



Towards End-to-End Spoken Dialogue Models

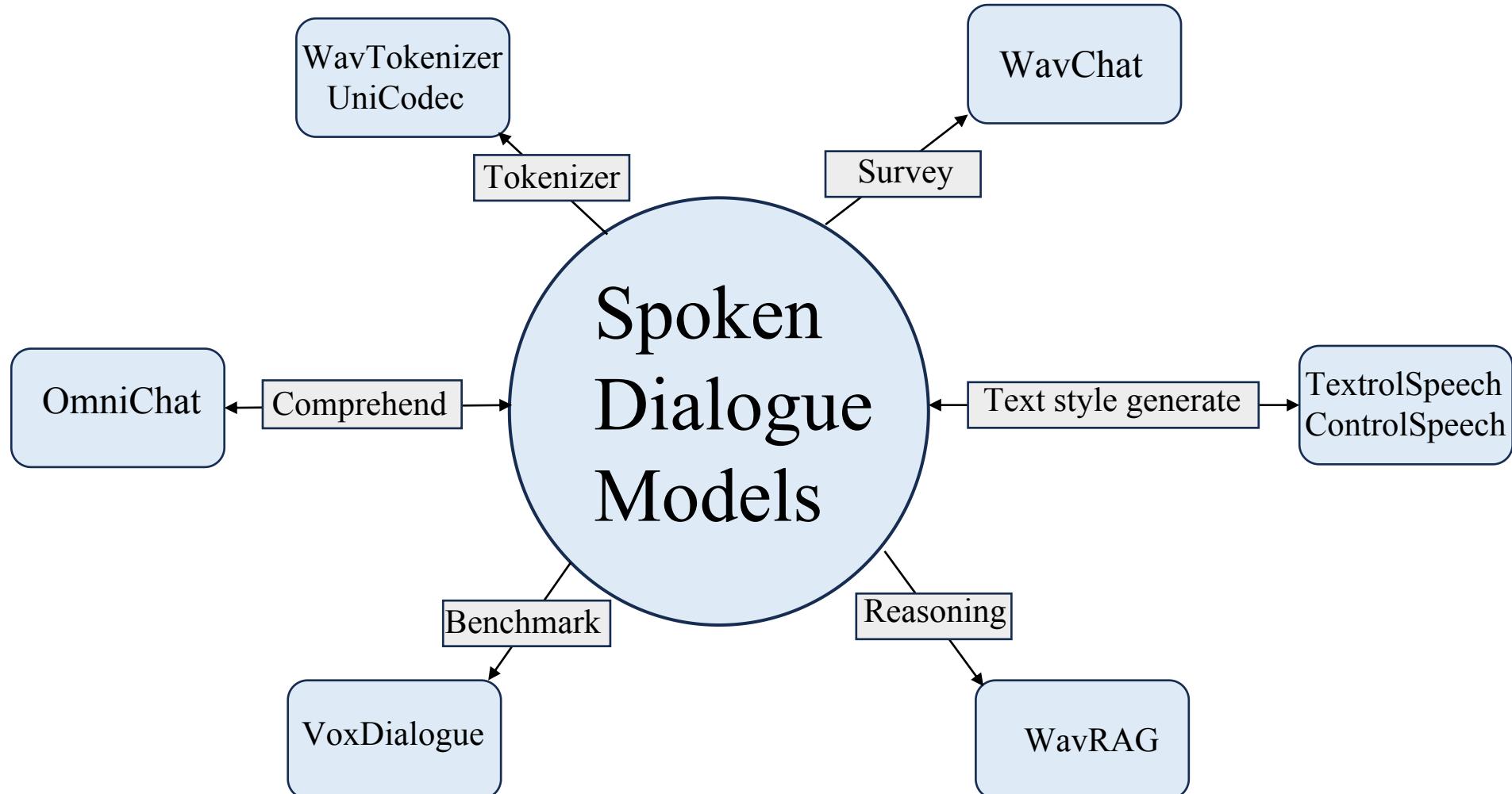


季圣鹏



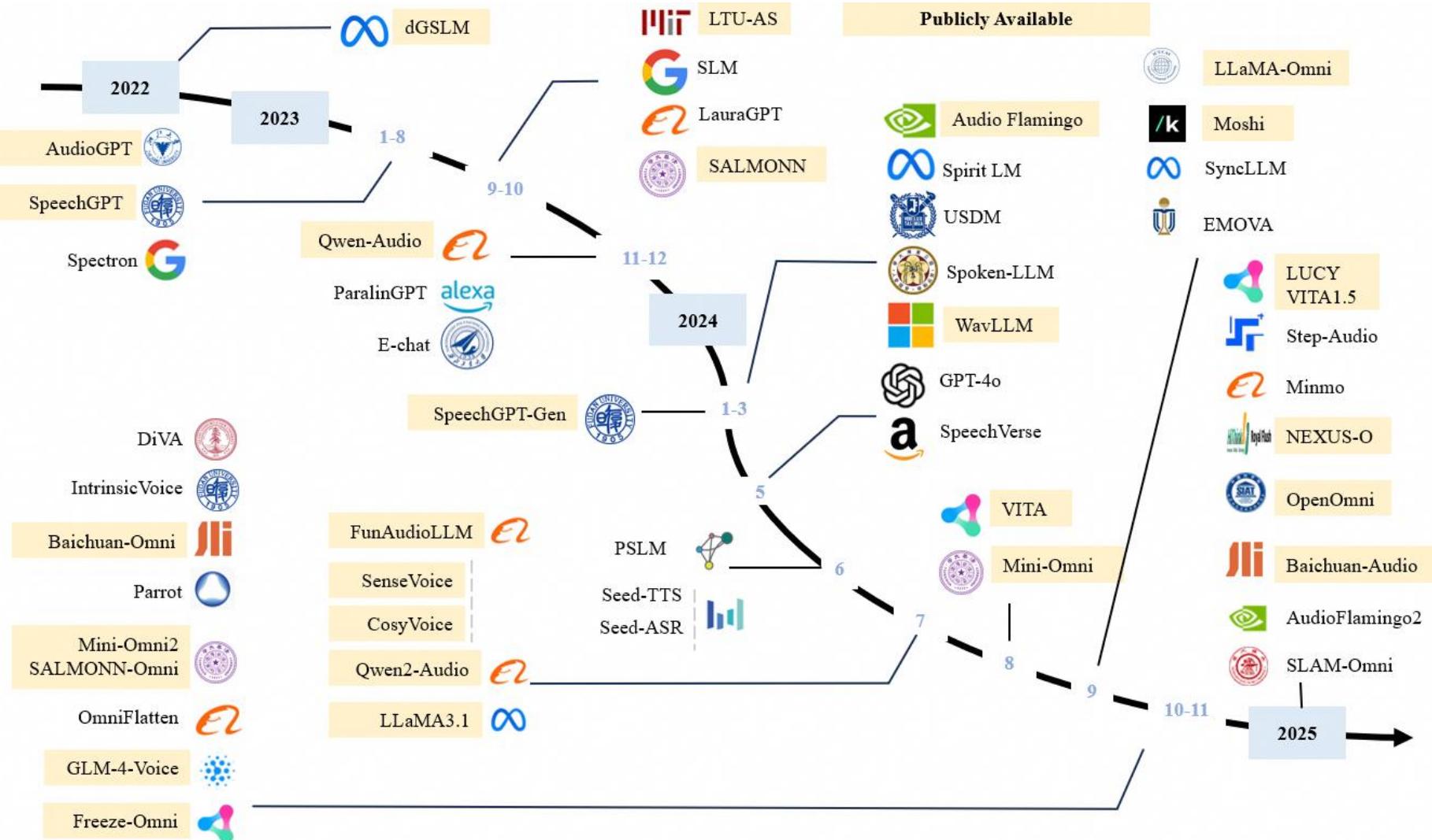
导师：赵洲

<https://novateurjsp.github.io>





概述





WavChat

- 右边是九种用例场景

- 对话的智商和情商（语音智能）非常重要

- 综合的对话能力->AGI

(a) Text Intelligence

Human: Hi, can you give me some travel tips for Japan?

Bot: Absolutely! Japan is a fantastic destination with a wonderful mix of traditional culture and modern attractions. If you're heading to Tokyo, make sure to visit the historic Senso-ji Temple and experience the lively atmosphere at Shibuya Crossing. Don't miss out on the incredible views from Tokyo Tower and be sure to try some authentic sushi and ramen while you're there. Kyoto is perfect for immersing yourself in traditional Japanese culture, with beautiful sites like the Golden Pavilion and Kiyomizu-dera Temple. You can also enjoy a serene tea ceremony in one of the local tea houses. If you're a foodie, Osaka is a must-visit with its delicious takoyaki and okonomiyaki...

(b) Speech Intelligence

Human: (sad tone): I've been feeling down lately. Can you talk with me?

Bot: (gentle, encouraging tone, gentle mezzo-soprano): Of course, I'm here for you. I'm sorry to hear you're feeling this way. Remember, you're strong enough to overcome them. Anything specific on your mind that I can help with?

Human: Can you tell me a joke in a humorous tenor voice?

Bot: (humorous tone, tenor): Sure thing! Here goes: How do trees get on the Internet? They log in!

(c) Audio and Music Generation

Human: I love the sound of birds chirping. Can you play that for me?

Bot: Certainly, here's a relaxing bird song for you.

(Playing bird chirping sound)

Human: That's perfect! Now, could you play some soothing piano music?

Bot: Sure thing, here's a gentle piano melody.

(Playing calming piano music)

(d) Audio and Music Understanding

Human: Hey, I just recorded an audio of my dog chasing his tail and barking excitedly. Can you write a funny poem based on it?

Bot: (An audio of a dog chasing his tail and barking excitedly)

Sounds like a lot of fun! Here's a playful poem inspired by your energetic pup:

(lyrical, human tone):

Round and round he spins with glee,
A fluffy tornado, wild and free.
With barks that echo, loud and clear,
He's the happiest pup far and near!
His tail's his goal, he chases fast,
In his mind, he'll catch it at last.
But round he goes, a fury spree,
A doggy dance, pure jubilee!

(e) Multilingual Capability

Human: Hi, could you translate this into French for me? "The weather is beautiful today, and I'd like to go for a walk in the park."

Bot: Absolutely! The translation in French is: "Il fait beau aujourd'hui, et j'aimerais aller me promener dans le parc."

Human: Thanks! Now translate this to Spanish: "I'm very happy to meet you."

Bot: In Spanish, that would be: "Estoy muy feliz de conocerte."

(f) Context Learning

Human: So far, you have a pair of running shoes, a water bottle, and a set of wireless earbuds in your shopping list. Need anything else?

Bot: Can you add milk and bread to my shopping list?

Human: Here's the updated list: you've got running shoes, a water bottle, wireless earbuds, milk, and bread. Anything else you'd like to add, or are we all set?

Bot: Actually, let's throw in a box of granola bars too.

Human: Got it—adding a box of granola bars. So now your cart has running shoes, a water bottle, wireless earbuds, milk, bread, and granola bars. Looking good!

(g) Interaction Capability

Human: So, I was thinking we could try that new-sushi place.

Bot: (interrupts): Oh, would you like a recommendation for a new Italian restaurant nearby?

Human: Oh! You mean the Italian place downtown?

Bot: Yes, there's a popular Italian restaurant nearby. Would you like — to give that a shot?

Human: (interrupts): Eh, actually, scratch that! I was kind of leaning toward sushi.

Bot: Understood. I can suggest several sushi places nearby. How about we check out that one with great reviews?

(h) Streaming Latency

Human: I'd like some help writing a birthday message for a friend. Could you help me?

Bot: Absolutely! Here's a draft for you: I hope your special day is filled with joy, laughter, and wonderful memories that you'll cherish for years to come. Today is all about celebrating you and the amazing person you are. Remember that I'm always here for you, supporting you through every high and low, and I truly cherish every moment we have!

LLM stream

Bot: Absolutely! Here's a draft for you: I hope your special day is filled with joy, laughter, and wonderful memories that you'll cherish for years to come. Today is all about celebrating you and the amazing person you are. Remember that I'm always here for you, supporting you through

(i) Multimodal Capability

Human: I just took a video of my hike. Can you help me describe it?

Bot: Based on your video, I see you were hiking through a forested area with some beautiful mountain views in the distance. The lighting suggests it was a sunny morning, and the atmosphere looks serene and refreshing. Would you like me to add any background music to your video?



级联式语音对话模型

- 高时延
- 累计误差
- 互动局限性
- 以文本为中心，无法处理非文本信息

右图的代表性模型，自上而下

- AudioGPT
- Qwen-Audio
- SLAM-Omni
- LLaMA-Omni
- Mini-Omni
- Moshi

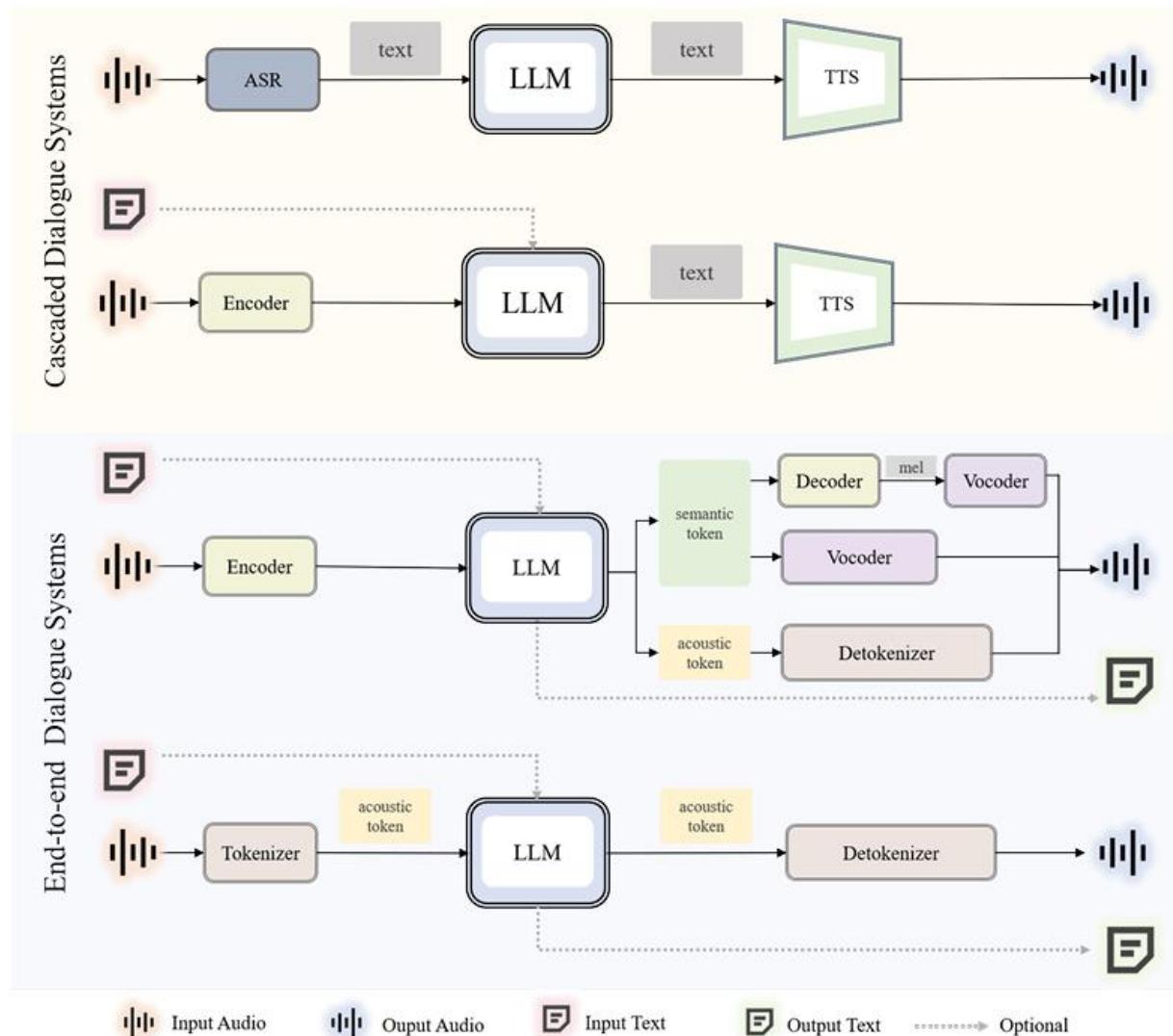


Figure 2: A general overview of current spoken dialogue systems. We categorize these systems into two paradigms, cascaded spoken dialogue models and end-to-end spoken dialogue models, based on whether the core language model can **directly** understand and generate speech representations. Additionally, we provide a visualization of the input and output methods used in different spoken dialogue systems.



1. Codec语音类表征领域的快速发展

SoundStream/Encoder开创了端到端 codec 模型的范式，DAC极大地提高了 acoustic codec 模型的重建质量，SpeechTokenizer/SemanticCodec 尝试在 codec 模型中加入 semantic 信息缩小 codec 与下游 LLM 的 Gap，HiFi-Codec/SNAC/Language-Codec 离散声学表征在生成模型上验证了可行性

ParlerTTS 在风格可控 TTS 上达到了很好的生成效果，VALL-E2 可以生成接近人类主观听觉的音频。

Codec 优势：重建的架构训练简单；包含语音特有声学信息；统一建模 audio, music, speech；离散化 Token 适配 LLM 的 CE loss；Token 化可以和文本结合。GPT-4o 端到端语音对话系统的火热，多模态大模型的一种重要范式，简单的 Tokenizer 和 DeTokenizer 的趋势

以 Qwen-Audio 和 Mini-Omni 为例，在理解端是 Whisper 的架构，在生成端是 codec 架构，历史信息的互动困难

以 Codec 模型为基础，长远来看完成 Tokenizer 化和 DeTokenizer 化

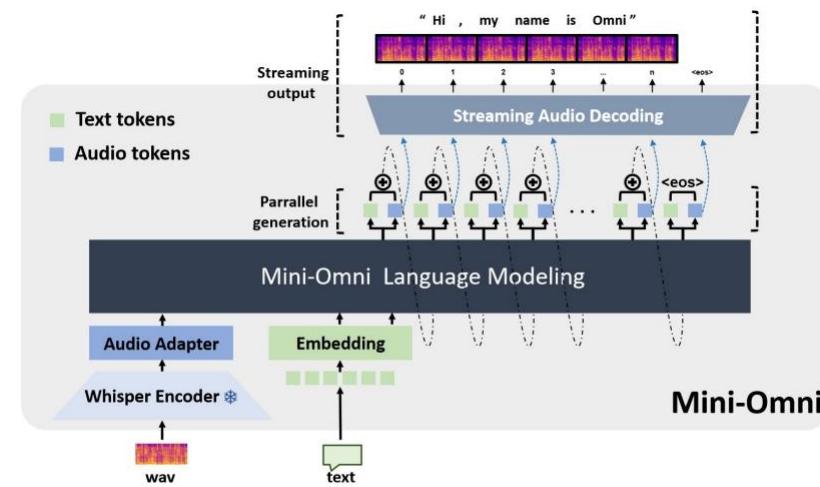
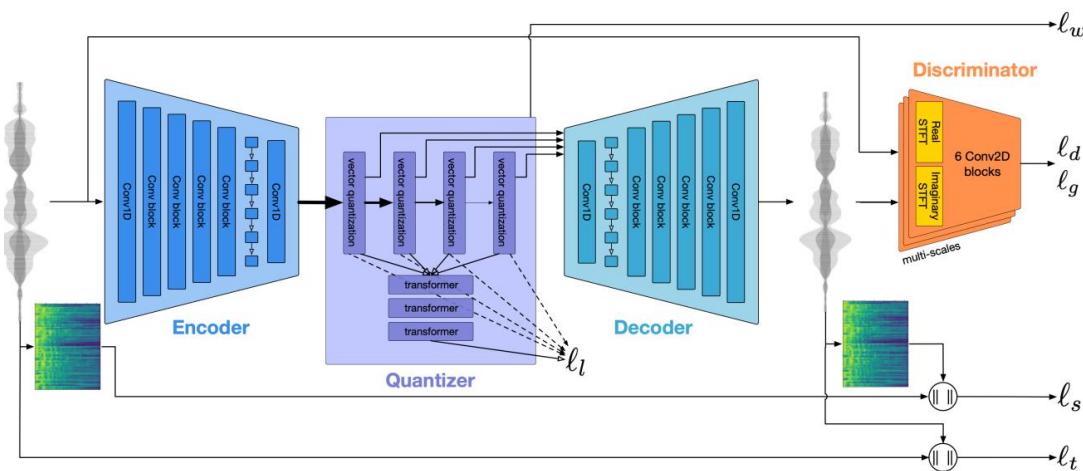


Figure 1: The Mini-Omni model architecture.



1. 现有Codec model Token数量生成过多, 引出下游CodecLLM的各种生成范式。 (需要探索低码率下的重建性能)

9层的44.1khzDAC模型需要900个Token, 4层的HiFi-Codec模型需要300个Token, 过多的Token限制语言模型的生成能力 (单层和多层具有本质区别)

2. 如何从Codec本身出发, 增强Semantic信息, 缩小重建范式和LLM之前的gap, 统一CodecLLM理解和生成

蒸馏的方式会限制Codec模型的质量上限

不优雅, 限制统一建模music, audio和music能力

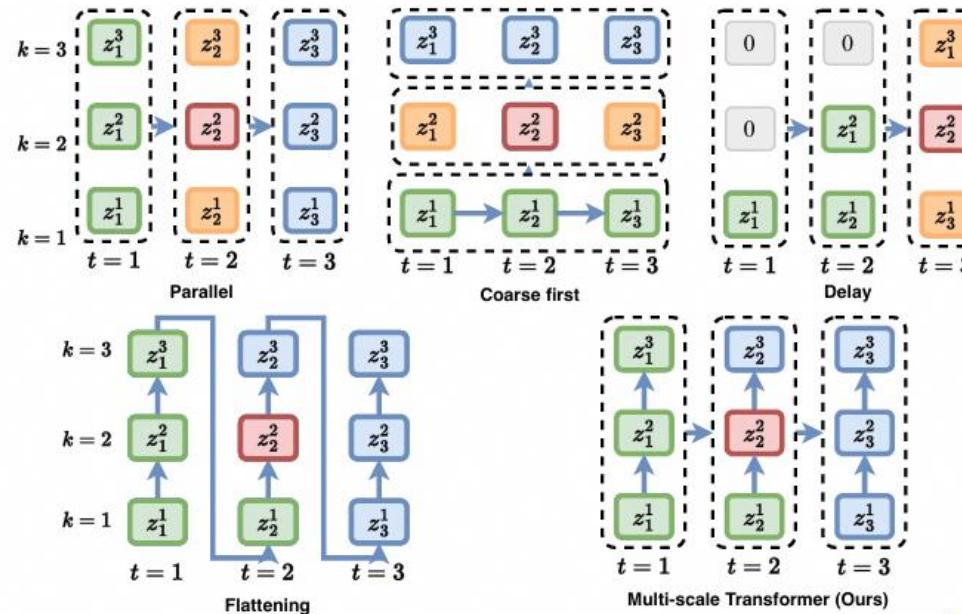


Figure 2: Order of token prediction for 4 representative methods in audio generation (Copet et al., 2023) and the proposed multi-scale Transformer. Assume $n_q = 3$ and $T = 3$. Current token prediction (red) is conditioned on prior tokens (in green). Tokens in orange are concurrently predicted with the current token. 0 is a special token indicating empty positions in the delay prediction.

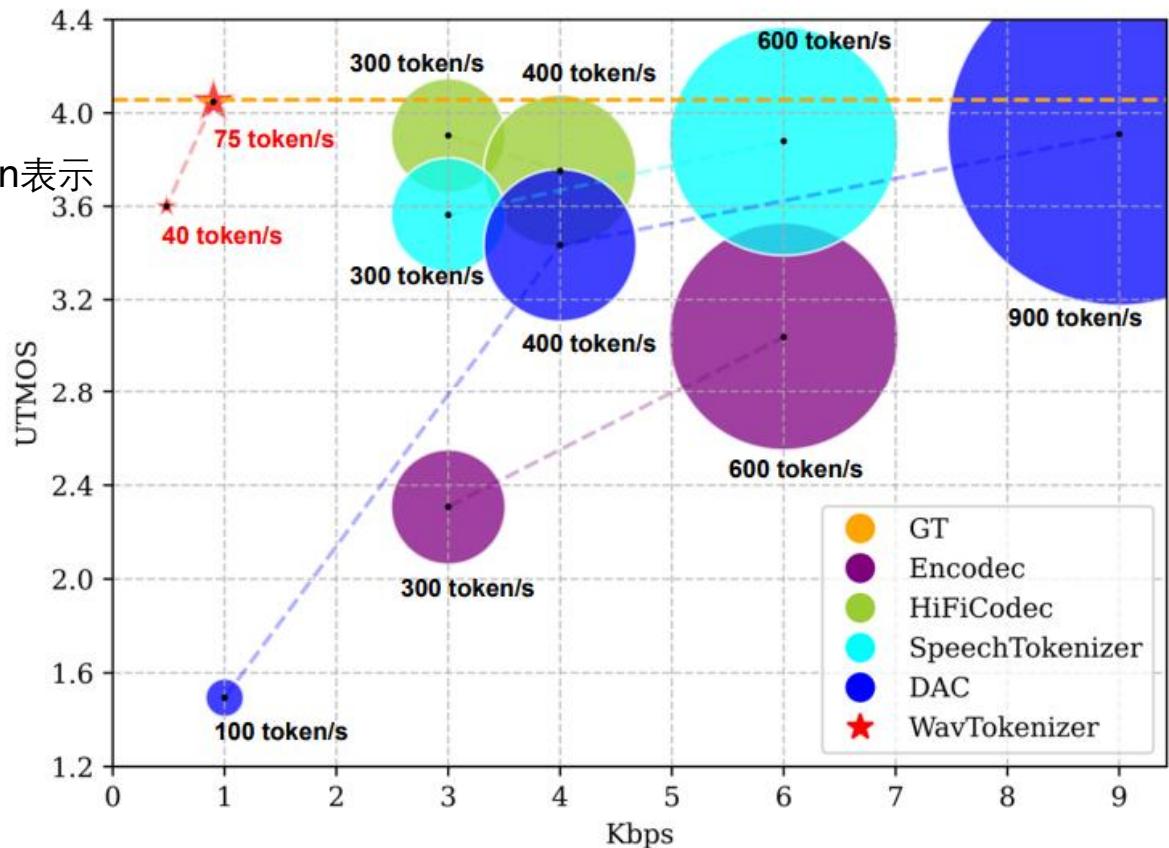


1. 实验效果上

- 极限的Token数量，一秒的语音类表征（24khz）可以仅有40个Token表示
- 良好的主观重建效果，UTMOS4.0，并且包含丰富的语义信息

2. 方法设计上

- 验证大码本空间VQ在极限压缩下的潜力
- 首次提出单层量化器的范式，潜在的语音作为特殊文本语言的对齐能力
- 验证Transformer架构在Codec重建任务上不会拥有长度外推的问题
- 引入更长的上下文建模窗口，引入直接逆傅里叶变换上采样模型，引入多尺度判别器





CodeBook的空间大小或者宽度?

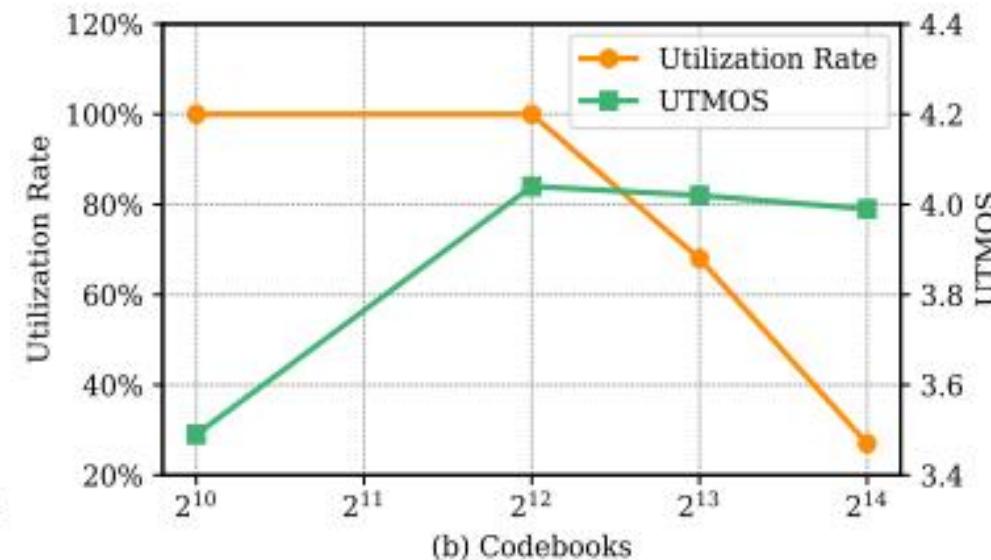
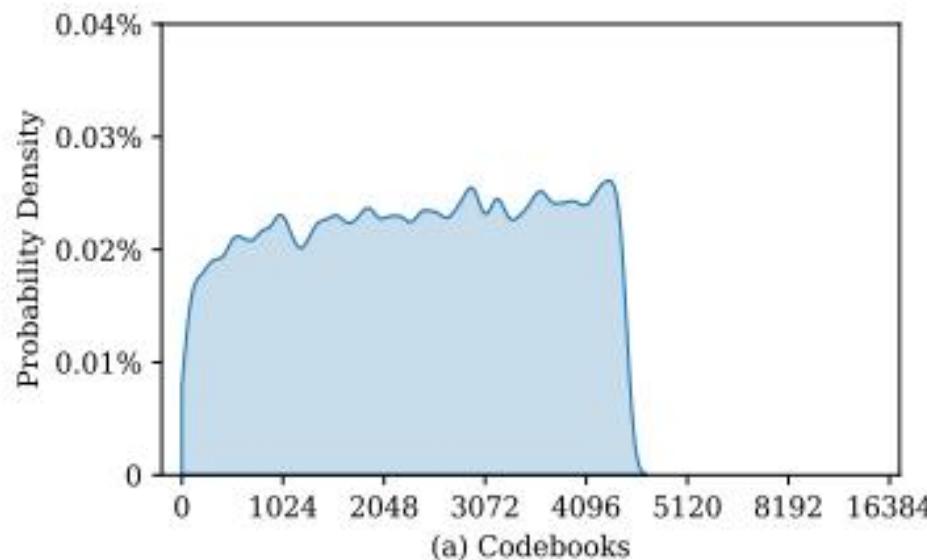


Table 8: The ablation study investigates the impact of dataset size on codebook utilization.

Model	Dataset	Codebooks	Utilization rate	UTMOS ↑	PESQ ↑	STOI ↑
WavTokenizer	585 Hours	16384	27%	3.9989	2.3600	0.8129
WavTokenizer	4000 Hours	16384	26.5%	3.9465	2.3721	0.8217

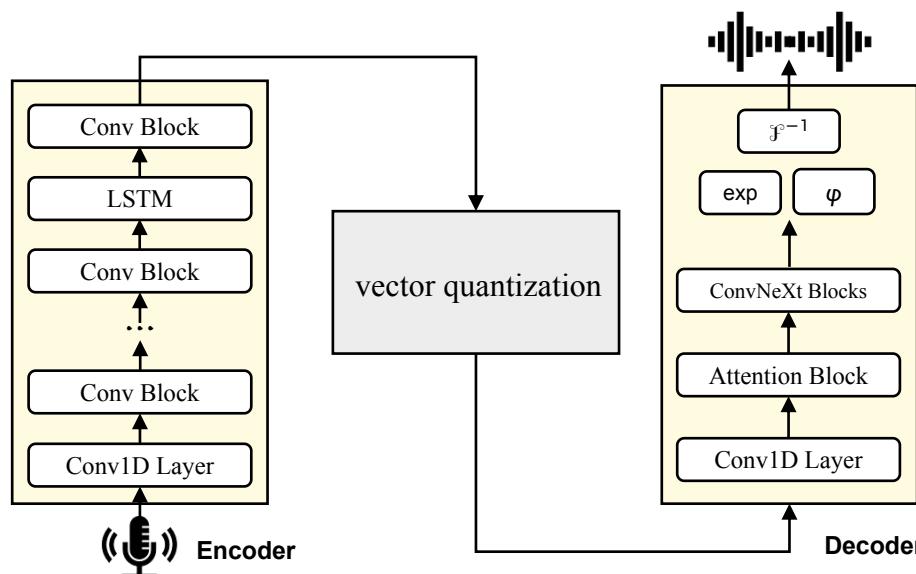


Table 6: Impact of the contextual modeling window size.

Model	Codebooks	windows	UTMOS ↑	PESQ ↑	STOI ↑
WavTokenizer	4096	1	3.7448	2.0112	0.8944
WavTokenizer	4096	3	4.0486	2.3730	0.9139
WavTokenizer	4096	5	4.0448	2.3556	0.9127

Table 7: Ablation on the multi-scale STFT discriminator (MSTFTD), the attention module, and switching from our improved decoder to a mirror decoder, in WavTokenizer.

Model	UTMOS ↑	PESQ ↑	STOI ↑	V/UV F1 ↑
WavTokenizer	4.0486	2.3730	0.9139	0.9382
w/ mirror decoder	2.7782	1.5007	0.8249	0.8820
w/o attention module	3.6020	1.9332	0.8734	0.9067
w/o MSTFTD	3.7806	2.1270	0.9008	0.9269



Table 1: **Objective reconstruction results** of different codec models on LibriTTS *test-clean* (clean environment), LibriTTS *test-other* (noisy environment), and *LJSpeech dataset* (out-of-domain environment). **Nq** denotes the number of quantizers. **GT** denotes ground truth waveforms. Best results from models with a single quantizer (hence directly comparable to WavTokenizer) are in bold.

Dataset	Model	Bandwidth ↓	Nq ↓	token/s ↓	UTMOS ↑	PESQ ↑	STOI ↑	V/UV F1 ↑
LibriTTS <i>test-clean</i>	GT	-	-	-	4.0562	-	-	-
	DAC	9.0kbps	9	900	3.9097	3.9082	0.9699	0.9781
	Encodec	6.0kbps	8	600	3.0399	2.7202	0.9391	0.9527
	Vocos	6.0kbps	8	600	3.6954	2.8069	0.9426	0.9437
	SpeechTokenizer	6.0kbps	8	600	3.8794	2.6121	0.9165	0.9495
	DAC	4.0kbps	4	400	3.4329	2.7378	0.9280	0.9572
	HiFi-Codec	3.0kbps	4	400	3.7529	2.9611	0.9405	0.9617
	HiFi-Codec	4.0kbps	4	300	3.9035	3.0116	0.9446	0.9576
	Encodec	3.0kbps	4	300	2.3070	2.0517	0.9007	0.9198
	Vocos	3.0kbps	4	300	3.5390	2.4026	0.9231	0.9358
	SpeechTokenizer	3.0kbps	4	300	3.5632	1.9311	0.8778	0.9273
	DAC	1.0kbps	1	100	1.4940	1.2464	0.7706	0.7941
	WavTokenizer	0.5kbps	1	40	3.6016	1.7027	0.8615	0.9173
	WavTokenizer	0.9kbps	1	75	4.0486	2.3730	0.9139	0.9382
LibriTTS <i>test-other</i>	GT	-	-	-	3.4831	-	-	-
	DAC	9.0kbps	9	900	3.3566	3.7595	0.9576	0.9696
	Encodec	6.0kbps	8	600	2.6568	2.6818	0.9241	0.9338
	Vocos	6.0kbps	8	600	3.1956	2.5590	0.9209	0.9202
	SpeechTokenizer	6.0kbps	8	600	3.2851	2.3269	0.8811	0.9205
	DAC	4.0kbps	4	400	2.9448	2.5948	0.9083	0.9404
	HiFi-Codec	4.0kbps	4	400	3.0750	2.5536	0.9126	0.9387
	HiFi-Codec	3.0kbps	4	300	3.3034	2.6083	0.9166	0.9318
	Encodec	3.0kbps	4	300	2.0883	2.0520	0.8835	0.8926
	Vocos	3.0kbps	4	300	3.0558	2.1933	0.8967	0.9051
	SpeechTokenizer	3.0kbps	4	300	3.0183	1.7373	0.8371	0.8907
	DAC	1.0kbps	1	100	1.4986	1.2454	0.7505	0.7775
	WavTokenizer	0.5kbps	1	40	3.0545	1.6622	0.8336	0.8953
	WavTokenizer	0.9kbps	1	75	3.4312	2.2614	0.8907	0.9172
<i>LJSpeech</i>	GT	-	-	-	4.3794	-	-	-
	DAC	9.0kbps	9	900	4.3007	3.9022	0.9733	0.9757
	Encodec	6.0kbps	8	600	3.2286	2.6633	0.9441	0.9555
	Vocos	6.0kbps	8	600	4.0332	2.9258	0.9497	0.9459
	SpeechTokenizer	6.0kbps	8	600	4.2373	2.6413	0.9316	0.9452
	DAC	4.0kbps	4	400	3.8109	2.7616	0.9338	0.9524
	HiFi-Codec	4.0kbps	4	400	4.1656	2.7629	0.9446	0.9497
	HiFi-Codec	3.0kbps	4	300	4.2692	2.9091	0.9485	0.9469
	Encodec	3.0kbps	4	300	2.3905	2.0194	0.9058	0.9326
	Vocos	3.0kbps	4	300	3.7880	2.5006	0.9310	0.9388
	SpeechTokenizer	3.0kbps	4	300	3.9908	2.0458	0.9021	0.9299
	DAC	1.0kbps	1	100	1.4438	1.2084	0.7822	0.8095
	WavTokenizer	0.5kbps	1	40	4.0186	2.1142	0.9093	0.9406
	WavTokenizer	0.9kbps	1	75	4.2580	2.4923	0.9312	0.9397

Table 2: The **subjective reconstruction results** using MUSHRA (comparative scoring of samples) of codec models on speech, music and audio domains. **Nq** denotes the **number of quantizers**.

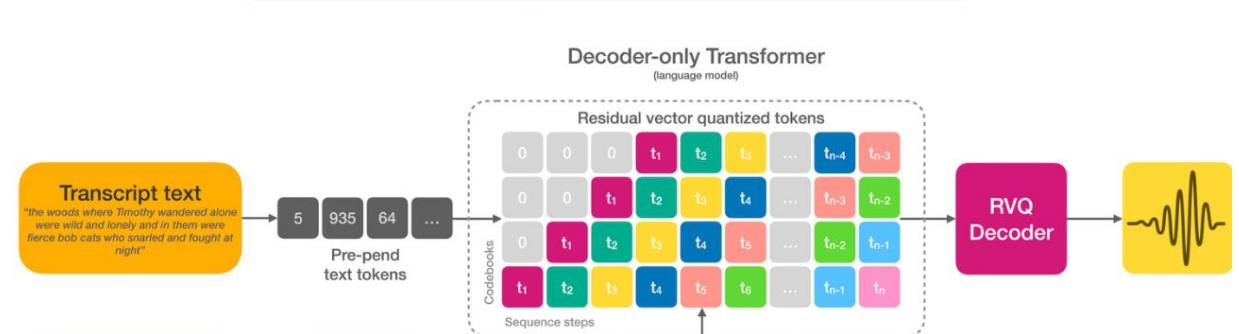
Model	Bandwidth ↓	Nq ↓	token/s ↓	LibriTTS test-clean ↑	MusicDB ↑	Audioset ↑
GT	-	-	-	96.4±1.2	95.3±1.7	95.8±2.1
DAC	9.0kbps	9	900	92.8±1.8	92.6±2.4	92.7±1.5
Encodec	6.0kbps	8	600	78.6±1.9	76.9±1.6	81.2±1.8
DAC	1.0kbps	1	100	58.4±2.4	57.6±2.1	56.8±1.4
WavTokenizer	0.9kbps	1	75	96.1±2.3	92.9±2.2	94.4±1.6

Table 3: The **semantic representation (speech, music, audio)** evaluation of different codec models on ARCH Benchmark in terms of **classification accuracy**. **Nq** represents the **number of quantizers**.

Model	Nq ↓	token/s ↓	RAVDESS ↑	SLURP ↑	EMOVO ↑	AM ↑	FMA ↑	MTT ↑	IRMAS ↑	MS-DB ↑	ESC50 ↑	US8K ↑	FSD50K ↑	VIVAE ↑
DAC	9	900	0.3750	0.0779	0.2363	0.6926	0.3504	0.2805	0.4023	0.6014	0.2594	0.4032	0.1297	0.3440
Encodec	8	600	0.2881	0.0636	0.2261	0.4388	0.2790	0.1993	0.3671	0.3917	0.1925	0.3055	0.1091	0.3005
DAC	4	400	0.3194	0.0782	0.2346	0.6838	0.3379	0.2784	0.3833	0.5942	0.2580	0.3824	0.1293	0.3342
Encodec	4	300	0.2951	0.0660	0.2193	0.4301	0.2728	0.1934	0.3684	0.3656	0.1790	0.3097	0.1099	0.2710
Encodec	2	150	0.2743	0.0627	0.2193	0.3649	0.2816	0.1900	0.3704	0.3245	0.1699	0.2960	0.1065	0.2630
DAC	1	100	0.2500	0.0713	0.2278	0.3287	0.3304	0.2502	0.3572	0.5137	0.2065	0.3350	0.1295	0.2991
WavTokenizer	1	75	0.3255	0.0802	0.3163	0.6957	0.3417	0.2835	0.4117	0.5764	0.2550	0.3975	0.1392	0.3563

Table 4: The **subjective evaluations** of various acoustic codec models for downstream speech synthesis models. **GT** denotes ground truth waveforms.

Model	Bandwidth ↓	Nq ↓	CMOS-Q↑	CMOS-P↑
GT	-	-	0.22	0.26
DAC	9.0kbps	9	-0.35	-0.29
WavTokenizer	0.9kbps	1	0.00	0.00





进一步增强在统一Audio, Music和Speech情况下的Codec重建性能和Semantic语义丰富性

Table 1: Comparison of recent codec models based on single codebook, compatibility with speech, music, and sound domains, and whether they use *separate* models for different domains or a *unified* model.

Model	Single Codebook	Speech	Music&Sound	Separate/Unified model
DAC (Kumar et al., 2023)	✗	✓	✓	Unified
Encodec (Défossez et al., 2023)	✗	✓	✓	Unified
Mimi (Défossez et al., 2024)	✗	✓	✓	Unified
SemantiCodec (Liu et al., 2024)	✗	✓	✓	Unified
SpeechTokenizer (Zhang et al., 2023)	✗	✓	✗	-
BigCodec (Xin et al., 2024)	✓	✓	✗	-
TAAE (Parker et al., 2024)	✓	✓	✗	-
Wavtokenizer (Ji et al., 2024c)	✓	✓	✓	Separate&Unified
UniCodec	✓	✓	✓	Unified



- Encoder的transformer网络加SimVQ量化
- Partitioned Codebook
- MOE网络
- 多阶段的semantic stage对比学习loss

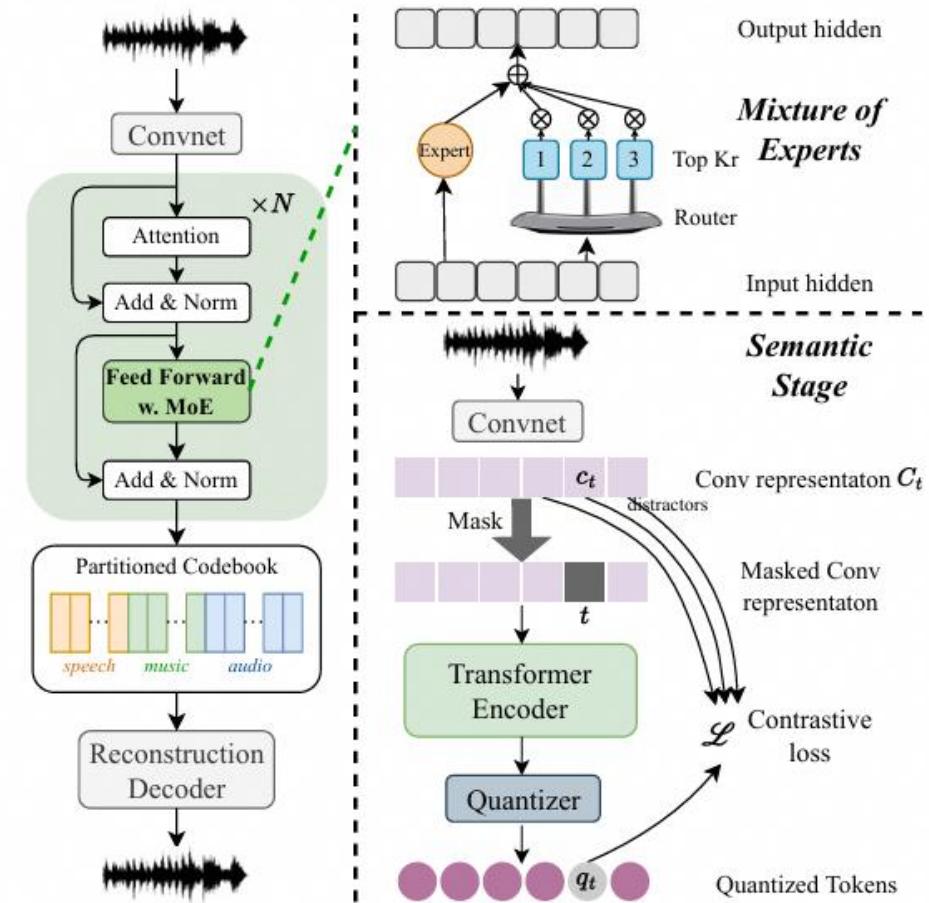


Figure 1: Left: Overview of the proposed UniCodec. Upper-right: the domain MoE encoder structure. Lower-right: the semantic training stage.



Table 2: **Objective reconstruction results** of UniCodec and baselines on **speech, music and audio** domains on LibriTTS test-clean, MusicDB test set, and AudioSet eval set, in terms of Mel Distance and STFT Distance. **TPS** denotes token per second. We **bold** the best results in all the models, and **bold and underline** the best results in single-codebook codec models.

Model	Unified	TPS↓	LibriTTS test-clean		MusicDB test		AudioSet eval	
			Mel Dist↓	STFT Dist↓	Mel Dist↓	STFT Dist↓	Mel Dist↓	STFT Dist↓
DAC	✓	600	0.3697	1.5525	0.3578	1.9621	0.4581	2.1378
Encodec	✓	600	0.5367	1.8271	0.5565	2.1678	0.7601	2.6273
Mimi	✓	100	0.6709	1.9859	0.6714	2.2526	0.8406	2.6639
TAAE	✗	50	0.7508	2.2426	1.4067	4.1340	1.9335	5.2897
DAC	✗	75	0.7217	2.1662	1.8894	6.2476	1.7063	5.2923
BigCodec	✗	80	0.4427	1.7385	1.3803	4.2366	1.8632	5.6171
Wavtokenizer (speech)	✗	75	0.5001	1.7879	0.6586	3.0335	0.5990	2.5479
Wavtokenizer (music/audio)	✗	75	0.5451	1.8649	0.4516	2.2450	0.4536	2.1871
Wavtokenizer (unified)	✓	75	0.5308	1.8614	0.5435	2.5451	0.5193	2.3727
UniCodec (Ours)	✓	75	0.3442	1.5147	0.3959	2.1822	0.3820	2.1065

Table 3: **Objective reconstruction results** on the **Speech** domain from UniCodec and baselines on LibriTTS test-clean, in terms of naturalness, distortion, and intelligibility. **DR** denotes the Downsample Rate (the input audio sample rate division by Tokens Per Second (TPS)). **Unified** denotes the codec model can support all three domains of speech, music, and sound. The results of models marked by [†] are cited from the Wavtokenizer paper (Ji et al., 2024c) and others are reproduced by us based on the checkpoints released by the corresponding work.

Model	Unified	DR (↑)	TPF (↓)	TPS (↓)	PESQ (↑)	STOI (↑)	F1 (↑)	UTMOS (↑)
Ground Truth [†]	-	-	-	-	-	-	-	4.0562
DAC	✓	40	8	600	3.5197	0.9709	0.9546	3.6905
Encodec [†]	✓	40	8	600	2.7202	0.9391	0.9527	3.0399
SpeechTokenizer [†]	✗	40	8	600	2.6121	0.9165	0.9495	3.8794
Mimi	✓	240	8	100	2.2695	0.9118	0.912	3.5731
TAAE	✗	320	2	50	1.8955	0.8816	0.9260	4.1389
DAC	✗	320	1	75	1.1763	0.7739	0.7560	1.3531
BigCodec	✗	200	1	80	2.6872	0.9293	0.9480	4.0367
Wavtokenizer (speech) [†]	✗	320	1	75	2.3730	0.9139	0.9382	4.0486
Wavtokenizer (unified)	✓	320	1	75	1.8379	0.8718	0.9175	3.6115
UniCodec (Ours)	✓	320	1	75	3.0266	0.9493	0.9486	3.9873

Table 4: **Subjective MUSHRA test reconstruction results** from codec models on **speech, music and audio** domains, on LibriTTS test-clean, MusicDB test set and AudioSet eval set. We report mean and standard deviation.

Model	Unified	LibriTTS test-clean (↑)	MusicDB test (↑)	AudioSet eval (↑)
Ground Truth	-	93.52 ± 1.99	96.18 ± 1.47	95.28 ± 2.18
Wavtokenizer (speech)	✗	85.44 ± 2.29	-	-
Wavtokenizer (music & audio)	✗	-	75.24 ± 2.38	80.19 ± 2.43
Wavtokenizer (unified)	✓	80.40 ± 2.54	56.10 ± 3.74	62.21 ± 3.42
UniCodec (Ours)	✓	90.74 ± 2.06	77.77 ± 2.45	82.43 ± 2.56

Table 5: **Semantic representation evaluation results** on the ARCH benchmark, in terms of classification accuracy. The results of models marked by [†] are cited from the Wavtokenizer paper (Ji et al., 2024c).

Model	TPS (↓)	Speech		Music		Audio	
		RAVDESS (↑)	AM (↑)	MTT (↑)	MS-DB (↑)	ESC50 (↑)	VIVAE (↑)
Encodec [†]	150	27.43	36.49	19.00	32.45	16.99	26.30
DAC [†]	100	25.00	62.87	25.02	51.37	20.65	29.91
Wavtokenizer (speech) [†]	75	32.55	69.57	-	-	-	-
Wavtokenizer (music&audio) [†]	75	-	-	28.35	57.64	25.50	35.63
UniCodec	75	40.28	70.94	29.55	59.29	26.00	34.17
w/o semantic stage	75	36.81	69.84	28.09	54.05	20.80	30.21

Table 6: Ablation study of UniCodec by evaluating the effects of domain ID during evaluation, the domain MoE module, domain-adaptive codebook, and the semantic training stage and the fine-tuning stage.

Model	LibriTTS test-clean		MusicDB test		AudioSet eval	
	Mel Dist ↓	STFT Dist ↓	Mel Dist ↓	STFT Dist ↓	Mel Dist ↓	STFT Dist ↓
UniCodec	0.3442	1.5147	0.3959	2.1822	0.3820	2.1065
w. domain id	0.3474	1.5151	0.3912	2.1818	0.3824	2.1061
w/o finetune stage	0.4476	1.7005	0.4490	2.2505	0.4366	2.1659
w/o semantic&finetune stage	0.4481	1.6978	0.4534	2.2690	0.4380	2.1723
w/o MoE	0.4883	1.8024	0.4592	2.3153	0.4548	2.2633
w/o partitioned codebook	0.4873	1.7742	0.5064	2.3031	0.5135	2.2382



最近有很多工作尝试进一步增加decoder参数量，更好的encoder-decoder结构，流式的codec，更好的VQ利用率策略，更好的统一acoustic和semantic，期待获得更好的重建效果，更好的平衡CodecLLM的重建和生成，以及更好的统一理解和生成等等。

摘自《Recent Advances in Discrete Speech Tokens: A Review》 离散和连续，AR和Diffusion (Transfusion)

- Low-Bitrate Tokens (在保证参数量和RTF的前提下，语音的压缩上限是多少，是否能实现和文本的时序对齐)
- Streaming Ability and Efficiency
- Disentanglement in Acoustic Tokens
- Variable Frame Rate Tokens
- Combining Acoustic and Semantic Tokens (Acoustic能否突破蒸馏上限/ (对比模型层次) 在Tokenizer的层次统一)
- Paralinguistics in Semantic Tokens
- Noise Preservation vs. Noise Robustness
- Timbre Control in Token Vocoders
- Adaptivity

↑
↓
show-o





《TokenFlow: Unified Image Tokenizer for Multimodal Understanding and Generation》

$$d_{sem,i} = \|\hat{z}_{sem} - z_{sem,i}\|_2^2, \text{ for } i = 1, \dots, K \quad (1)$$

$$d_{pix,i} = \|\hat{z}_{pix} - z_{pix,i}\|_2^2, \text{ for } i = 1, \dots, K \quad (2)$$

$$i^* = \arg \min_i (d_{sem,i} + w_{dis} \cdot d_{pix,i}) \quad (3)$$

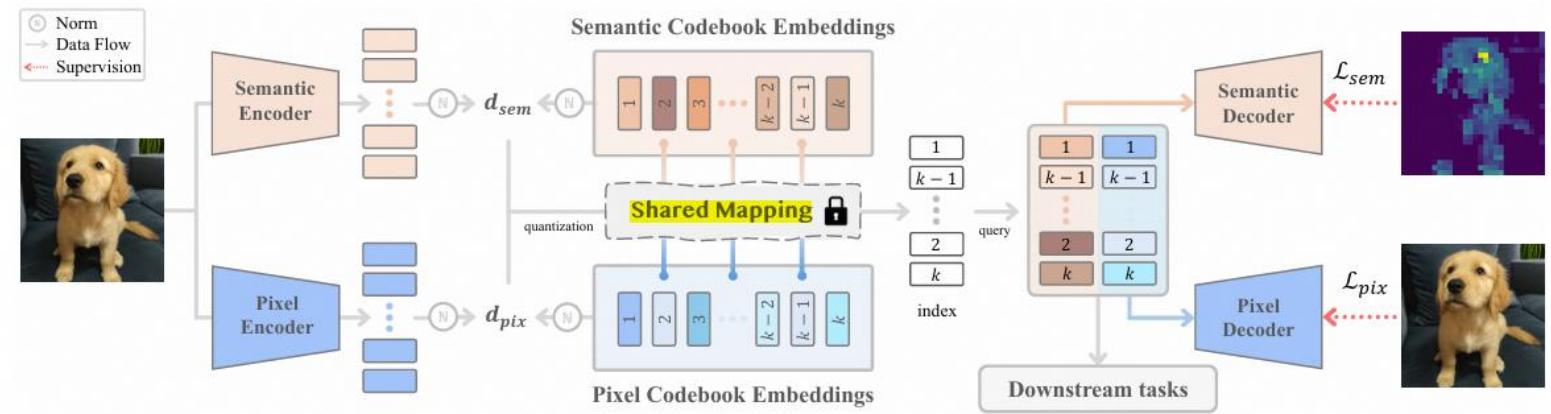


Figure 3. Overview of TokenFlow. We incorporate dual encoders and codebooks with a shared mapping, enabling the joint optimization of high-level semantics and low-level pixel details. For a given input image, distances d_{sem} and d_{pix} are calculated from the pixel-level and semantic-level codebooks, respectively, with the final codebook index and features determined by minimizing the weighted sum $d_{sem} + w_{dis} \cdot d_{pix}$. The resulting quantized features are independently decoded for both semantic alignment and image reconstruction training, and then concatenated to provide a unified representation for downstream tasks in understanding and generation.



Token 讨论2

《UniTok: A Unified Tokenizer for Visual Generation and Understanding》

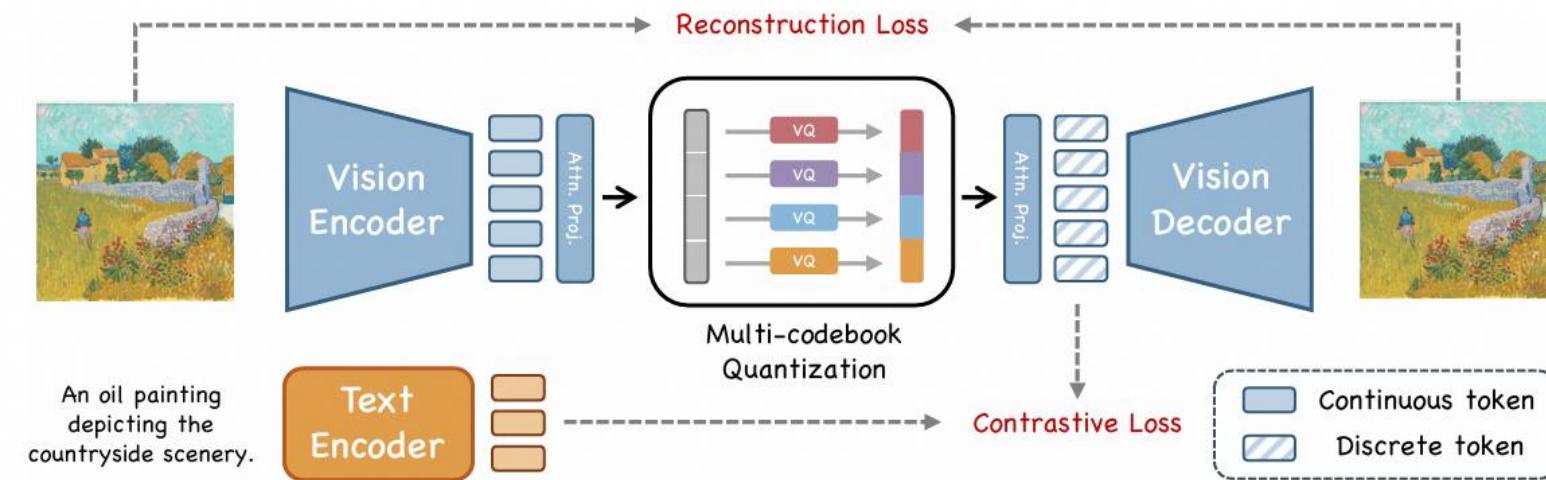
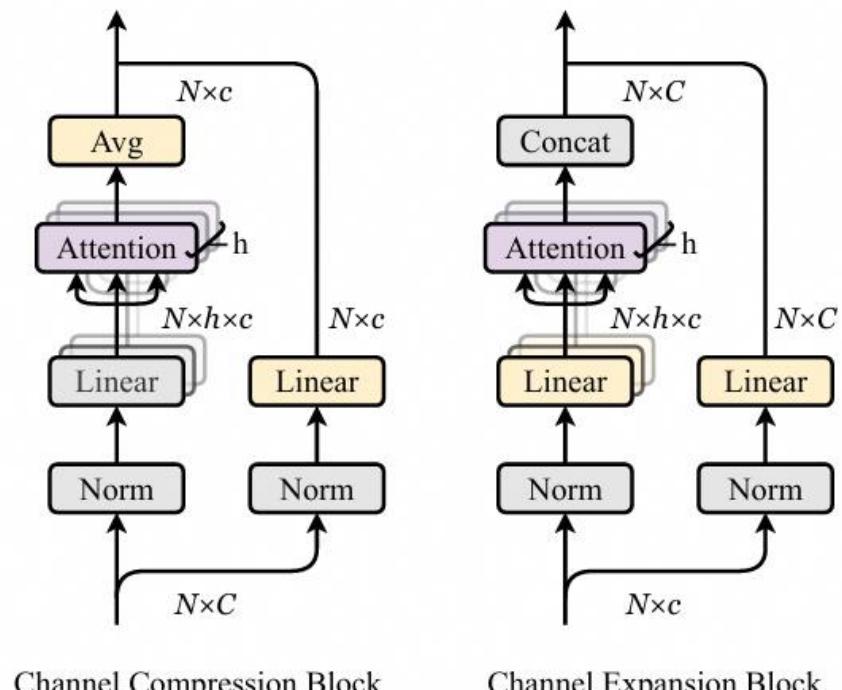


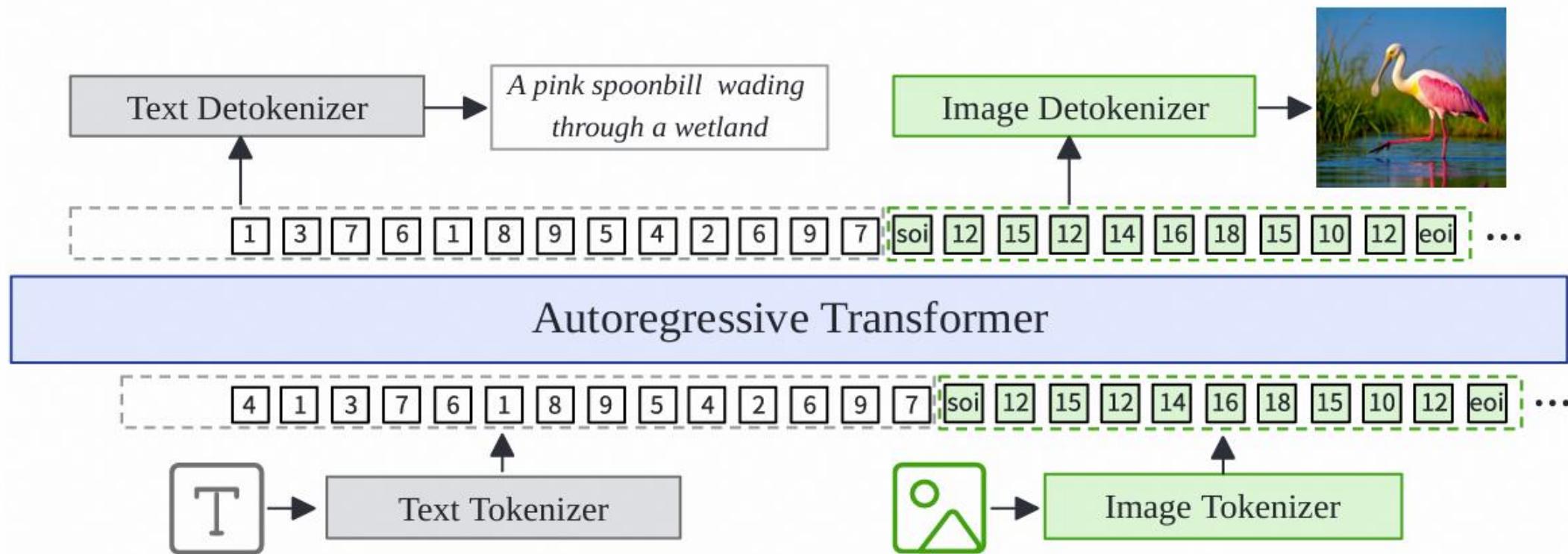
Figure 2. An overview of UniTok. The tokenizer is trained to faithfully reconstruct the input image while aligning its discrete latent features with the text caption. For vector quantization, each visual token is split into multiple chunks, which then undergo code index lookup on corresponding sub-codebooks concurrently.





Token 讨论2

《MUSE-VL:Modeling Unified VLM through Semantic Discrete Encoding》





Token 讨论2

《MUSE-VL:Modeling Unified VLM through Semantic Discrete Encoding》

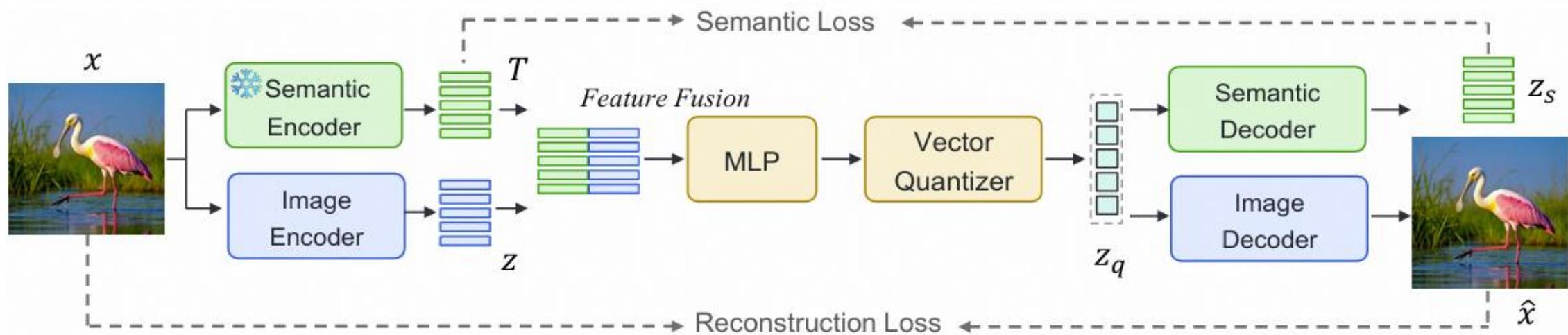


Figure 3. The overview of Semantic Discrete Encoding. The image is encoded and quantized into semantic discrete tokens, which are then separately reconstructed by the semantic decoder and the image decoder into semantic features and the original image.



Token 讨论2

《MUSE-VL:Modeling Unified VLM through Semantic Discrete Encoding》

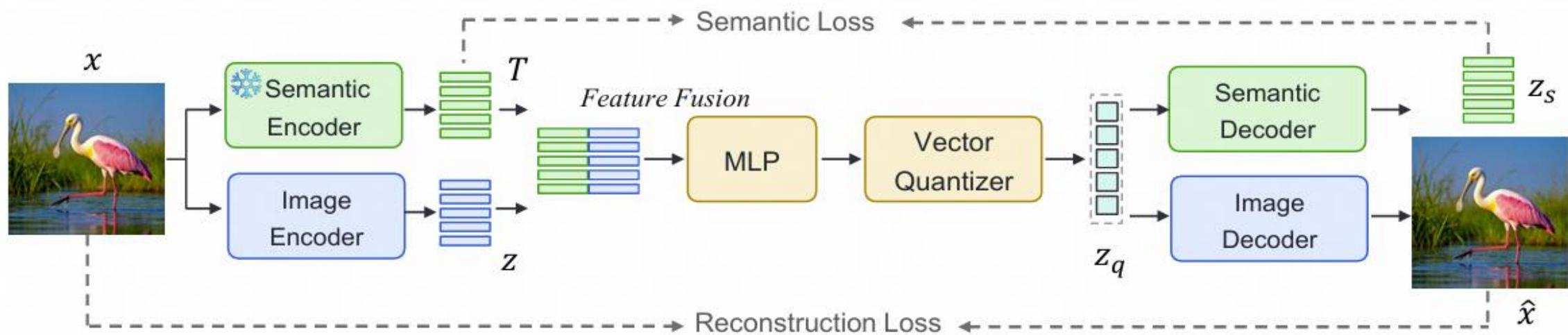


Figure 3. The overview of Semantic Discrete Encoding. The image is encoded and quantized into semantic discrete tokens, which are then separately reconstructed by the semantic decoder and the image decoder into semantic features and the original image.



Table 1: Comparison of Spoken Dialogue Datasets. **E** indicates whether the dataset emphasizes emotional information, **A** indicates the presence of audio events in the dialogue, and **M** indicates the involvement of music. The dialogue data is derived from three scenarios: controlled environments (**Env**), in-the-wild collection (**Wild**), and AI generation (**AI-Gen**). **#Avg.** represents the average number of turns per dialogue. [†]All responses in E-chat200 are in text format, duration only includes speech on the query side. The dialogues in AF-Dialogue are all text-based, with duration reflecting only audio and music segments.

Datasets	Scens.			Source	# Turns	#Dialog.	#Avg.	#Dur.
	E	A	M					
<i>Speech-to-Speech Dialogue Dataset</i>								
IEMOCAP (Busso et al., 2008)	✓	X	X	Env	10,039	151	66.48	12
SwitchBoard (Godfrey et al., 1992)	X	X	X	Wild	-	2,500	-	250
Fisher (Cieri et al., 2004)	X	X	X	Wild	-	11,699	-	1,960
DSTC2 (Henderson et al., 2014)	X	X	X	Wild	23,354	1,612	14.49	32
MELD (Poria et al., 2018)	✓	X	X	Wild	13,000	1,433	9.07	14
Expresso (Nguyen et al., 2023)	✓	X	X	Env	2,400	391	6.14	47
DailyTalk (Lee et al., 2023)	✓	X	X	Env	23,774	2,514	9.46	22
SpokenWOZ (Si et al., 2024)	X	X	X	Env	203,074	5,700	35.63	249
StyleTalk (Lin et al., 2024)	✓	X	X	AI-Gen	12,056	2,967	4.06	12
ShareChatX (ours)								
- ShareChat-Emotion	✓	X	X	AI-Gen	588,174	80,152	7.34	672
- ShareChat-Audio	✓	✓	X	AI-Gen	199,034	27,005	7.37	217
- ShareChat-Music	✓	X	✓	AI-Gen	160,028	21,443	7.46	242
- Overall	✓	✓	✓	AI-Gen	947,236	128,600	7.37	1,130
<i>Non-Speech-to-Speech Dialogue Dataset</i>								
E-chat200 (Xue et al., 2023)	✓	X	X	AI-Gen	356,000	178,000	2.00	193 [†]
AF-Dialogue (Kong et al., 2024)	X	✓	✓	AI-Gen	657,600	82,200	8.00	228 [†]

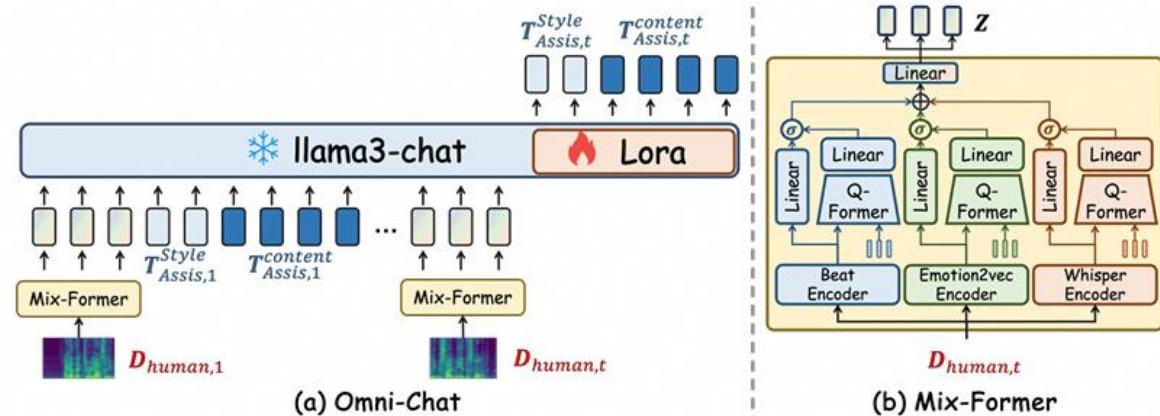
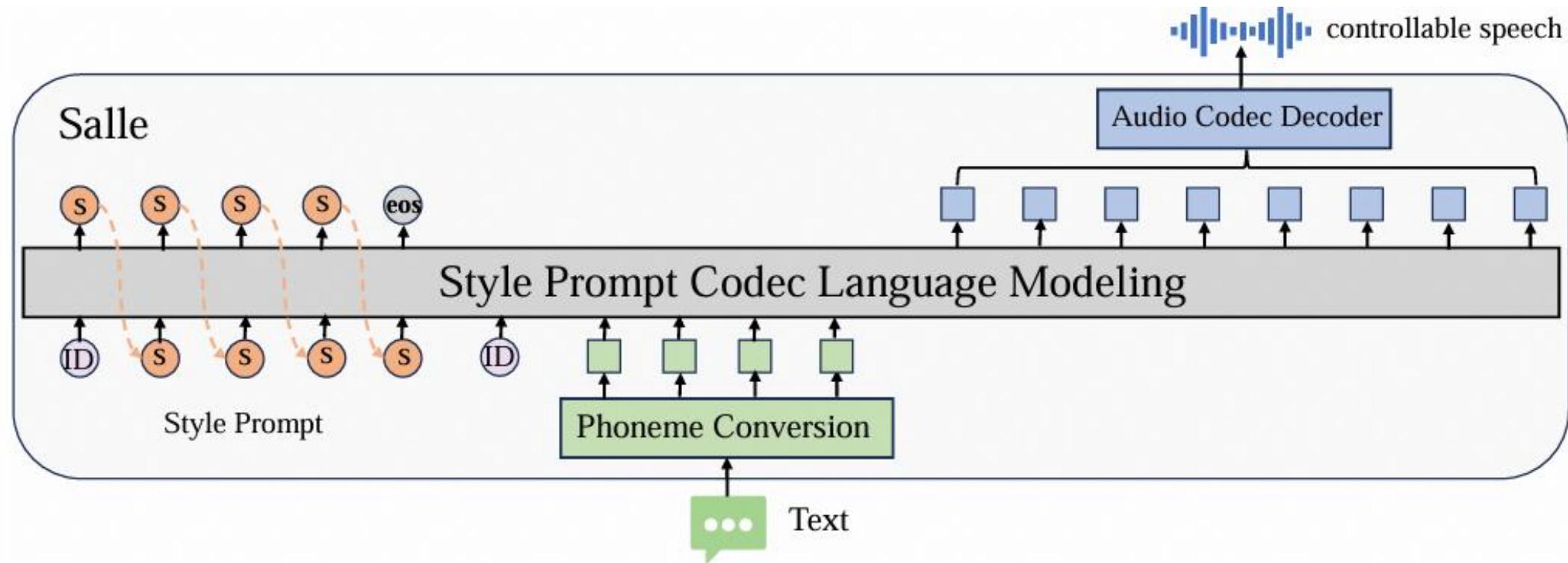


Table 2: Performance Comparison of Various Spoken Dialogue Systems on the DailyTalk Dataset. The content metrics include **@B** (BLEU), **@R** (ROUGE-L), **@M** (METEOR), and **@BS** (BERTScore). The Style metrics include **@F1_e** for emotion prediction accuracy.

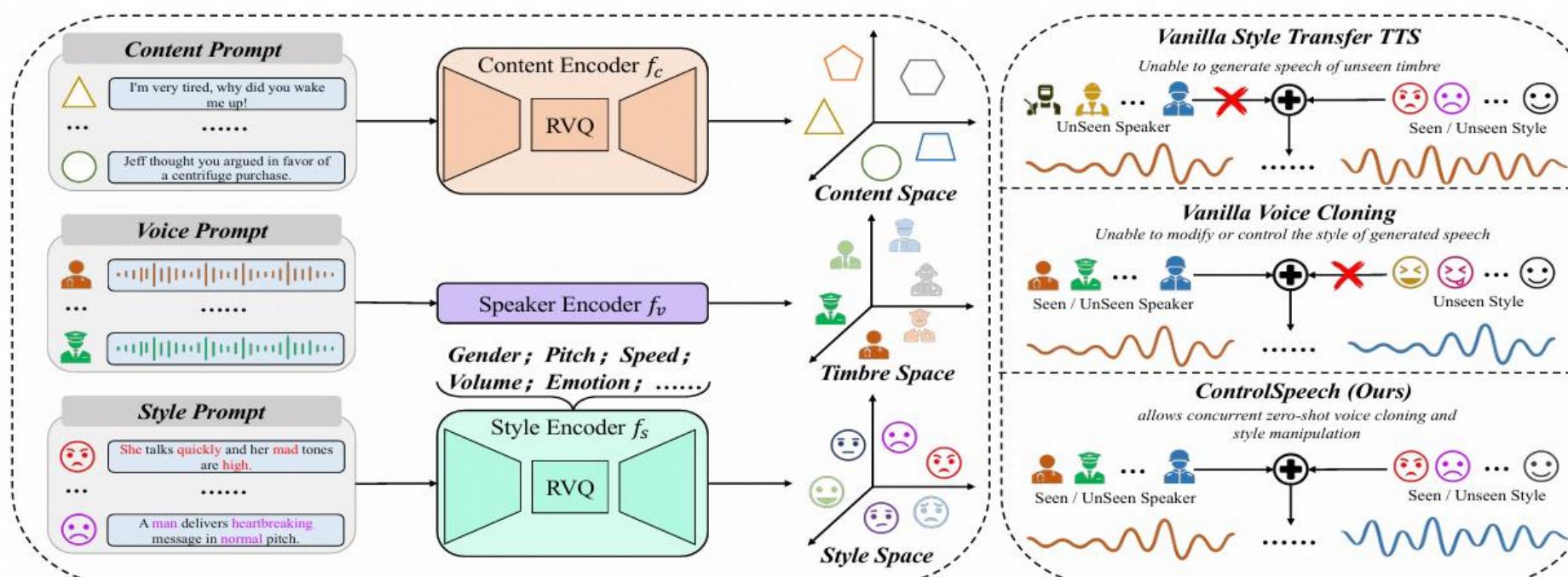
Methods	@B	@R	@M	@BS	F1 _e	GPT-eval	MOS
<i>ASR-Based Spoken Dialogue System</i>							
StyleTalk (Lin et al., 2024)	2.01	9.42	10.95	82.82	49.63	3.51	3.42±0.23
FunAudioLLM (SpeechTeam, 2024)	2.65	12.53	11.82	84.76	61.02	3.82	3.85±0.18
<i>Direct Spoken Dialogue System</i>							
Audio-Flamingo (Kong et al., 2024)	1.47	5.01	10.23	83.94	-	2.35	2.53±0.25
SpeechGPT (Zhang et al., 2023)	1.42	7.85	9.42	84.11	-	2.68	2.45±0.32
Qwen-Audio (Chu et al., 2023)	2.04	7.43	11.21	84.33	-	3.01	3.23±0.18
Salmonn (Tang et al., 2023)	2.32	11.78	11.56	85.47	-	3.41	3.05±0.22
Qwen2-Audio (Chu et al., 2024)	3.03	12.81	13.89	86.14	-	4.01	3.87±0.25
OmniChat (ours)	3.54	12.63	12.57	86.24	71.87	3.96	3.97±0.22
OmniChat + Real Data (ours)	4.95	12.95	14.24	86.99	75.46	4.15	3.99±0.18





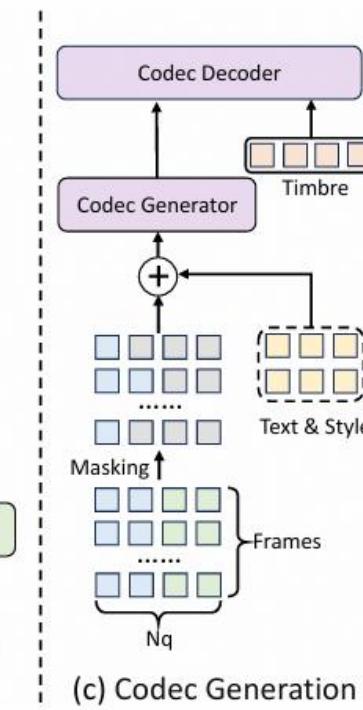
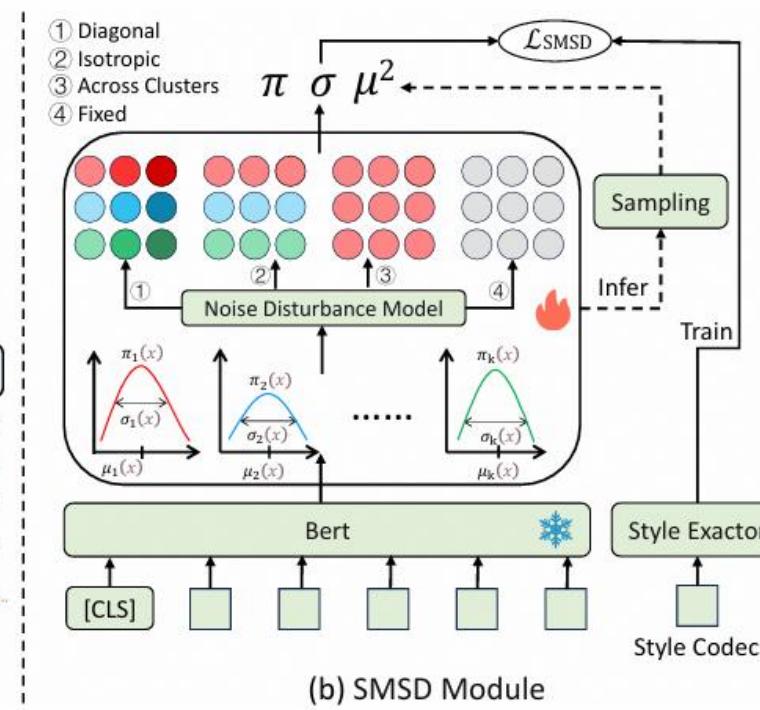
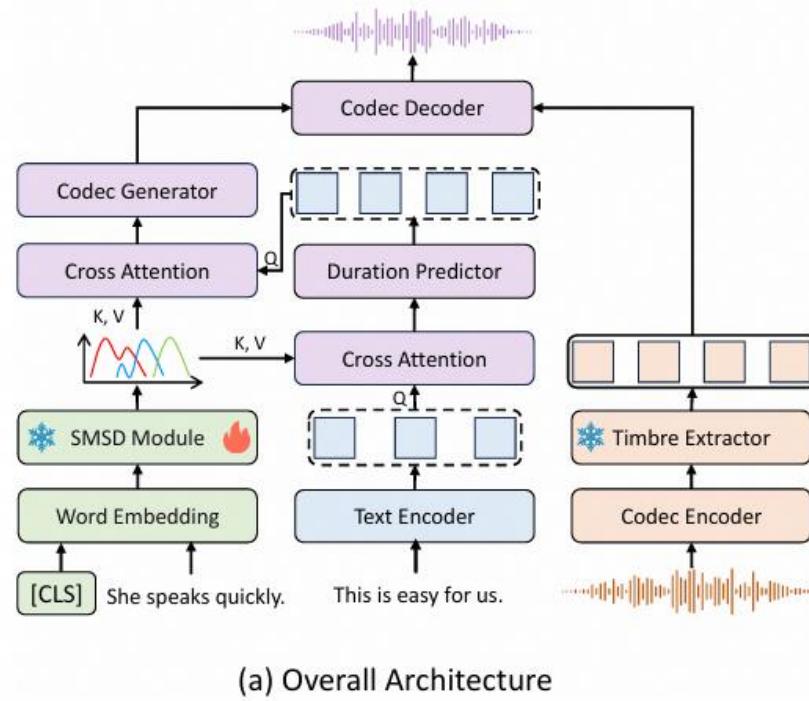
1. 语音克隆模型例如VALL-E只能克隆说话人的声音，没办法做到任意的风格调整
2. 文本风格控制模型，虽然能控制说话人的风格，但是不能克隆说话人的音色（用说话人ID）

例如“请你用特朗普的声音，小孩的语调悲伤地说，今天是星期一”，这个需求涉及到语音合成模型需要同时，独立zero-shot 控制音色，内容，风格。





ControlSpeech的架构设计难点:两段Prompt会冲突/zero-shot能力与训练数据不匹配的问题





ControlSpeech的风格映射多对多问题: 一段文本风格描述会对应多段语音, 多段文本风格描述会对应一段音频

Specifically, one raw style prompt $X_s = [X_1, X_2, X_3, \dots, X_L]$ is prepended with a $[CLS]$ token, then converted into word embedding, and fed into the BERT model, where L denotes the length of the style prompt. The hidden vector corresponding to the $[CLS]$ token is regarded as the global style semantic representation X_s' , which guides generation and sampling of subsequent modules.

Based on the MDN network (Zen & Senior, 2014; Duan, 2019; Du & Yu, 2021), we aim to regress the target style representation $Y_s' \in \mathbb{R}^d$, using the style semantic input representation $X_s' \in \mathbb{R}^n$ as covariates, where d and n are the respective dimensions. We model the conditional distribution as a mixture of Gaussian distribution, as follows:

$$P_\theta(Y_s' | X_s') = \sum_{k=1}^K \pi_k \mathcal{N}(\mu^{(k)}, \sigma^{2(k)}) \quad (5)$$

where K is a hyperparameter as the number of independent Gaussian distribution, and other mixture distribution parameters $\pi_k, \mu^{(k)}, \sigma^{2(k)}$ are output of a neural MDN network f_θ based on the input style semantic representation X_s' , as follows:

$$\pi \in \Delta^{K-1}, \mu^{(k)} \in \mathbb{R}^d, \sigma^{2(k)} \in S_+^d = f_\theta(X_s') \quad (6)$$

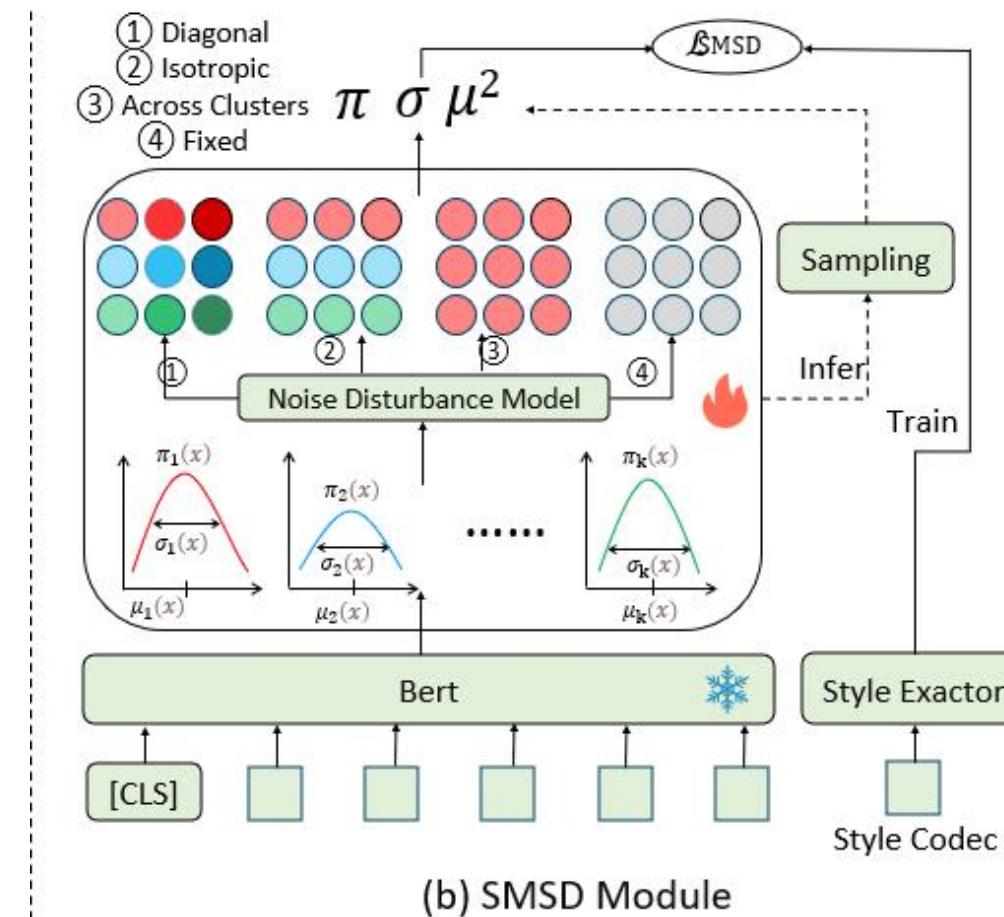
Note that the sum of the mixture weights is constrained to 1 during the training phase, which is achieved by applying a softmax function on the corresponding neural network output α_k , as follows:

$$\pi_k = \frac{\exp(a_k)}{\sum_{k=1}^K \exp(a_k)} \quad (7)$$

To further enhance the diversity of style control, we design a specialized noise perturbation module within the SMSD module to constrain the noise model. As illustrated by the circles within the SMSD module in Figure 2 (b), this noise perturbation module regulates the isotropy of perturbations ε across different dimensions in variance $\sigma^{2(k)}$. The four types of perturbations from left to right in Figure 2 (b) are as follows:

- Fully factored:** $\sigma^{2(k)} = f_\theta(X_s') + f_\theta(\varepsilon) = \text{diag}(\sigma^{2(k)}) \in \mathbb{R}_+^d$, which predicts the noise level for each dimension separately.
- Isotropic:** $\sigma^{2(k)} = f_\theta(X_s') + f_\theta(\varepsilon) = \sigma^{2(k)} I \in \mathbb{R}_+$, which assumes the same noise level for each dimension over d .
- Isotropic across clusters:** $\sigma^{2(k)} = f_\theta(X_s') + f_\theta(\varepsilon) = \sigma^2 I \in \mathbb{R}_+$, which assumes the same noise level for each dimension over d and cluster.
- Fixed isotropic** is the same as Isotropic across clusters but does not learn σ^2 .

As shown in the experimental results in Appendix I, *isotropic across clusters* outperforms the other types for striking a balance between accuracy and diversity and is used as the mode for noise perturbation. We obtain more robust mean, variance, and weight parameters for the mixture of Gaussian distributions with the noise perturbation module. The training objective of the SMSD module is the negative log-likelihood of the observation Y_s' given its input X_s . The loss function is formulated as follows. Details for deriving the non-convex \mathcal{L}_{SMSD} are in Appendix C.





理解语音中的副语言信息

- 一个老人和一个小孩问我一天应该喝多少水
- 晴天和雨天下问应不应该晒衣服
- 不同的情感说我今天把手机丢了

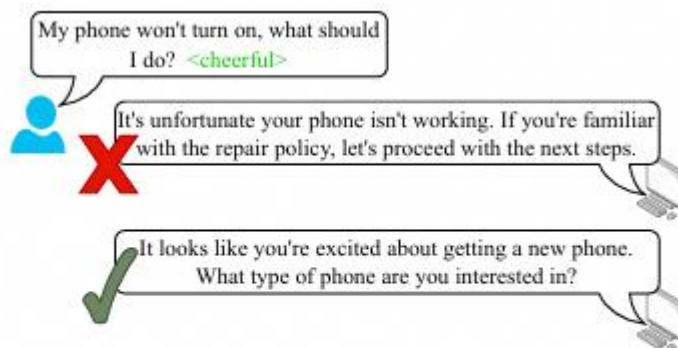


Figure 1: Emotion-sensitive spoken dialogue scenario. <> denotes the emotion of the speech.

Table 2: Dialogue example of VoxDialog. The bold black text highlights the acoustic and background information unique to spoken dialogue, while the blue text emphasizes responses that depend on these information.

Attributes Example	
I. Speaker Information	
Gender	A: Excuse me, I'm looking for a reliable smartphone that suits my needs. (Man Voice) B: Certainly, Sir! ... excellent performance and a long-lasting battery, perfect for business tasks .
Age	A: What dishes do you recommend? (Young Voice) B: Sure! You might enjoy our specialty beef burger , it's fantastic with fresh juice .
Accent	A: Could you help me with my schedule? (British Accent) B: Of course, I'll assist you in organizing your timetable .
Language	A: 今天天气怎么样? (Speaking in Chinese) B: 今天天气晴朗, 气温适中, 适合外出活动。
II. Paralinguistic Information	
Emotion	A: Can you recommend a good movie? (Sad Tone) B: I get it. Maybe Soul—it's emotional and might bring you some comfort.
Volume	A: Can you tell me how to solve this problem? (Loud Volume) B: Could you speak more quietly? People around are studying.
Speed	A: I'd like to get the burger, a soda, and also the pasta, please. (Speaking Quickly) B: Sorry, could you say that more slowly? I didn't catch everything you said.
Fidelity	A: Can you tell me the time of the meeting? (Poor Audio Quality) B: I'm having trouble hearing you. Could you improve the audio quality?
Stress	A: I really don't like making sushi. (Emphasis on "making sushi") B: Oh, I see. If you don't like making sushi, what other types of cooking do you enjoy?
Non-verbal Expressions	A: Could you help me move this box? (With a sigh before speaking) B: Are you feeling okay? It seems like you're really tired . I can take care of it for you.
III. Environmental Information	
Audio Events	What was that sound just now? (Background sound: airplane engine sound, explosion sound) That was a loud explosion. It sounded like the plane exploded. Hope no one was hurt.
Music	A: Hey, what instrument is this song played on? (Music: Piano Song, Sad Song) B: It should be the piano, it sounds so sad.



测试数据集制作过程和数据集具体类别

- Stage1: Dialogue Script Synthesis.
- Stage2: Spoken Dialogue Generation.
- Stage3: Automatic Verification for Spoken Dialogue.
- Stage4: Post-processing for Specific Acoustic Attributes.
- Stage5: Human Verification.

Table 3: Detailed statistics of the corresponding subsets of each attribute in VoxDialogue. Gray fonts indicate that samples of this attribute are included in other subsets. IN (India), CA (Canada), ZA (South Africa), GB (United Kingdom), SG (Singapore), US (United States), and AU (Australia). **Turns** represents the total number of turns in each subset, **Dialog.** indicates the number of dialogues in each subset, **Avg** denotes the average number of turns per dialogue in each subset, and **Dur.** refers to the total duration (in hours) of all dialogues in each subset.

			Turns	Dialog.	Avg	Dur.
I. Speaker Information						
Gender	Male, Female		2040	340	6.0	3.17
Age	Youth (15-30), Middle-Aged (30-60), Elderly (60+)		3096	447	6.9	6.05
Accent	IN, CA, ZA, GB, SG, US, AU		1440	240	6.0	2.20
Language	Chinese, English		2892	482	6.0	3.51
II. Paralinguistic Information						
Emotion	Neutral, Happy, Sad, Angry, Surprised, Fearful, Disgusted		1980	330	6.0	2.41
Volume	Loud Volume, Low Volume, Normal Volume		1824	304	6.0	2.08
Speed	High Speed, Low Speed, Normal Speed		2184	364	6.0	2.93
Fidelity	Low Fidelity, Normal Fidelity		2196	366	6.0	3.36
Stress	Stress, No Stress		2354	392	6.0	2.51
NVE	Laughter, No Laughter		2046	341	6.0	3.68
III. Environmental Information						
Audio	The caption of different audio. (e.g., The wind is blowing and rustling occurs.)		5000	500	10.0	5.25
Music	The aspect list of different music pieces. (e.g., [steel drum, higher register, amateur recording])		3734	420	8.9	5.42
Overall			30.7K	4.5K	6.8	42.56



如何评测各个spoken dialogue models的能力？

Table 4: GPT-based Metric Comparison of Different Spoken Dialogue Models on VoxDialogue.

Method	Speaker Info			Paralinguistic Info				Env Info				
	Age	Gen	Acc	Lan	Emo	Vol	Spd	Fid	Str	NVE	Aud	Mus
<i>ASR-Based Spoken Dialogue System</i>												
FunAudioLLM (SpeechTeam, 2024)	4.32	4.39	3.57	4.61	4.09	1.82	1.92	1.79	3.13	2.87	3.47	3.59
<i>Direct Spoken Dialogue System</i>												
Audio-Flamingo (Kong et al., 2024)	1.00	1.00	1.04	1.72	1.00	1.20	1.14	1.26	1.34	1.06	1.37	1.11
SALMONN (Tang et al., 2023)	1.99	1.64	1.78	3.50	1.84	2.88	2.27	2.29	3.86	2.59	2.15	2.23
Qwen-Audio (Chu et al., 2023)	1.36	1.04	1.28	1.04	1.06	1.48	1.08	1.32	2.49	2.65	1.42	1.18
Qwen2-Audio (Chu et al., 2024)	3.46	4.18	2.71	4.43	3.73	3.06	3.29	2.98	3.93	3.46	3.81	3.98

relevant and natural, mentioning the attribute, but could be improved. **5**: Contextually relevant, smooth, natural, and accurately addresses the attribute. We have included all the evaluated prompt templates in supplementary materials. Please refer to the supplementary materials for more details.

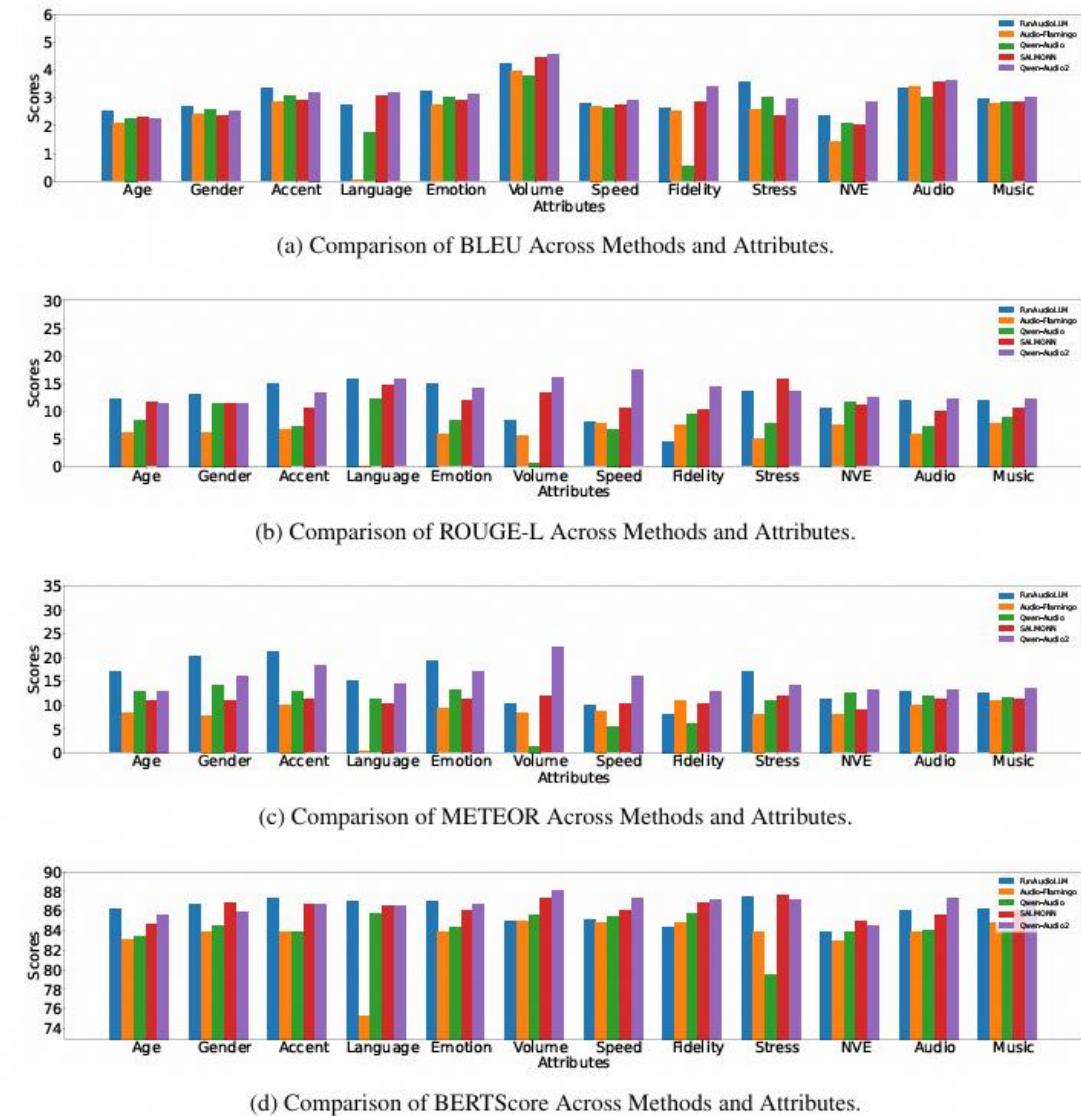


Figure 2: The comparison of spoken dialogue performance across 12 different attribute-specific test sets on the VoxDialogue dataset.



RAG技术的定义: Figure1

1. 最新的电影 (Latest Movies)

- 用户问题: 美国队长4什么时候上映?
- 内部推理: 《美国队长4》的上映日期尚未公布。 ✗
- 知识库检索: 导演关于《美国队长4》的声明:很高兴介绍《美国队长4》，它将于 **2025年2月14日** 上映.....
- 检索增强: “《美国队长4: 新世界秩序》将在 **2025年2月14日** 于北美和中国大陆上映。这是第一部由安东尼·麦凯 (Anthony Mackie) 饰演美国队长的电影。” ✓

2. 当前新闻 (Current News)

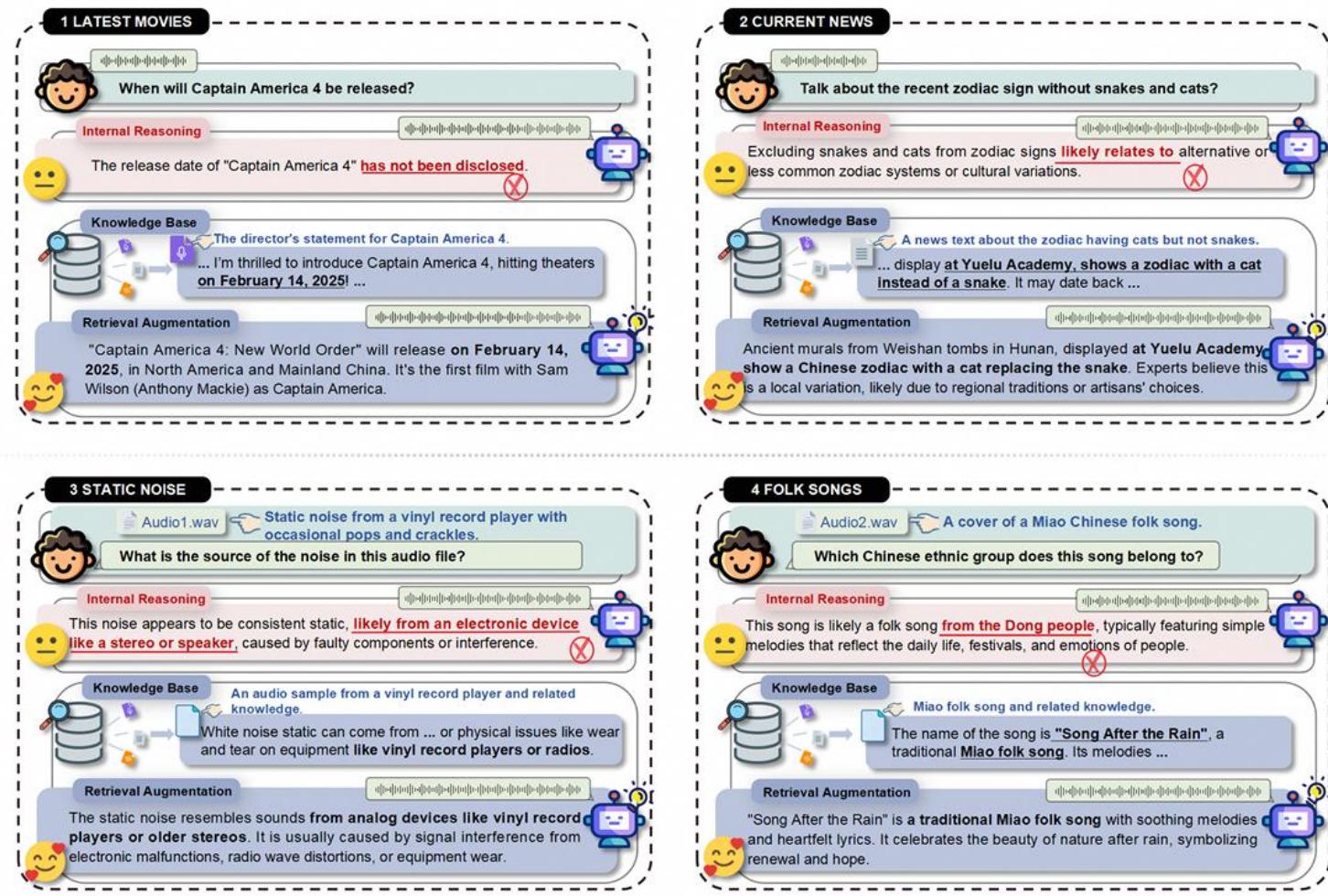
- 用户问题: 谈谈最近没有蛇和猫的十二生肖?
- 内部推理: 排除蛇和猫的生肖可能与不常见的生肖系统或文化变种有关。 ✗
- 知识库检索: 一篇关于生肖的新闻报道提到, 岳麓书院展示了一种用猫代替蛇的中国生肖.....
- 检索增强: 湖南省滩山古墓的壁画, 在岳麓书院展出, 显示了一种**用猫代替蛇的中国生肖**。专家认为这是一个地方变种, 可能与地区传统或工匠的选择有关。 ✓

3. 静态噪声 (Static Noise)

- 用户问题: 这段音频中的噪声来源是什么?
- 内部推理: 这段噪声似乎是持续性的静态噪声, 可能来自电子设备, 如立体声喇叭或扬声器, 由故障组件或干扰引起。 ✗
- 知识库检索: 一段来自黑胶唱片播放器的音频样本及相关知识: 白噪声静电可能来源于.....或物理问题, 如黑胶唱片播放器或收音机的磨损。
- 检索增强: *这种静态噪声类似于**模拟设备 (如黑胶唱片播放器或老式立体声设备) *发出的噪声。通常由信号干扰 (例如电子设备故障、无线电波干扰或设备磨损) 引起。 ✓

4. 民族民歌 (Folk Songs)

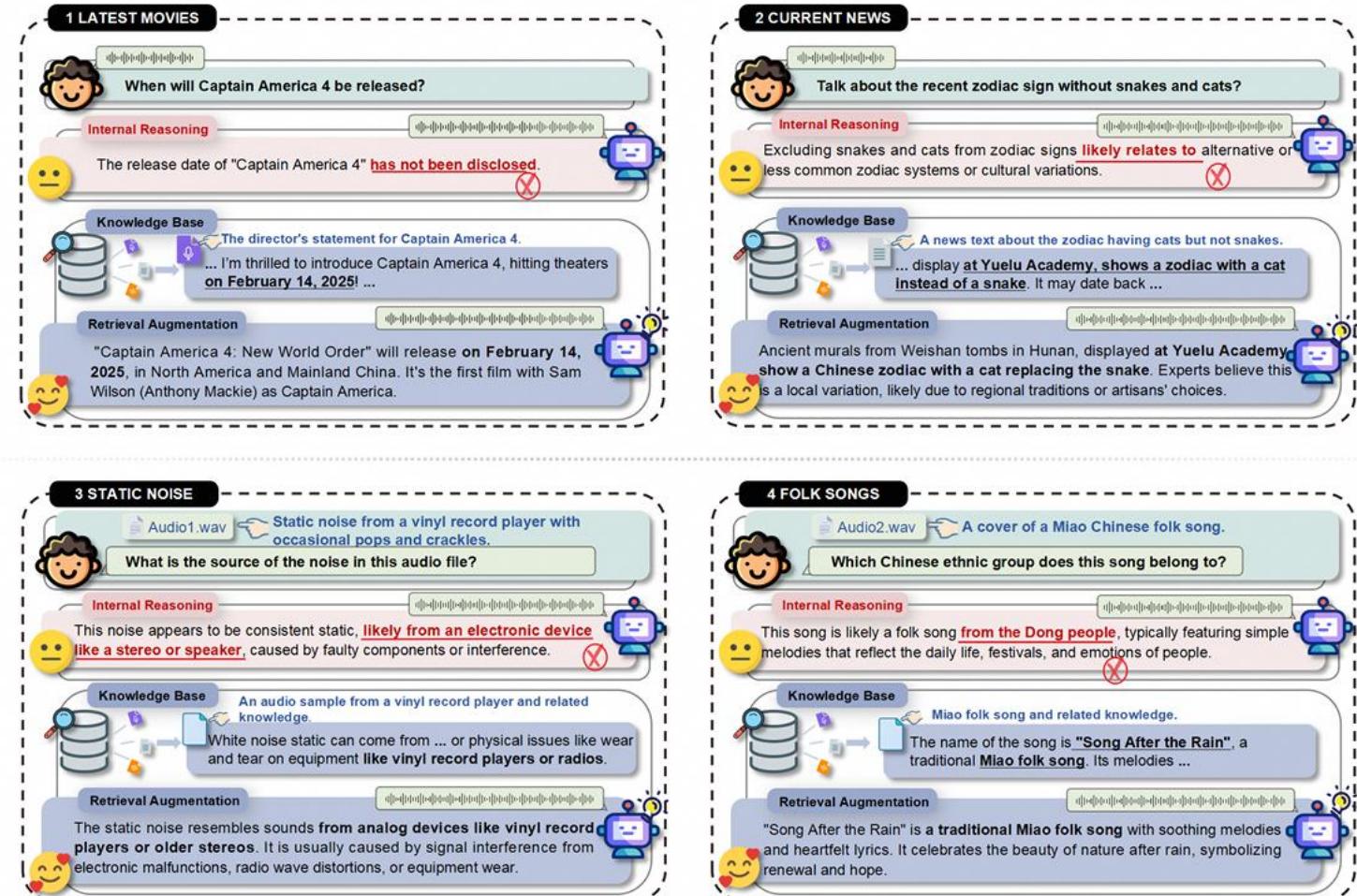
- 用户问题: 这首歌属于哪个中国少数民族?
- 内部推理: 这首歌可能是一首**侗族 (Dong)** 民歌, 通常具有简单旋律, 反映了人们的日常生活、节日和情感。 ✗
- 知识库检索: 苗族 (Miao) 民歌及相关知识: 这首歌的名字是《雨后之歌》 (Song After the Rain), 是一首传统的苗族民歌.....
- 检索增强: 《雨后之歌》是一首**传统苗族民歌**, 旋律优美, 歌词富有情感, 歌颂雨后大自然的美丽, 象征着重生与希望。 ✓





RAG技术在语音领域的应用: Figure2

- 级联式固有的问题
- 没有真正意义上和speech模态深入融合





WavRAG的核心pipeline

- 端到端的接入语音模态
- 知识库中包含speech, music, audio等

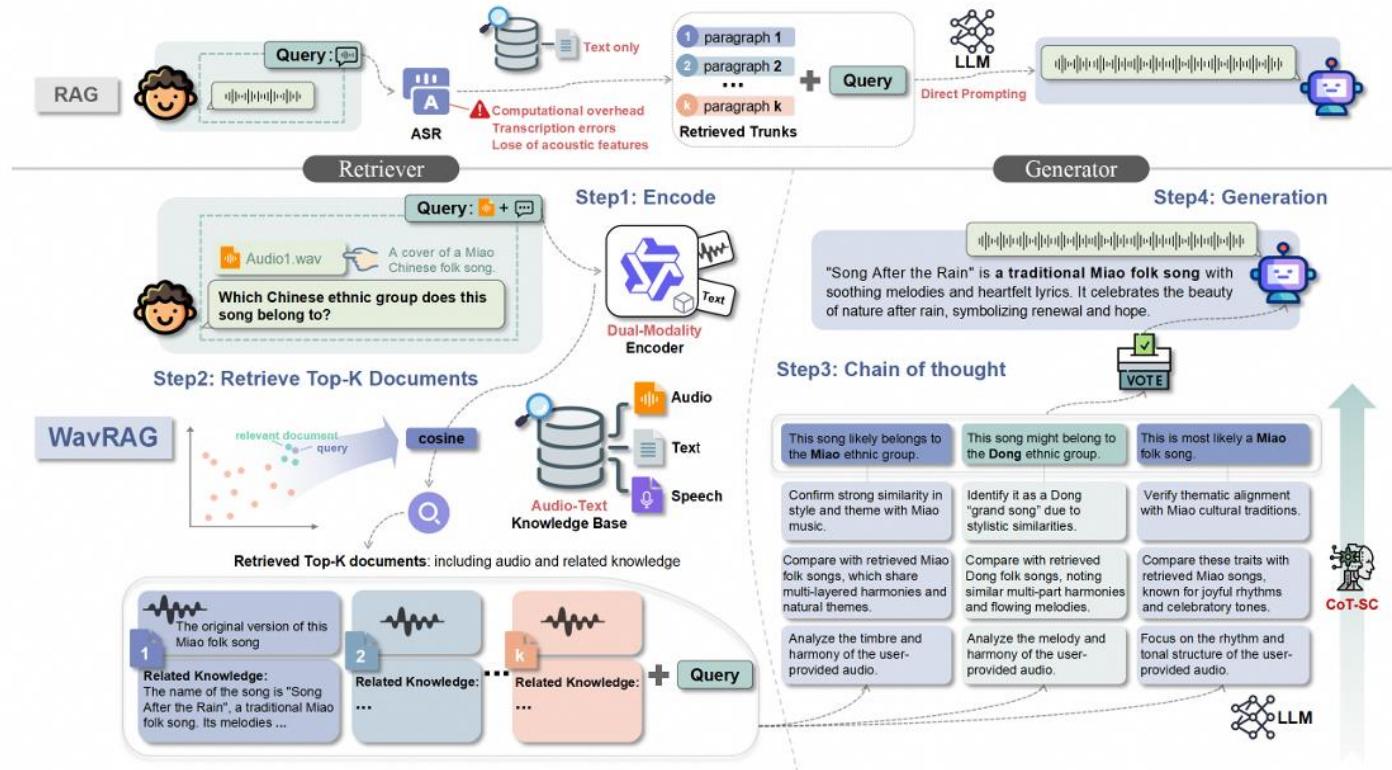


Figure 2: Architecture of the WavRAG framework. Top: Traditional RAG pipeline using ASR, highlighting its limitations. Bottom: WavRAG's four-step process: (1) A dual-modality encoder creates embeddings for both audio and text queries; (2) Top-K documents are retrieved from an audio-text knowledge base using cosine similarity; (3) A chain-of-thought reasoning process analyzes the retrieved information; (4) A large language model generates the final response, grounded in the retrieved knowledge.



如何统一编码text-speech-audio-music? 让大模型变成Encoder

受到《LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders》2024COLM启发

simply fine-tuning Qwen2-Audio on the downstream task is insufficient for optimal retrieval performance. While pre-trained MLLMs like Qwen2-Audio possess robust multimodal understanding, their pre-training objectives are not directly optimized for creating embeddings suitable for similarity-based retrieval.

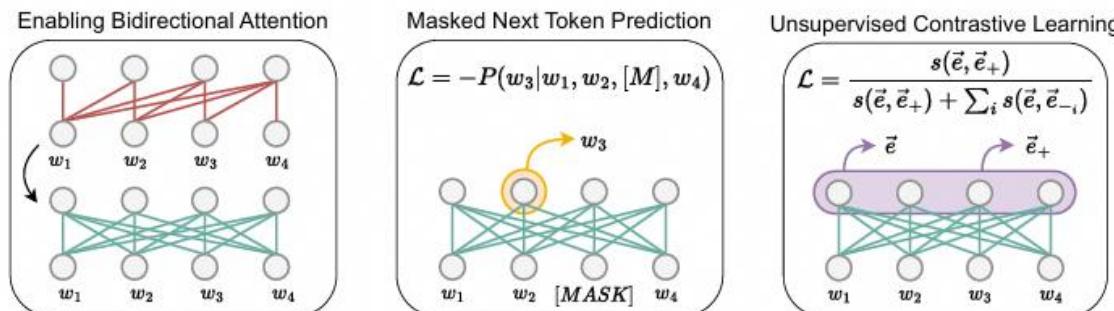
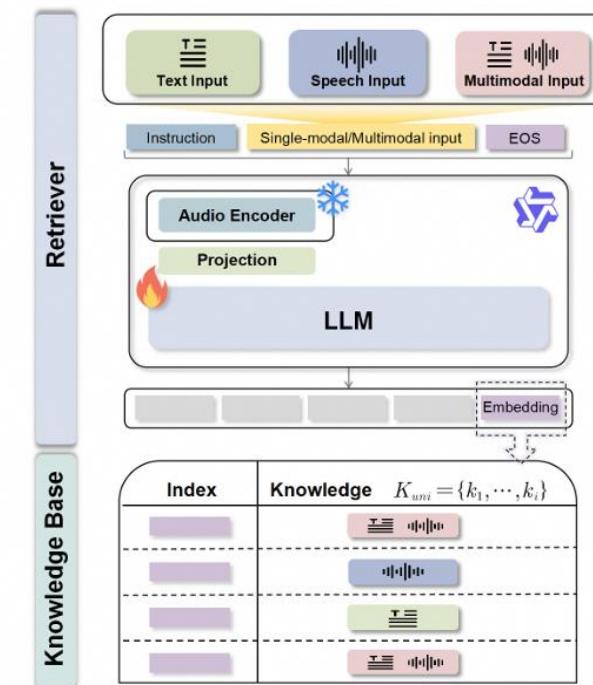


Figure 1: The 3 steps of LLM2Vec. First, we enable bidirectional attention to overcome the restrictions of causal attention (**Bi**). Second, we adapt the model to use bidirectional attention by masked next token prediction training (**MNTP**). Third, we apply unsupervised contrastive learning with mean pooling to learn better sequence representations (**SimCSE**).





检索和生成的实验结果

Task	Dataset	Model	Whisper size	WER	Avg. Time	Metric			
						R@1	R@5	R@10	nDCG@10
Speech2Text	HotpotQA	BGE	Tiny	37.55%	1.26	0.3741	0.7024	0.7509	0.4628
			Medium	21.67%	1.48	0.4440	0.8319	0.8736	0.5190
			Large	19.2%	1.92	0.4533	0.8519	0.8895	0.5252
	Comparison vs. BGE (Tiny / Medium / Large):	Ours	-	-	0.23	0.4532	0.8492	0.8898	0.5117
		Speed-up $\approx 5.49 \times / 6.43 \times / 8.35 \times$, $\Delta R@10 \approx +0.139 / +0.016 / +0.0003$.							
		CLS-R	-	-	-	0.4982	0.7963	0.8583	-
Text2Speech	Spoken-SQuAD	BGE	CLS-R	-	44.22%	-	0.5464	0.7767	0.8497
			Ours	-	-	0.11	0.6844	0.8374	0.9023
			CLS-R	-	16.69%	-	0.3065	0.6219	0.7443
	Comparison vs. BGE (Tiny / Medium / Large):	BGE	Tiny	45.34% / 53.66%	0.62 / 1.27	0.1696	0.3871	0.4828	0.2194
		SLUE-SQA-5	Medium	26.14% / 44.46%	0.87 / 3.44	0.3228	0.5940	0.6982	0.2989
			Large	23.59% / 42.19%	0.98 / 4.63	0.3312	0.6121	0.7196	0.3269
		Ours	-	-	0.17 / 0.22	0.3392	0.6308	0.7221	0.3623
Speech2Speech	CIAP(AT)	BGE	CIAP(AT)	-	-	0.05	0.1260	0.2940	0.3989
			CLAP(TA)	-	-	0.05	0.0998	0.2577	0.3588
			CLAP(AT2AT)	-	-	0.09	0.1345	0.2145	0.2379
	Audio+Text2Audio+Text	Ours	CIAP(ALL)	-	-	0.06	0.0001	0.0012	0.0018
			BGE(Caption)	-	-	1.99	0.0251	0.0585	0.0775
			Ours	-	-	0.19	0.2728	0.5184	0.6313

Table 1: Comparison of various models and configurations across multiple tasks including Speech2Text, Text2Speech, Speech2Speech, and Audio+Text2Audio+Text. Performance metrics include Word Error Rate (WER), Average Time, and multiple retrieval metrics (R@1, R@5, R@10, and nDCG@10). CLAP are evaluated for their respective tasks: CIAP (AT) is tested only on the Audio-to-Text retrieval task subset, CLAP (TA) on the Text-to-Audio retrieval task, CLAP (AT2AT) on Audio-to-Text-to-Audio tasks, and CLAP (ALL) on the entire dataset. Speed-up and performance changes relative to BGE (different Whisper model size Tiny/Medium/Large) configurations are reported.

Method	Model	Input	EM		Avg EM	FS (Ours)
			HotpotQA	SLUE-SQA-5		
(a) TextRAG						
GPT-4o	QwenAudio	top-1	0.3124	0.3237	0.3181	-
		top-2	0.3457	0.3359	0.3408	-
		top-3	0.3623	0.3531	0.3577	-
		Oracle	0.5853	0.5931	0.5892	-
(b) WavRAG						
GPT-4o	QwenAudio	top-1	0.4019	0.3904	0.3962	0.5732
		top-2	0.4186	0.4315	0.4249	0.6408
		top-3	0.4271	0.4007	0.4139	0.5129
		Oracle	0.5941	0.6164	0.6053	0.7096
(c) WavRAG-CoT						
GPT-4o	QwenAudio	top-1	0.4261	0.4520	0.4390	0.6412
		top-2	0.4286	0.5239	0.4983	0.6487
		top-3	0.4403	0.4918	0.4662	0.5981
		Oracle	0.5976	0.6849	0.6413	0.7389
QwenAudio	QwenAudio	top-1	0.2688	0.3132	0.2910	0.6386
		top-2	0.3026	0.3352	0.3189	0.6017
		top-3	0.3152	0.3397	0.3275	0.5612
		Oracle	0.5863	0.6103	0.5983	0.7122

Table 2: Generation experiment results: Performance comparison of TextRAG, WavRAG, and WavRAG-CoT on HotpotQA and SLUE-SQA-5 datasets. Metrics include Exact Match (EM) and F1-Score (FS). Results are shown for both GPT-4o and QwenAudio base models, with varying numbers of retrieved documents (top-1, top-2, top-3, and Oracle).



我们的目标：以spoken dialogue models为核心，借助各种上下游的技术，完善对话模型领域相关生态。通过多模态的深入融合，进一步涌现对话智能体的智商和情商。通过大模型中的Tokenizer、生成范式、推理、强化学习等等技术。通往真正的AGI。



季圣鹏



导师：赵洲

微信：18943038195



Thank you



季圣鹏



导师: 赵洲

微信: 18943038195