From Pixels to Voice:
A Simple and Efficient End-to-End Spoken Image Description Approach via Vision Codec Language Models

Chung Tran - Sakriani Sakti
Nara Institute of Science and Technology

Introduction

- ❖ Generative models in NLP excel in sequential input like VALL-E [1] (TTS), VioLa [2], LauraGPT [3] (audio generation)
- ❖ Challenge: Non-sequential input (Image) remains underexplored
- ❖ Application: Image-to-text/speech models can assist visually impaired people.
- ❖ Problem: Many languages lack standard writing systems [4], limiting text-based technology.

Related Works

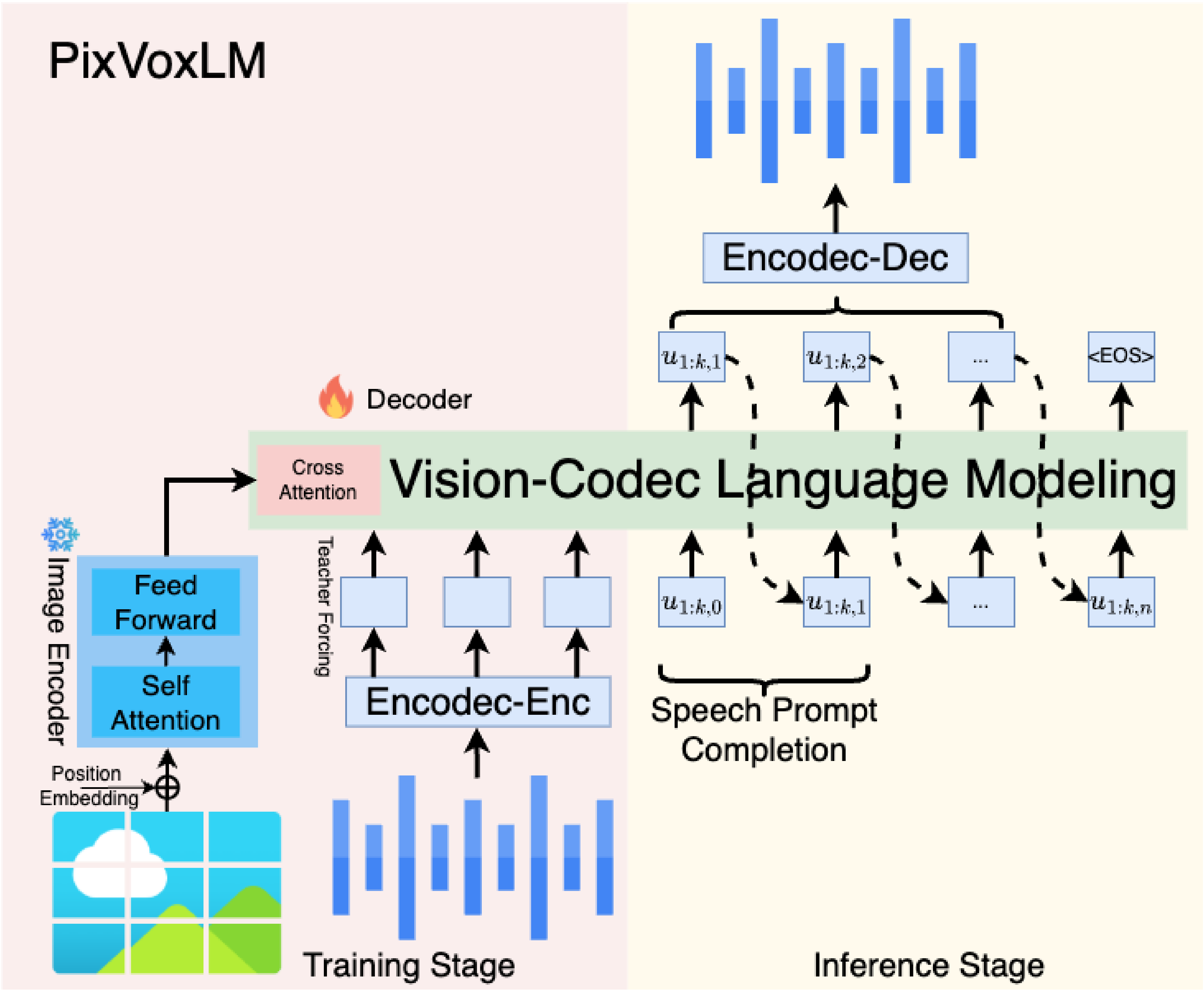
❖ Recent studies can describe images in speech
→ SAT[5], E-I2S [6], Im2p [7]: Train multiple components:
I2U + U2S or VQ-VAE + Vocoder
→ Show-and-Speak (SAS[8]) use Faster-RCNN + modified Tacotron2

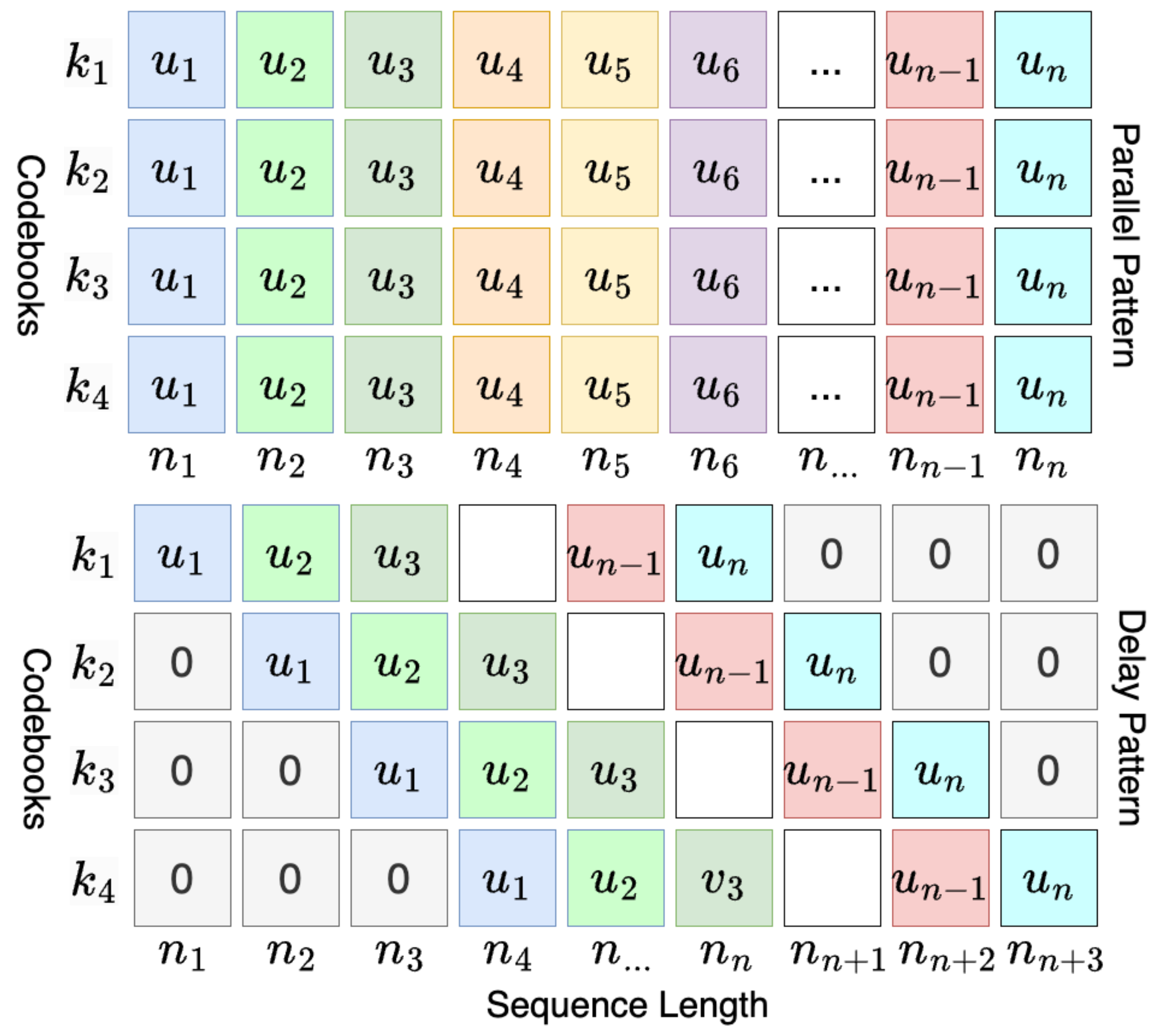
❖ Limitation:
→ Multi-component training is complex
→ Depend on the external model to extract feature

Proposal Approach

❖ Use audio codec model to extract discrete representations and reconstruct it into speech
→ Simplifies I2S training
+ (single model-E2E).

Methodology





k_1	u_1	u_2	u_3	u_4	u_5	u_6	...	u_{n-1}	u_n
k_2	u_1	u_2	u_3	u_4	u_5	u_6	...	u_{n-1}	u_n
k_3	u_1	u_2	u_3	u_4	u_5	u_6	...	u_{n-1}	u_n
k_4	u_1	u_2	u_3	u_4	u_5	u_6	...	u_{n-1}	u_n
	n_1	n_2	n_3	n_4	n_5	n_6	$n_{...}$	n_{n-1}	n_n

k_1	u_1	u_2	u_3		u_{n-1}	u_n	0	0	0
k_2	0	u_1	u_2	u_3		u_{n-1}	u_n	0	0
k_3	0	0	u_1	u_2	u_3		u_{n-1}	u_n	0
k_4	0	0	0	u_1	u_2	v_3		u_{n-1}	u_n
	n_1	n_2	n_3	n_4	$n_{...}$	n_n	n_{n+1}	n_{n+2}	n_{n+3}

Speech Encoding and Reconstruction

$$U = Encodec - Enc(S)$$

$$\hat{S} = Encodec - Dec(U)$$

Image-2-Unit (I2U) Mapping

$$\hat{U} = I2U(I)$$

Objective function:

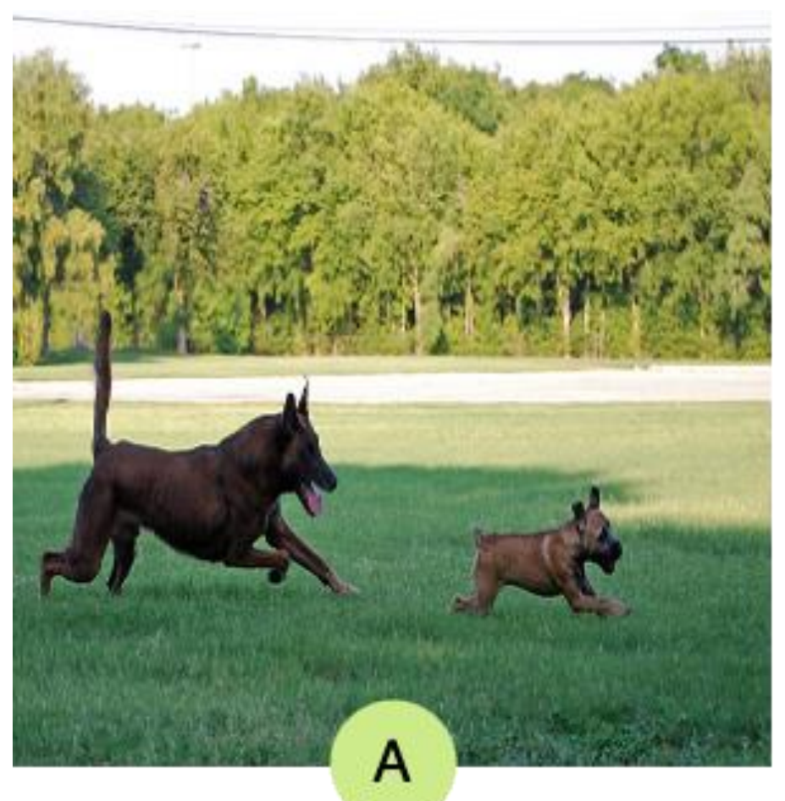
$$L = \sum_{i=1}^{k} \sum_{n=1}^{N} \hat{u} \log p(\hat{u}|u)$$

TABLE I
PERFORMANCE COMPARISON OF PIXVOXLM WITH EXISTING I2S MODELS ACROSS VARIOUS EVALUATION METRICS

Methods	B1↑	B2↑	B3↑	B4↑	M↑	R↑	C↑
Multiple-Model Training							
SAT [12]	-	-	-	11.60	14.10	39.00	23.20
SAT-FT [12]	-	-	-	12.60	14.50	39.10	24.20
E-I2S [14]	-	-	-	14.78	17.40	45.75	32.89
Single-Model Training							
SAS [18]	29.60	14.70	7.20	3.50	11.30	23.20	8.00
PixVoxLM-Parallel	34.52	18.75	10.65	6.22	10.51	26.30	9.43
PixVoxLM-Delay	48.08	30.59	18.92	11.49	15.19	35.76	25.54

Result

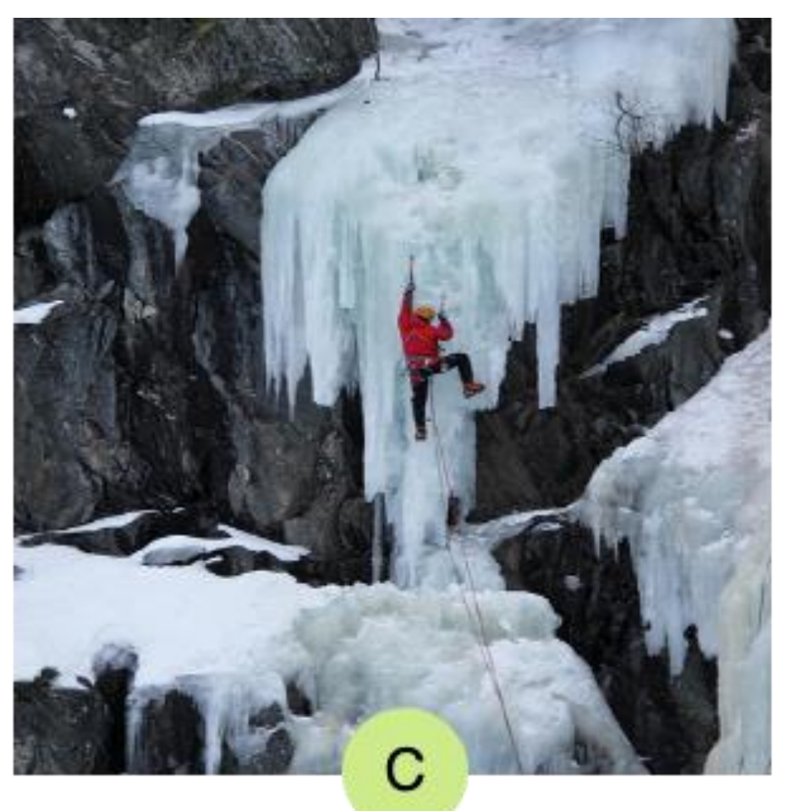
- ❖ Delay Pattern performs better than Parallel Pattern
- ❖ PixVoxLM outperforms the end-to-end SAS model
- ❖ PixVoxLM (delay pattern) is better than SAT and SAT-FT in the M and C metrics.


A

GT: Two dogs play in the grass
ASR: Two dogs running in grass

B

GT: Three children playing in sand at beach
ASR: Thre children playing in the sand


C

GT: A man climbs icy rocks
ASR: Clamber or climbing a neste

TABLE II
SPEECH PROMPT COMPLETION AT VARIOUS INFORMATION LEVELS

PixVoxLM	Prompt	B1↑	B2↑	B3↑	B4↑	M↑	R↑	C↑
Parallel	0%	34.52	18.75	10.65	6.22	10.51	26.30	9.43
	25%	37.11	23.30	14.58	8.87	13.05	29.71	15.24
	50%	46.26	34.00	26.25	20.42	20.54	40.83	37.64
Delay	0%	48.08	30.59	18.92	11.49	15.19	35.76	25.54
	25%	49.76	34.18	23.04	14.90	18.00	39.04	32.68
	50%	57.31	44.3	35.31	28.11	24.19	48.35	60.10

- ❖ Use image and partial speech inputs for more accurate and context-aware completions.
- ❖ Delay pattern have better result than Parallel

Conclusion

- ❖ PixVoxLM offers a simple and efficient solution for generating speech directly from images
- ❖ PixVoxLM outperform the recent end-to-end SAS model

Future work

- ❖ Subjective evaluations highlight several issues
- ❖ Need to improve the performance

[1] Alexandre Défossez et al, High Fidelity Neural Audio Compression
[2] Tianrui Wang et al, VioLA: Unified Codec Language Models for Speech Recognition, Synthesis, and Translation
[3] Zhihao Du et al, LauraGPT: Listen, Attend, Understand, and Regenerate Audio with GPT
[4] Gilles Adda et al, Breaking the Unwritten Language Barrier
[5] Wei-Ning Hsu et al, Text- Free Image-to-Speech Synthesis Using Learned Segmental Units
[6] Johanes Effendi et al, End-to-end image-to-speech generation for untranscribed unknown languages
[7] Minsu Kim et al, Towards practical and efficient image-to-speech captioning with vision-language pre-training and multi-modal tokens
[8] Xinsheng Wang et al, Show and Speak: Directly Synthesize Spoken Description of Images