43.35% (43.35%)
Qwen-Audio-Chat 7B	2.2	5.1	23.2	18.4	0.50	-	0.38	25.50% (54.25%)
Our WavLLM 7B	**2.0**	**4.8**	**23.6**	**21.7**	**0.91**	**0.72**	0.57	**67.55% (67.55%)**

WavLLM achieves better performance on 7B baseline models, even better than SALMONN-13B.

Hu et al., **WavLLM: Towards Robust and Adaptive Speech Large Language Model**. 2024

SLAM-Omni: From TTS to Voice Interaction



Chen et al., **SLAM-Omni: Timbre-Controllable Voice Interaction System with Single-Stage Training**. 2025

SLAM-Omni: From TTS to Voice Interaction

Models	LLM Scale	Understanding		Reasoning			Oral Conversation			Overall
		Repeat	Summary	StoralEval	TruthfulEval	MLC	AlpacaEval	CommonEval	WildchatEval	
Qwen2-7B-instruct†	7B	96.87	97.45	82.35	67.89	73.26	95.91	85.93	92.72	86.55
Freeze-Omni	7B	70.89	78.87	57.74	46.95	42.56	52.23	48.70	55.80	56.72
LLaMA-Omni	8B	45.62	80.68	50.65	45.13	44.44	64.36	58.40	72.19	57.68
GLM-4-Voice	9B	90.95	91.07	73.80	59.28	57.82	80.77	63.07	78.76	74.44
Qwen2-0.5B-instruct†	0.5B	60.12	78.59	49.82	39.73	52.92	58.93	57.50	63.97	57.70
Mini-Omni	0.5B	5.07	32.20	23.25	25.06	2.82	30.99	29.80	31.42	22.58
Mini-Omni2	0.5B	8.10	40.06	28.49	26.92	6.97	34.81	30.70	36.43	26.56
SLAM-Omni (ours)	0.5B	**12.26**	**66.21**	**36.95**	**34.65**	**21.85**	**48.98**	**41.03**	**52.61**	**39.32**

Stronger than speech baselines

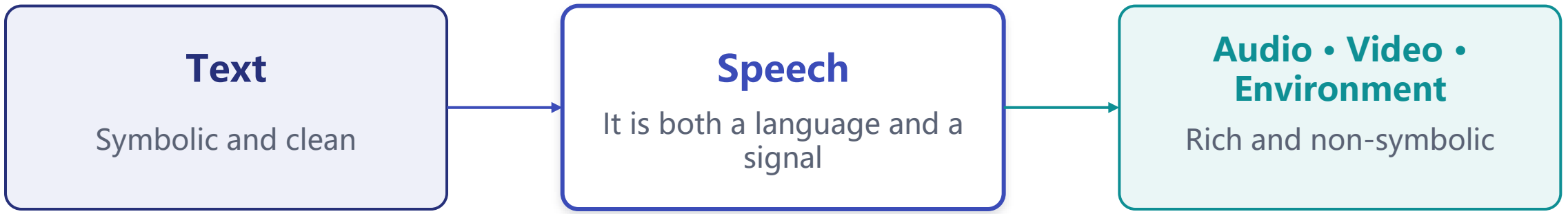
SLAM-Omni substantially outperforms Mini-Omni and Mini-Omni2.

Modality Gap

Performance remains well below the text-only Qwen2-0.5B-Instruct model.

Chen et al., **SLAM-Omni: Timbre-Controllable Voice Interaction System with Single-Stage Training**. 2025

Natural Interaction Requires Multimodality



Real interaction is inherently multimodal

Users can speak, display pictures and videos, and expect to understand their environment.

Modal bridges

Speech is both language and signal: close to text, yet similar to visual modality.

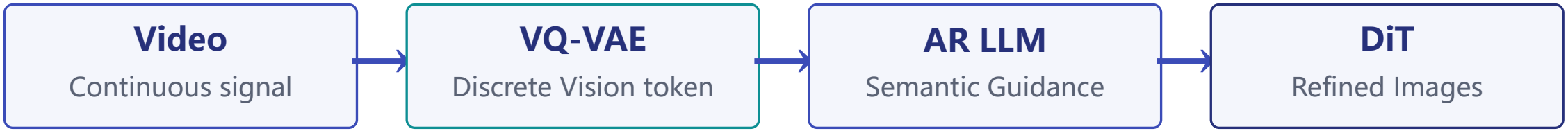
Stronger reasoning also brings new challenges

Multimodal large models have greater potential but introduce challenges of alignment and robustness.

If discrete tokens can unify speech, the same approach can also be extended to video.

From Speech Tokens to Visual Tokens

Compressed continuous signals into discrete tokens, then allowed to provide coarse-grained guidance from the autoregressive model.



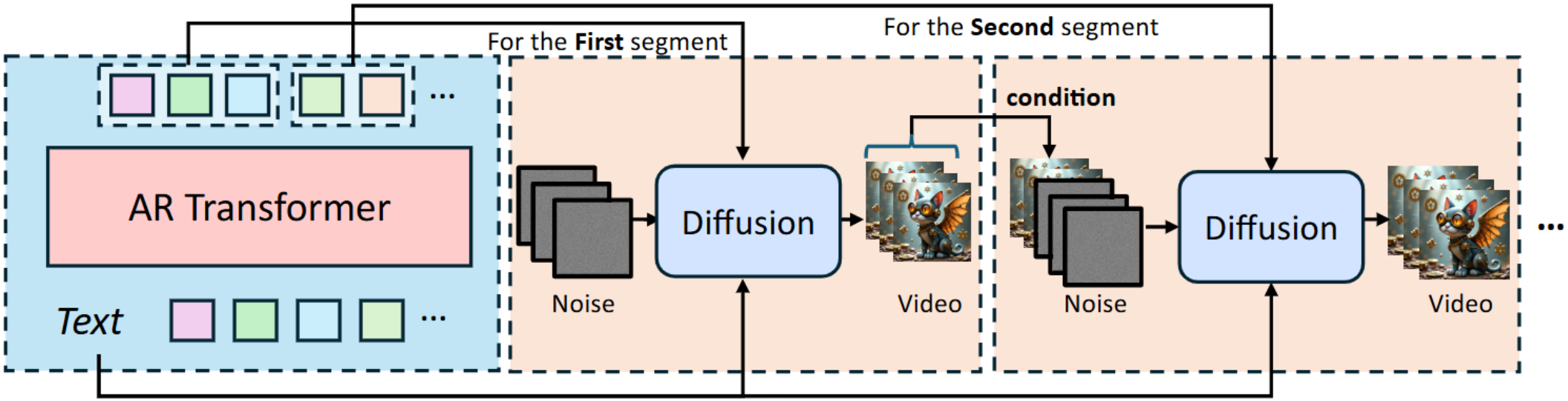
Connecting LLM and diffusion models

The token layer of VQ-VAE connects the autoregressive "planner" with the DiT "refiner."

Solving long-range consistency

LLM keeps long videos coherent in timing, and DiT noise reduction produces high-fidelity images.

ARLON: LLM-based Video Generation



ARLON first predicts long-range, coarse visual tokens (AR codes). These codes guide the DiT model to generate video one segment at a time. Each new segment uses the next AR codes and the final M seconds of the previous segment, helping the video remain consistent over time.

Li et al., **ARLON: Boosting Diffusion Transformers with Autoregressive Models for Long Video Generation**. 2024

ARLON: LLM-based Video Generation

ONE-MINUTE VIDEO CONSISTENCY • COMPARISON WITH LONG-VIDEO BASELINES

Models	Subject Consist	Background Consist	Motion Smooth	Dynamic Degree	Aesthetic Quality	Imaging Quality	Overall Consist
FreeNoise	96.59	97.48	98.36	17.44	47.39	**63.88**	25.78
StreamingT2V	87.31	94.64	93.83	**85.64**	44.57	53.64	23.65
OpenSora-V1.2	96.30	97.39	**98.94**	44.79	56.68	51.64	26.36
ARLON (Ours)	**97.11**	**97.56**	98.50	50.42	**56.85**	53.85	**26.55**

Stronger continuity

ARLON preserves subject and scene continuity better than long-video baselines.

VBENCH PERFORMANCE • COMPARISON WITH TEXT-TO-VIDEO MODELS

Models	Dynamic Degree	Aesthetic Quality	Imaging Quality	Subject Consist	Background Consist	Motion Smooth	Temporal Flicker	Temporal Style	Overall Consist	Scene	Object
Kling	46.9	61.2	65.6	**98.3**	97.6	99.4	99.3	24.2	26.4	50.9	87.2
Gen-2	18.9	**67.0**	**67.4**	97.6	97.6	**99.6**	99.6	24.1	26.2	48.9	90.9
Pika-V1.0	47.5	62.0	61.9	96.9	97.4	99.5	**99.7**	24.2	25.9	49.8	88.7
VideoCrafter-2.0	42.5	63.1	67.2	96.8	**98.2**	97.7	98.4	25.8	**28.2**	**55.3**	92.6
LaVie	49.7	54.9	61.9	91.4	97.5	96.4	98.3	**25.9**	26.4	52.7	91.8
LaVie-Interpolation	46.1	54.0	59.8	92.0	97.3	97.8	98.8	26.0	26.4	52.6	90.7
Show-1	44.4	57.4	58.7	95.5	98.0	98.2	99.1	25.2	27.5	47.0	**93.1**
CogVideo	42.4	38.2	41.0	92.2	96.2	96.5	97.6	7.8	7.7	28.2	73.4
OpenSoraPlan-V1.1	47.7	56.9	62.3	95.7	96.7	98.3	99.0	23.9	26.5	27.1	76.3
OpenSora-V1.2	47.2	56.2	60.9	94.5	97.9	98.2	99.5	24.6	27.1	42.5	83.4
ARLON (Ours)	**52.8**$^{\uparrow 5.6}$	61.0$^{\uparrow 4.8}$	61.0$^{\uparrow 0.1}$	93.4$^{\downarrow 1.1}$	97.1$^{\downarrow 0.8}$	98.9$^{\uparrow 0.7}$	99.4$^{\downarrow 0.1}$	25.3$^{\uparrow 0.7}$	27.3$^{\uparrow 0.2}$	54.4$^{\uparrow 11.9}$	89.8$^{\uparrow 6.4}$

Competitive Quality

Performance remains strong across visual-quality and semantic measures versus text-to-video systems.

Li et al., **ARLON: Boosting Diffusion Transformers with Autoregressive Models for Long Video Generation**. 2024

Text LLMs vs. Multimodal LLMs

Pure text large models

- Input is clean and symbolic
- Strong reasoning skills
- Stable and reliable

Multimodal large models

- Signal with noise, and redundancy
- Long-range timing structure
- Reasoning and robustness are relatively poor

Root Cause

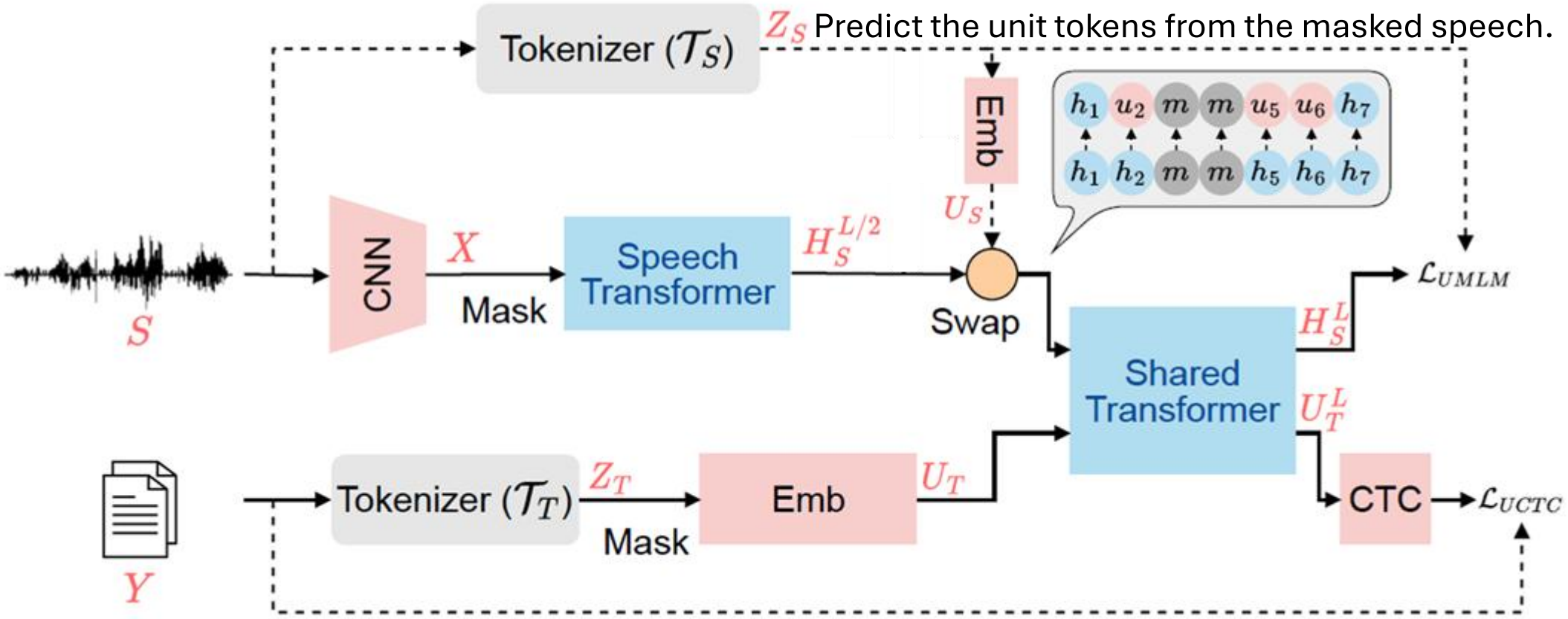
Modals are only added without true alignment: representing mismatches and noise signals forces the model to focus its modeling power on processing noise rather than inference. The real issue isn't *modal combination*, but modal **alignment**.

Modality Alignment : SpeechLM, TARS, C-Gate

Making non-text modalities compatible with large models: semantic alignment, structure alignment, and interpretability.

Stated	Sequence and Semantics	Cross-modal interface
SpeechLM	**TARS**	**C-Gate**
Unified statement	**Sequence alignment**	**Gating and routing**
Place voice tokens and text tokens in the same semantic space, supporting text-speech hybrid modeling and cross-modal migration.	Enable the model to understand the timing structure of multimodal sequences, rather than simply stitching them together.	A universal modal interface that represents information from other modalities in text modal form.

SpeechLM: Alignment for Speech and Text Input



Reconstruct the whole text sequences from the masked unit sequences.

Zhang et al., **SpeechLM: Enhanced Speech Pre-Training with Unpaired Textual Data**. 2022.

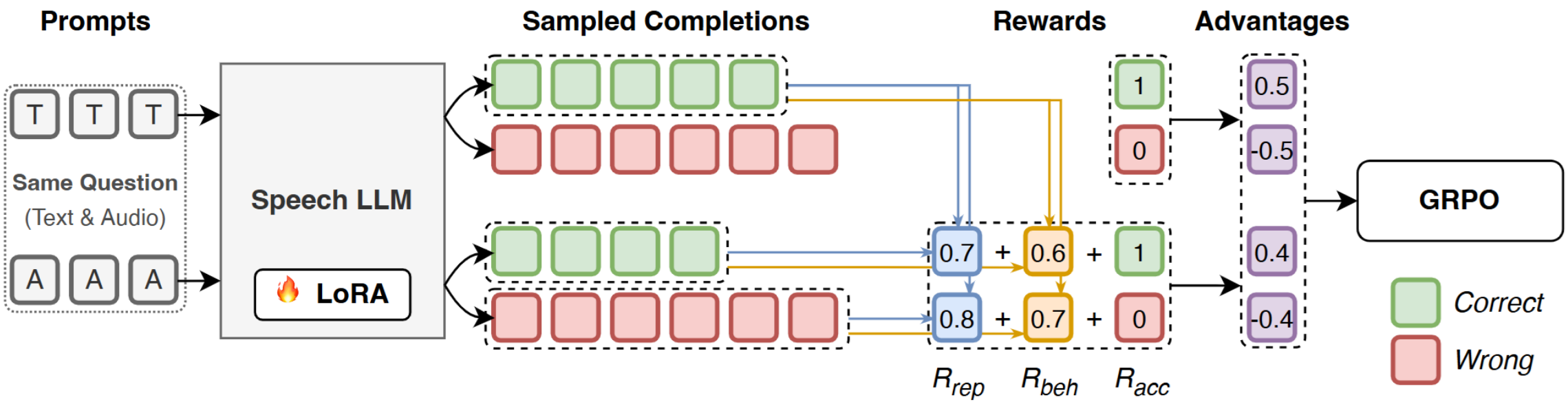
SpeechLM: Alignment for Speech and Text Input

Method	#Params	Corpus	Speaker			Content					Semantics				ParaL
			SID	ASV	SD	PR	ASR	OOD-ASR	KS	QbE	ST	IC	SF		ER
			Acc ↑	EER ↓	DER ↓	PER ↓	WER ↓	WER ↓	Acc ↑	MTWV ↑	BLEU ↑	Acc ↑	F1 ↑	CER ↓	Acc ↑
FBANK	0	-	8.5E-4	9.56	10.05	82.01	23.18	63.58	8.63	0.0058	2.32	9.10	69.64	52.94	35.39
PASE+ (Ravanelli et al., 2020)	7.83M	LS 50 hr	37.99	11.61	8.68	58.87	25.11	61.56	82.54	0.0072	3.16	29.82	62.14	60.17	57.86
APC (Chung et al., 2019)	4.11M	LS 360 hr	60.42	8.56	10.53	41.98	21.28	63.12	91.01	0.0310	5.95	74.69	70.46	50.89	59.33
VQ-APC (Chung et al., 2020)	4.63M	LS 360 hr	60.15	8.72	10.45	41.08	21.20	63.56	91.11	0.0251	4.23	74.48	68.53	52.91	59.66
NPC (Liu et al., 2020a)	19.38M	LS 360 hr	55.92	9.40	9.34	43.81	20.20	61.66	88.96	0.0246	4.32	69.44	72.79	48.44	59.08
Mockingjay (Liu et al., 2020c)	85.12M	LS 360 hr	32.29	11.66	10.54	70.19	22.82	65.27	83.67	6.6E-04	4.45	34.33	61.59	58.89	50.28
TERA (Liu et al., 2020b)	21.33M	LS 960 hr	57.57	15.89	9.96	49.17	18.17	58.49	89.48	0.0013	5.66	58.42	67.50	54.17	56.27
DeCoAR 2.0 (Ling and Liu, 2020)	89.84M	LS 960 hr	74.42	7.16	6.59	14.93	13.02	53.62	94.48	0.0406	9.94	90.80	83.28	34.73	62.47
modified CPC (Rivière et al., 2020)	1.84M	LL 60k hr	39.63	12.86	10.38	42.54	20.18	62.54	91.88	0.0326	4.82	64.09	71.19	49.91	60.96
wav2vec (Schneider et al., 2019)	32.54M	LS 960 hr	56.56	7.99	9.9	31.58	15.86	55.86	95.59	0.0485	6.61	84.92	76.37	43.71	59.79
vq-wav2vec (Baevski et al., 2020a)	34.15M	LS 960 hr	38.80	10.38	9.93	33.48	17.71	60.66	93.38	0.0410	5.66	85.68	77.68	41.54	58.24
Wav2vec 2.0 Base (Baevski et al., 2020b)	95.04M	LS 960 hr	75.18	6.02	6.08	5.74	6.43	46.95	96.23	0.0233	14.81	92.35	88.30	24.77	63.43
HuBERT Base (Hsu et al., 2021)	94.68M	LS 960 hr	81.42	5.11	5.88	5.41	6.42	46.69	96.30	0.0736	15.53	93.34	88.53	25.20	64.92
WavLM Base (Chen et al., 2022a)	94.70M	LS 960 hr	84.51	4.69	4.55	4.84	6.21	42.81	96.79	0.0870	20.74	93.63	89.38	22.86	65.94
SpeechLM-H Base	94.70M	LS 960 hr	76.90	5.79	6.10	3.70	**4.85**	45.82	95.91	0.0485	**21.90**	93.52	88.80	24.20	63.77
SpeechLM-P Base	94.70M	LS 960 hr	75.24	5.97	7.34	**3.10**	4.98	49.04	94.09	0.0410	19.20	97.68	87.67	25.90	61.84

Performance improved on **content-related and semantic-related** tasks (36% and 20% relative WER reductions on PR and ASR tasks), with performance degradation for the **speaker and paralinguistics-related tasks.**

Zhang et al., **SpeechLM: Enhanced Speech Pre-Training with Unpaired Textual Data**. 2022.

TARS: Closing the Reasoning Gap for Speech LLMs



$$R_{\text{base}} = R_{\text{acc}} + \lambda R_{\text{fmt}},$$

$$R_{\text{rep}} = \frac{1}{L}\sum_{l=1}^{L} \text{CosSim}(\bar{\mathbf{h}}^{(l)}_{\text{speech}}, \bar{\mathbf{h}}^{(l)}_{\text{text}})$$

Hidden states mean pooling

$$R_{\text{beh}} = \text{CosSim}(\mathcal{E}(y_{\text{speech}}), \mathcal{E}(y^{*}_{\text{text}}))$$

External embedding model

Wang et al., **Closing the Modality Reasoning Gap for Speech Large Language Models**. 2026.

TARS: Closing the Reasoning Gap for Speech LLMs

Model	Backbone	MMSU		OBQA		Average		MRR (%)
		A	T	A	T	A	T	
Proprietary & Cascaded Systems								
GPT-4o-mini-Audio	GPT-4o-mini	72.90	81.23	84.84	90.11	78.87	85.67	92.06
ASR† + Llama3.1-8B	Llama3.1-8B	58.78	65.65	72.53	80.88	65.66	73.27	89.61
ASR† + Qwen2.5-7B	Qwen2.5-7B	67.1	71.65	84.0	83.74	75.55	77.70	97.23
ASR† + Phi-4-7B	Phi-4-7B	69.00	74.92	77.80	83.96	73.40	79.44	92.40
Existing Baselines								
DeSTA2.5-Audio	Llama3.1-8B*	60.87	65.65	74.06	80.88	67.47	73.27	92.08
SALAD-7B	Qwen2.5-7B	57.5	71.6	75.1	90.1	66.30	80.85	85.33
MiniCPM-o 2.6	Qwen2.5-7B	54.78	59.42	78.02	82.86	66.40	71.14	85.46
Knowledge Distillation	Qwen2.5-7B	63.09	69.15	82.64	84.62	72.87	76.89	93.78
AlignChat	Qwen2.5-7B*	69.65	71.65	85.49	83.74	77.57	77.70	99.83
Base & Aligned Models								
Qwen2.5-Omni	Qwen2.5-7B	61.51	67.94	81.09	84.40	71.30	76.17	91.76
Phi-4-MM	Phi-4-7B	54.81	72.15	71.65	84.62	63.23	78.39	79.59
TARS (Qwen2.5-Omni)	Qwen2.5-7B	67.96	68.54	85.71	88.57	76.84	78.56	98.89
TARS (Phi-4-MM)	Phi-4-7B	**70.14**	**75.76**	**89.45**	**91.87**	**79.80**	**83.82**	**100.45**

1 Modality gap
Base models consistently perform worse on speech than on text.

2 Alignment helps
Prior methods narrow the gap, but most still remain below 100% MRR.

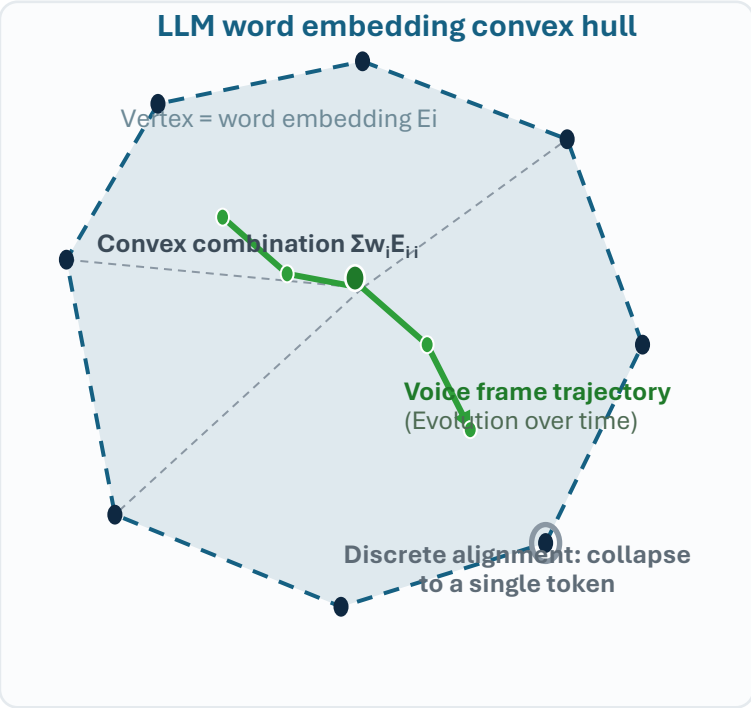
3 Best among 7B models
TARS + Qwen2.5-Omni reaches 76.84% audio accuracy and 98.89% MRR—above SALAD (66.30%), MiniCPM-o (66.40%), and KD (72.87%).

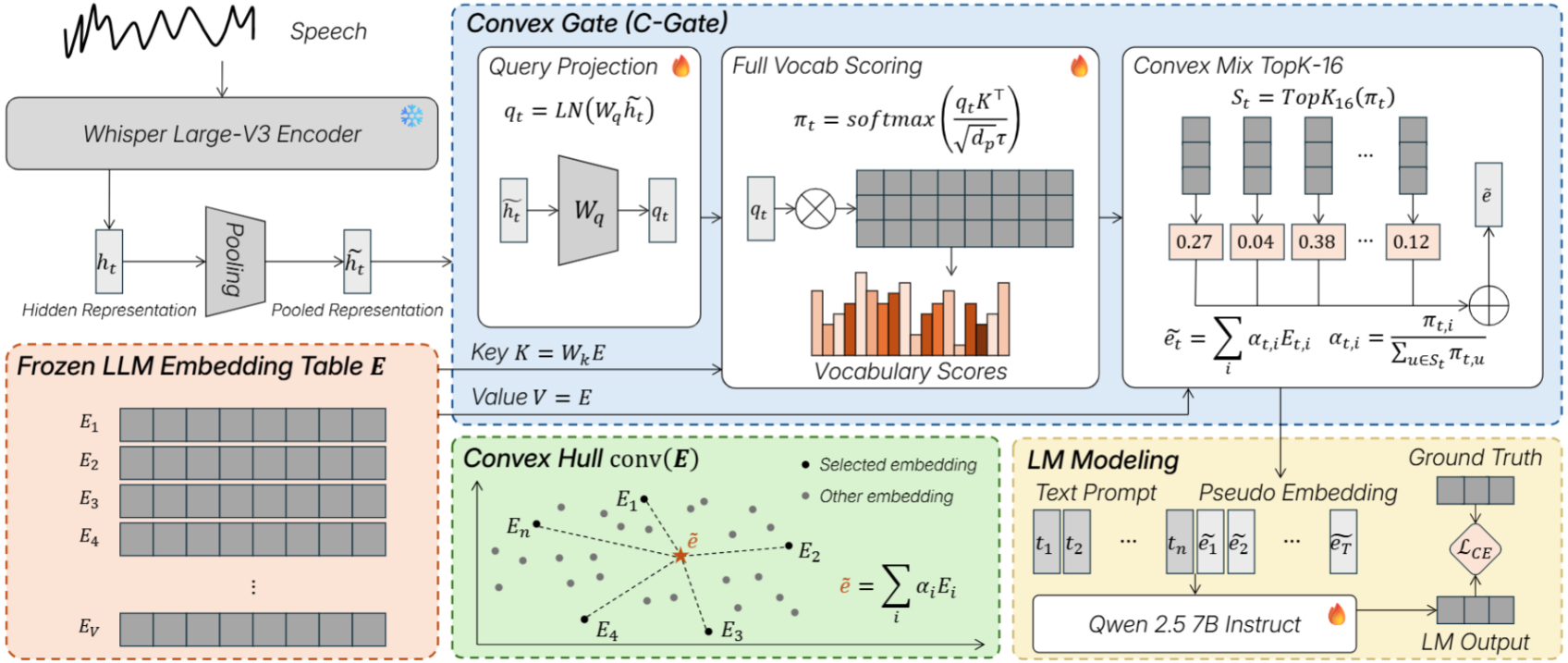
4 Full recovery on Phi-4-MM
TARS reaches 79.80% audio accuracy, above its 78.39% text result, with 100.45% MRR.

Wang et al., **Closing the Modality Reasoning Gap for Speech Large Language Models**. 2026.

C-Gate: Text as Unified Modal Interface

The limits of my language mean the limits of my world. ----Ludwig Wittgenstein





Hsu et al., **Is Text All You Need? Text as a Universal Information Bottleneck for Speech LLMs**. 2026

C-Gate: Text as Unified Modal Interface

System	LS WER ↓	MMSU ↑	MMAU ↑
Large-scale open-weight or foundation-model calibration			
Qwen-Audio-Chat [Chu et al., 2023]	2.0	46.9	41.9
Qwen2-Audio-Instruct [Chu et al., 2024]	1.6	53.3	52.5
Qwen2.5-Omni-7B [Xu et al., 2025]	1.8	61.3	65.6
Kimi-Audio-7B-Instruct [KimiTeam et al., 2025]	1.28	62.2	65.2
Audio Flamingo 3 [Goel et al., 2025]	1.57	61.4	72.4
Academic, open, or public-data speech/audio LLMs			
LTU-AS [Gong et al., 2023]	4.9	N.R.	N.R.
BLSP+RP [Wang et al., 2023]	6.4	N.R.	N.R.
WavLLM [Hu et al., 2024]	**2.0**	N.R.	N.R.
SALMONN [Tang et al., 2024]	2.1	30.0	32.8
AlignFormer [Fan et al., 2025]	3.52	N.R.	N.R.
C-Gate-3T (ours)	3.98	**60.6**	**48.3**

LS WER: 3.98%
Competitive with academic systems trained on public data.

MMSU: 60.6%
Only 1.6pp below the best; 7.3pp above Qwen2-Audio-Instruct.

Scale matters
The gap to SOTA mainly reflects far less training data and model capacity.

Hsu et al., **Is Text All You Need? Text as a Universal Information Bottleneck for Speech LLMs**. 2026

Responsible Assessment

A strong model requires equally robust and responsible evaluation.

Position Article

Beyond MOS

A stance on responsible TTS assessment: moving from a single MOS score to more comprehensive metrics: naturalness, speaker consistency, and content credibility.

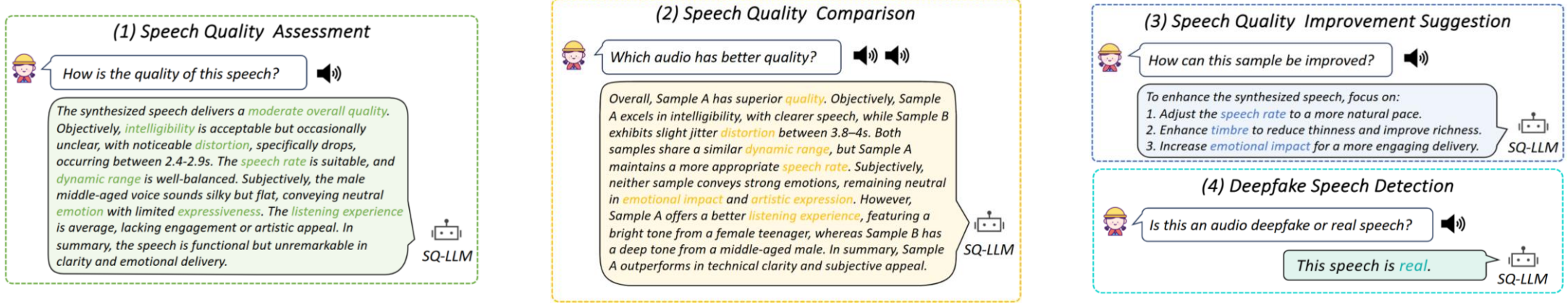
SpeechLLM-as-Judges

Using SpeechLLM to Evaluate Speech

Automatically assessing speech quality with large speech models—while carefully considering the evaluator's own biases, robustness, and interpretability.

What we need to improve is not just the models we build,
but also the way we measure how we "listen, speak, and see."

SpeechLLM-as-Judges: Beyond Scalar Scores



FOUR SPEECH EVALUATION TASKS

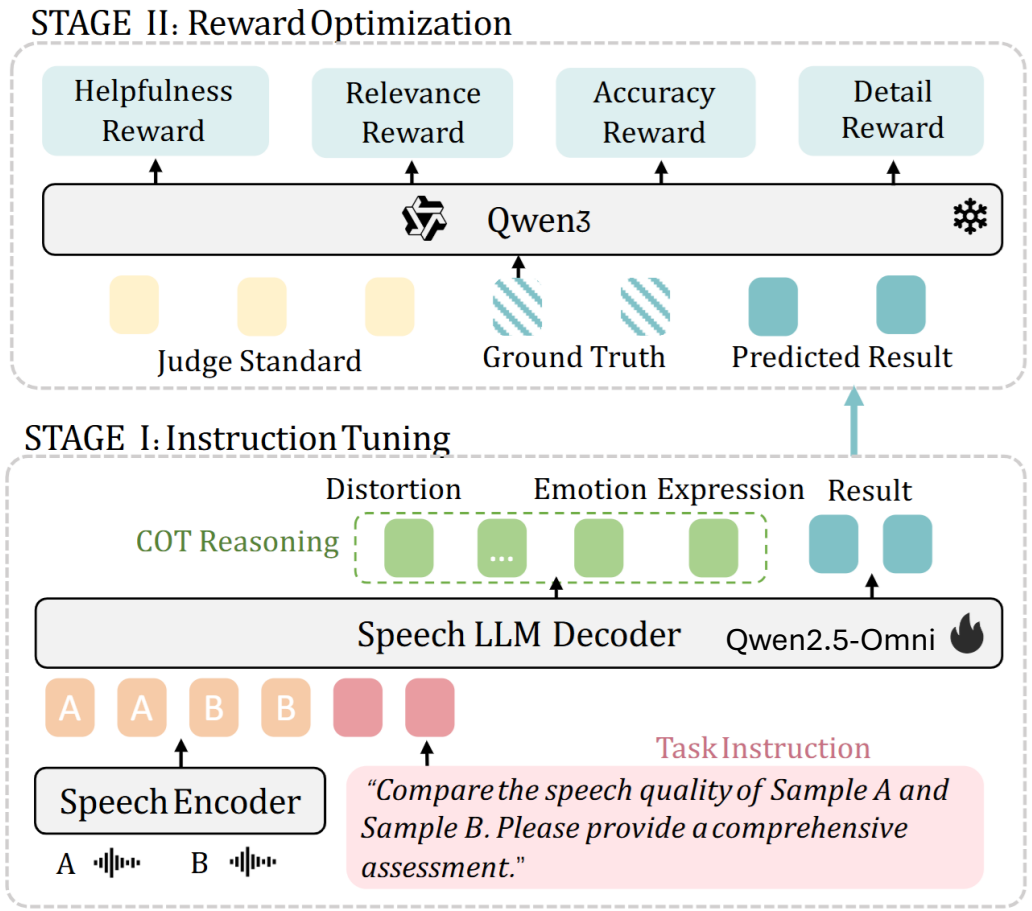
Dataset	# Label	Languages	Quality Assessment		Quality Comparison		Quality Suggestion	Deepfake Detection
			Form	# Dim	Form	# Dim		
ASVspoof2019-LA (Wang et al., 2020)	121,461	English	✗	—	✗	—	✗	✓
VCC2018 (Lorenzo-Trueba et al., 2018)	113,168*	English	Scores	2	✗	—	✗	✗
BC2019 (Wu et al., 2019)	812	Chinese	Scores	1	✗	—	✗	✗
BVCC (Cooper and Yamagishi, 2021)	7,106	English	Scores	1	✗	—	✗	✗
NISQA (Mittag et al., 2021)	14,672	English	Scores	5	✗	—	✗	✗
QualiSpeech (Wang et al., 2025c)	14,577	English	Natural-language (Human-annotated)	7+4	✗	—	✗	✗
ALLD-dataset (Chen et al., 2025a)	25,680	English	Natural-language (LLM-generated)	5	Natural-language (LLM-generated)	5	✗	✗
SpeechEval(OURS)	**128,754**	**Chinese, English, Japanese & French**	**Natural-language (Human-annotated)**	**8+3+5**	**Natural-language (Human-annotated)**	**8+3+5**	✓	✓

SPEECHEVAL VS. PRIOR SPEECH EVALUATION DATASETS

Wang et al., **SpeechLLM-as-Judges:Towards General and Interpretable Speech Quality Evaluation**, 2025

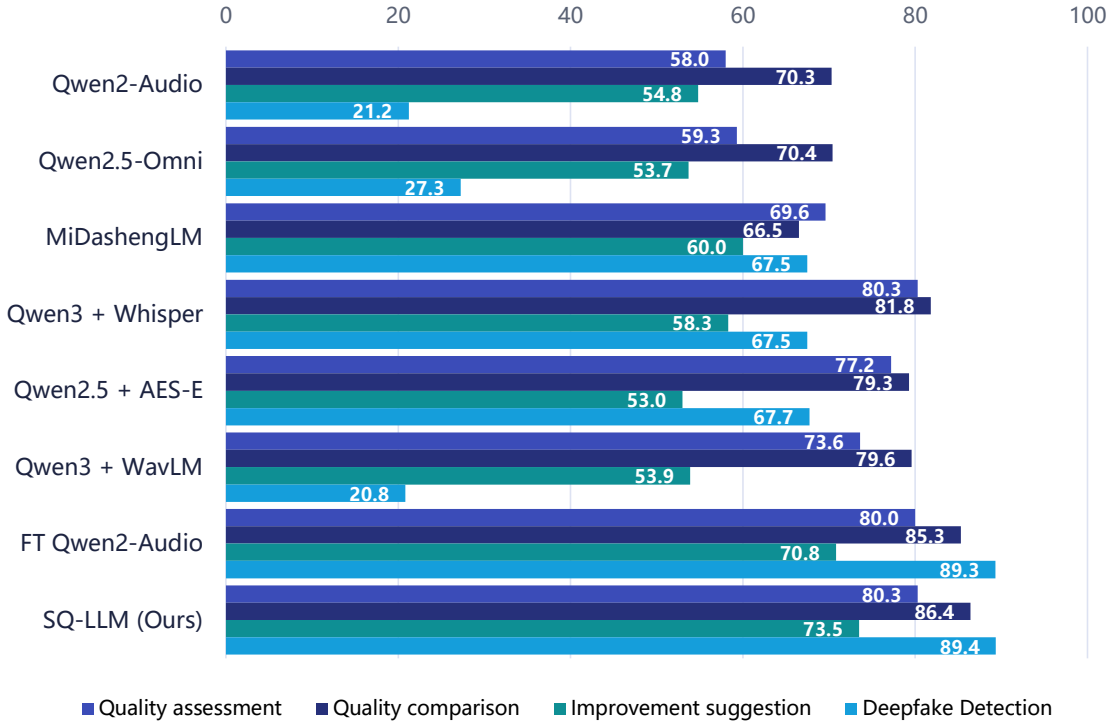
SpeechLLM-as-Judges: Beyond Scalar Scores

SQ-LLM JUDGE MODEL ARCHITECTURE AND TRAINING.



PERFORMANCE ON ASSESSMENT, COMPARISON, SUGGESTION, AND DEEPFAKE DETECTION.

SQ-LLM is consistently strongest across the four tasks



SBERT for quality assessment, comparison and improvement suggestion, and accuracy for deepfake detection.

Wang et al., **SpeechLLM-as-Judges:Towards General and Interpretable Speech Quality Evaluation**, 2025

From Multimodal LLMs to Multimodal Agents

Represent

Encode speech and video as model-ready token sequences.

Align

Preserve semantics, timing, and context across modalities.

Evaluate

Measure capability, robustness, and responsibility—not one scalar score.

Perceive in Real Time

Understand speech, video, and the environment with low latency.

Joint Reasoning

Fuse signals without sacrificing the reasoning strength of text LLMs.

Act Reliably

Use tools and take grounded actions with interpretable safeguards.

Thank you

Multimodal LLMs → Multimodal Agents