From Text to Talk: Audio-Language Model Needs Non-Autoregressive Joint Training

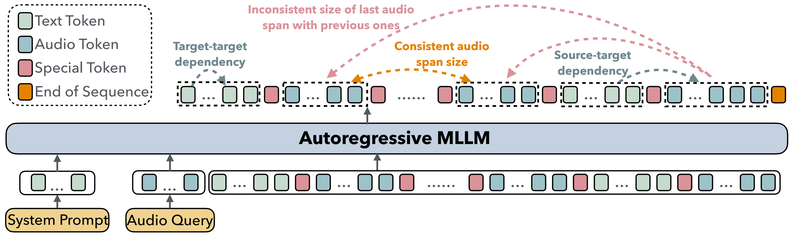
Tianqiao Liu[1,2], Xueyi Li[1], Hao Wang[3], Haoxuan Li[3], Zhichao Chen[3], Weiqi Luo[1], and Zitao Liu[1]

[1]Guangdong Institute of Smart Education, Jinan University, Guangzhou China
[2]TAL Education Group, Beijing China
[3] Peking University, Beijing China

Introduction

Current Problems In Audio-Language Multimodal Large Language Model (MLLM): Existing audio-language MLLMs generate interleaved text and audio tokens with a uniform autoregressive (AR) objective, overlooking a fundamental.

- **Dependency Mismatch:** Text needs causal target-to-target modeling, while audio is mainly conditioned on the source text; within-span audio order is flexible.
- **Error Propagation:** Left-to-right audio AR suffers exposure bias: early token errors cascade through the span.



Preliminary

Interleaved Sequence Notation: We consider interleaved text-audio sequences with vocabulary $\mathbf{V} = \mathbf{V}_{\text{text}} \cup \mathbf{V}_{\text{audio}} \cup \mathbf{S}$, where $\mathbf{S} = \langle\text{SOA}\rangle, \langle\text{EOA}\rangle, \langle\text{EOS}\rangle, [\mathbf{M}]$ contains control and mask tokens. Each sequence alternates between text and audio spans: $x = (\mathbf{T}_1, \mathbf{A}_1, \ldots, \mathbf{T}_M, \mathbf{A}_M, \langle\text{EOS}\rangle)$ where $\mathbf{T}_m = (t_{m,1}, \ldots, t_{m,|\mathbf{T}_m|})$ and $\mathbf{A}_m = (a_{m,1}, \ldots, a_{m,|\mathbf{A}_m|})$. A natural question is how to model this sequence.

AR Modeling: AR models factorize the joint probability via the chain rule $p(x) = \prod_{i=1}^{L} p(x^i \mid x^{<i})$. Applied to our interleaved setting:

$$\mathcal{L}_{\text{AR}}(x) = -\sum_{m=1}^{M} \sum_{j=1}^{|\mathbf{T}_m|} \log\, p_\theta\big(t_{m,j} \mid \mathbf{T}_{<m}, \mathbf{A}_{<m}, t_{m,<j}\big)$$

Absorbing Discrete Diffusion = Any-Order AR

Forward process: Each token is independently replaced with $[\mathbf{M}]$ with probability $1 - e^{-\bar{\sigma}(t)}$:

$$q(x_t^i \mid x_0^i) = \begin{cases} x_0^i & \text{with prob. } e^{-\bar{\sigma}(t)} \\ [\mathbf{M}] & \text{with prob. } 1 - e^{-\bar{\sigma}(t)} \end{cases}$$

Reverse process: The model learns to predict the original token at each masked position, while keeping visible tokens unchanged. This objective is time-independent and the concrete score decomposes as:

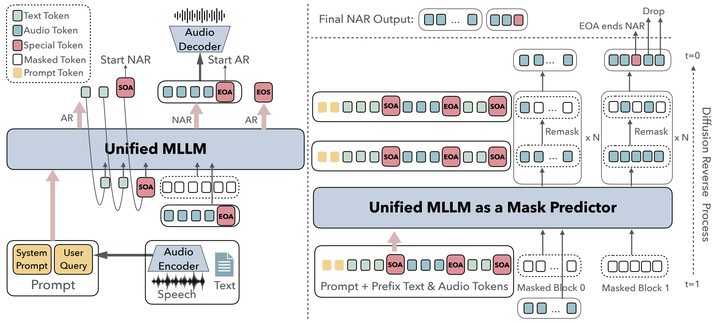
$$\underbrace{\frac{p_t(\ldots, x^i{=}v, \ldots)}{p_t(\ldots, x^i{=}[\mathbf{M}], \ldots)}}_{\text{concrete score}} = \underbrace{\frac{e^{-\bar{\sigma}(t)}}{1 - e^{-\bar{\sigma}(t)}}}_{\text{time scalar}} \cdot \underbrace{p_0(v \mid x_{\text{vis}})}_{\text{clean conditional}}$$

This is equivalent to an **any-order AR model, averaged over all possible token orderings**:

$$\mathcal{L}_{\text{AO}}(x) = \mathbb{E}_{\pi \sim U_\pi} \sum_{l=1}^{L} -\log\, q_\theta\big(x^{\pi(l)} \mid x^{\pi(<l)}\big)$$

Our Method - TtT

Core Idea: TtT is a single Transformer (initialized from a pretrained LLM) that unifies AR text generation and non-autoregressive (NAR) audio diffusion, alternating AR decoding and block-wise diffusion via $\langle\text{SOA}\rangle$ and $\langle\text{EOA}\rangle$.



Partial-Order Formulation: We model the interleaved sequence as a poset $(V, \preceq)$, where text is causal $t_{m,j} \preceq t_{m,j+1}$, span order is preserved (all tokens in span m precede span $m{+}1$), and each audio span $\mathbf{A}_m$ is an anti-chain (no mandatory within-span order). Under this construction, the predecessor sets are:

$$\text{Pa}(t_{m,j}) = \mathbf{T}_{<m} \cup \mathbf{A}_{<m} \cup t_{m,<j}, \quad \text{Pa}(a_{m,j}) = \mathbf{T}_{\leq m} \cup \mathbf{A}_{<m}$$

Within $\mathbf{A}_m$, we sample $\pi_m \sim U_{\pi_m}$ and use $a_{m,\pi_m(<j)}$ to denote tokens generated before index j under π_m.

Training objective: Combining fixed-order AR for text with **order-marginalized** any-order AR for audio:

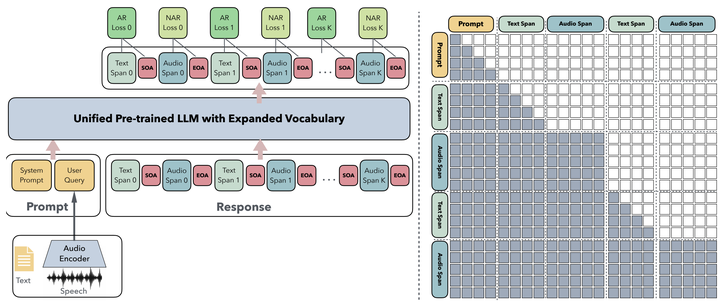
$$\tilde{p}_\theta(x) = \prod_{m=1}^{M} \Bigg[\underbrace{\prod_{j=1}^{|\mathbf{T}_m|} p_\theta\big(t_{m,j} \mid \mathbf{T}_{<m}, \mathbf{A}_{<m}, t_{m,<j}\big)}_{\text{fixed-order AR for text}} \cdot \underbrace{\tilde{p}_\theta\big(\mathbf{A}_m \mid \mathbf{T}_{\leq m}, \mathbf{A}_{<m}\big)}_{\text{order-marginalized for audio}} \Bigg]$$

We theoretically show that our practical unified objective upper-bounds the negative log-likelihood of the order-marginalized hybrid AR-NAR joint distribution (see the paper for the full derivation). Therefore, our training objective is:

$$\mathcal{L}_{\text{Unified}}(x) \triangleq \mathcal{L}_{\text{AR}}(x) + \mathcal{L}_{\text{AO}}(x) \geq -\log\, \tilde{p}_\theta(x)$$

Training Strategy: Bridging train-test discrepancy with only one-forward pass during training:

- **BANOM:** With probability p_{mix}, skip diffusion and apply $\mathcal{L}_{\text{AR}}$ only.
- **PPM:** With probability p_{prefix}, keep $\mathcal{A}_{<m}$ clean and apply diffusion loss on $\mathcal{A}_{\geq m}$.
- **SST:** With probability p_{trunc}, truncate $\mathcal{A}_M$ by removing $\langle\text{EOA}\rangle$ and suffix tokens.



Experiments

Datasets: We use the following evaluation datasets for Audio-QA, automatic speech recognition (ASR) and automatic audio caption (AAC) tasks:

Dataset	Language	Task Type	Abbreviation
AlpacaEval	English	Audio-QA	AE.
LLaMAQuestions	English	Audio-QA	LQ.
TriviaQA	English	Audio-QA	TQA.
WebQuestions	English	Audio-QA	WQ.
Fleurs-zh	Chinese	ASR	Fzh.
AISHELL-2	Chinese	ASR	A2.
AISHELL-1	Chinese	ASR	A1.
WenetSpeech-test_meeting	Chinese	ASR	WS_m.
WenetSpeech-test_net	Chinese	ASR	WS_n.
Fleurs-en	English	ASR	Fen.
Clotho-v2	English	AAC	Clo.
MACS	English	AAC	MACS

Exp. 1: Our proposed TtT consistently outperforms both pure AR and NAR variants across all metrics.

Models	Audio-QA (↑)				ASR (↓)					
	AE.	LQ.	TQA.	WQ.	Fzh.	A2.	A1.	WS_m.	WS_n.	Fen.
Main Results										
Qwen2.5-1.5B (AR)	10.85	1.00	0.00	0.10	103.18	81.84	95.96	103.15	95.54	105.62
Qwen2.5-1.5B (NAR)	10.70	0.00	0.40	0.20	86.97	224.37	191.11	123.96	143.76	108.25
TtT-1.5B (AR–NAR)	**15.68**	**23.75**	**3.47**	**7.70**	**44.36**	**14.89**	**16.72**	**52.23**	**41.52**	**49.00**
Qwen2.5-3B (AR)	14.42	10.00	0.60	0.70	90.32	54.94	72.01	80.01	73.64	74.47
Qwen2.5-3B (NAR)	11.31	0.67	1.21	0.70	68.94	212.27	160.58	89.22	111.29	83.51
TtT-3B (AR–NAR)	**17.46**	**34.68**	**6.53**	**11.61**	**55.67**	**12.53**	**13.65**	**53.83**	**44.29**	**64.31**
Ablation Study										
TtT-3B w/o BANOM	13.87	19.87	2.81	5.12	58.25	18.58	21.35	58.48	49.52	68.87
TtT-3B w/o PPM	14.27	22.79	2.71	5.54	58.86	15.63	18.83	57.76	47.92	67.37
TtT-3B w/o SST	14.12	10.20	1.30	3.72	56.39	25.43	31.03	64.41	56.70	**62.60**
TtT-3B (AR–NAR)	**17.46**	**34.68**	**6.53**	**11.61**	**55.67**	**12.53**	**13.65**	**53.83**	**44.29**	64.31
Training Strategy Comparison										
TtT-3B (AR–NAR)	17.46	34.68	6.53	11.61	55.67	12.53	13.65	53.83	44.29	64.31
Pretrain+AR	**29.45**	15.93	3.61	11.45	23.37	9.79	12.67	**26.75**	20.91	19.49
Pretrain+TtT	26.73	**40.07**	**11.07**	**21.43**	**18.99**	**6.80**	**5.78**	27.59	**19.85**	**19.10**

Exp. 2: Our proposed Pretrain+TtT (3B) achieves performance that is competitive with much larger 7B-9B audio-language models.

Models	Size	Audio-QA (↑)				ASR (↓)						AAC (↑)	
		AE.	LQ.	TQA.	WQ.	Fzh.	A2.	A1.	WS_m.	WS_n.	Fen.	Clo.	MACS
Large Models (> 7B)													
Moshi	7B	25.63	48.30	16.75	16.85	-	-	-	-	-	-	4.32	12.01
SpeechGPT	7B	10.00	30.96	16.53	24.53	101.45	120.77	111.81	123.15	124.86	45.15	2.10	3.95
Kimi-Audio	7B	19.49	**57.53**	43.51	43.20	**2.87**	**2.53**	**0.61**	**6.34**	**5.39**	**4.87**	**55.92**	**64.90**
VITA-Audio	7B	40.20	54.30	18.59	30.75	6.35	5.56	4.58	20.38	15.88	9.58	6.18	7.94
LLaMA-Omni	8B	39.59	48.46	21.80	30.28	-	-	-	-	-	-	2.53	4.56
GLM-4-Voice	9B	**44.87**	**62.67**	**44.99**	**48.47**	-	-	-	-	-	-	13.15	12.67
Efficient Models (≤ 3B)													
Mini-Omni	0.5B	15.73	2.00	1.10	2.42	182.73	342.40	442.06	294.42	335.80	22.74	3.61	4.45
SLAM-Omni	0.5B	17.47	24.75	3.51	7.90	-	-	-	-	-	-	**54.52**	**50.46**
Qwen2.5-3B (AR)	3B	14.42	10.00	0.60	0.70	90.32	54.94	72.01	80.01	73.64	74.47	9.73	48.64
Qwen2.5-3B (NAR)	3B	11.31	0.67	1.21	0.70	68.94	212.27	160.58	89.22	111.29	83.51	9.54	27.40
TtT	3B	17.46	34.68	6.53	11.61	55.67	12.53	13.65	53.83	44.29	64.31	12.63	48.87
Pretrain+TtT	3B	**26.73**	**40.07**	**11.07**	**21.43**	**18.99**	**6.80**	**5.78**	**27.59**	**19.85**	**19.10**	11.55	42.86

Conclusion

Text and audio have fundamentally different dependency structures, and respecting this asymmetry through hybrid AR-NAR training yields consistent gains. (We should rethink visual-language MLLMs)

Contact: lixueyi@stu2021.jnu.edu.cn