1P2-C10

session: Communication Robot

time: 30th May, 14:30-16:00

presenter: QUANG TIEN DAM
(Ritsumeikan University)

Facial Reaction Generation with Finite Scalar Quantization and Cross-modality Transformer

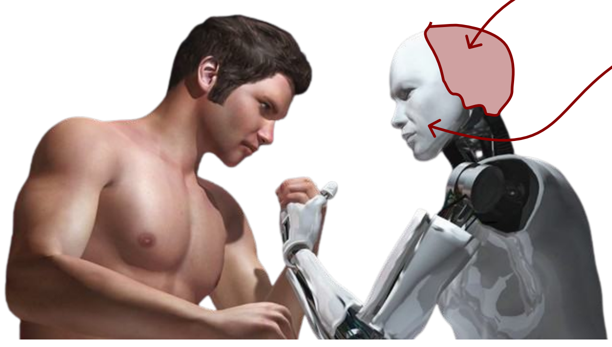
- Quang Tien DAM,
  Tri Tung Nguyen NGUYEN,
  Yasuyuki FUJII,
  Dinh Tuan TRAN,
  Joo-Ho LEE

RITSUMEIKAN UNIVERSITY

Advanced Intelligent System Lab

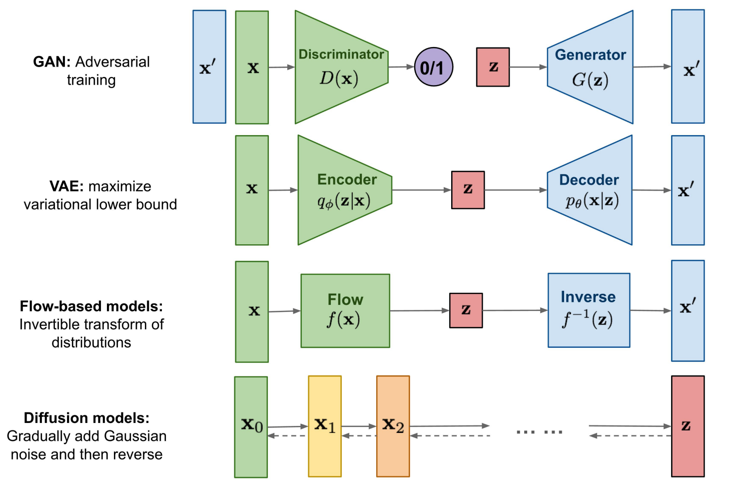
Research background & problem definition

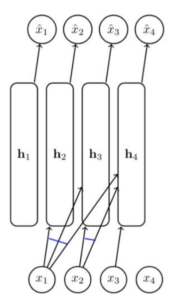
Human-robot conversation



- Language and sound model is advancing very fast OpenAI Gemini
- Facial expression make interaction more natural

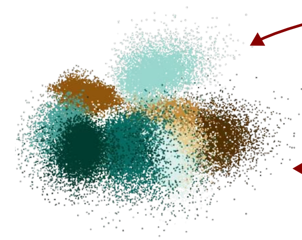
Generative model taxonomy





Auto-regressive model

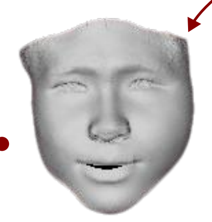
Image source: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/



Encode

Regress

Image space
TransVAE



Encode

Mesh space
BeLFusion

Latent space

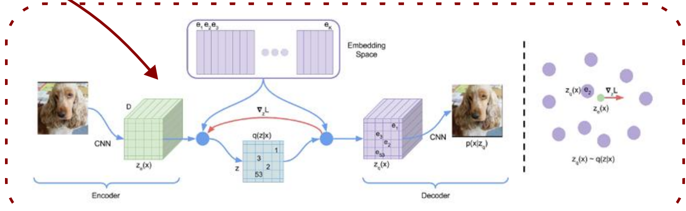
Learn2Listen
QPGesture
Bailando

Quantize

75 76 77 78

Facial expression token

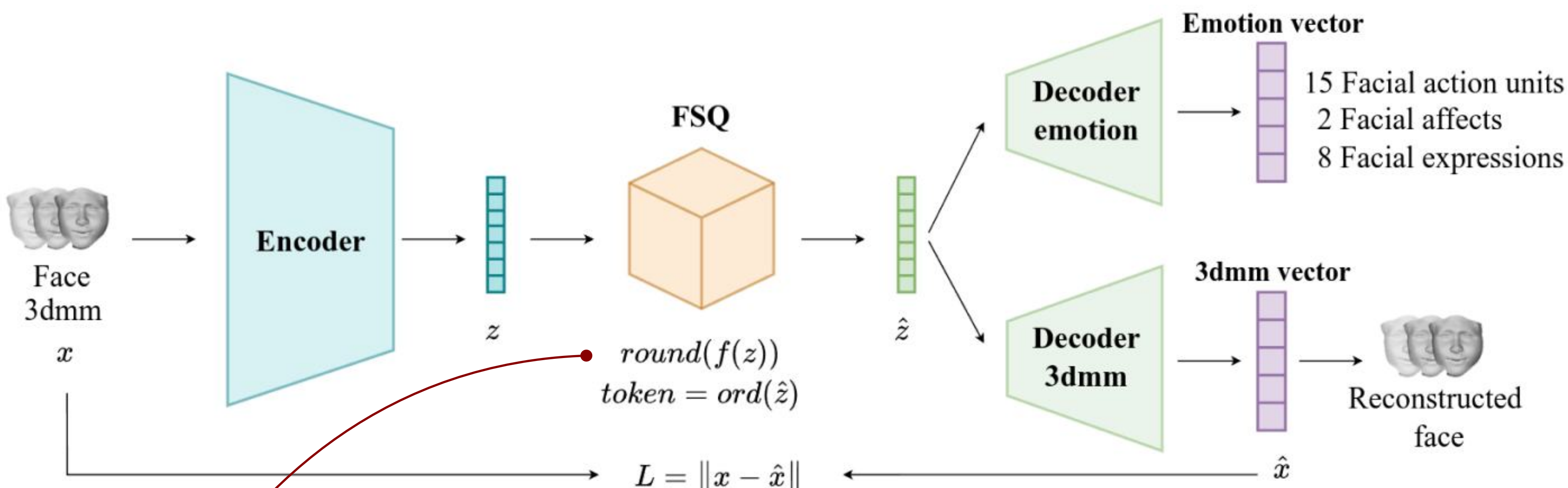
Use **VQ-VAE** as the vector quantizer.



Problems with VQ-VAE
- Highly sparse codebook
- Overlapping embedding
- Codebook collapse

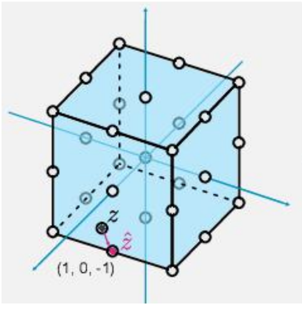
Methodology

A face state is a token.



Blendshape loss optimizes with more addressing to face features.

$$L_{Blendshape}(E, D_{3dmm}) = \|x_{eyebrow} - \hat{x}_{eyebrow}\| + \|x_{eyemovement} - \hat{x}_{eyemovement}\| + \dots + \|x_{rotation} - \hat{x}_{rotation}\| + \|x_{translation} - \hat{x}_{translation}\| .$$

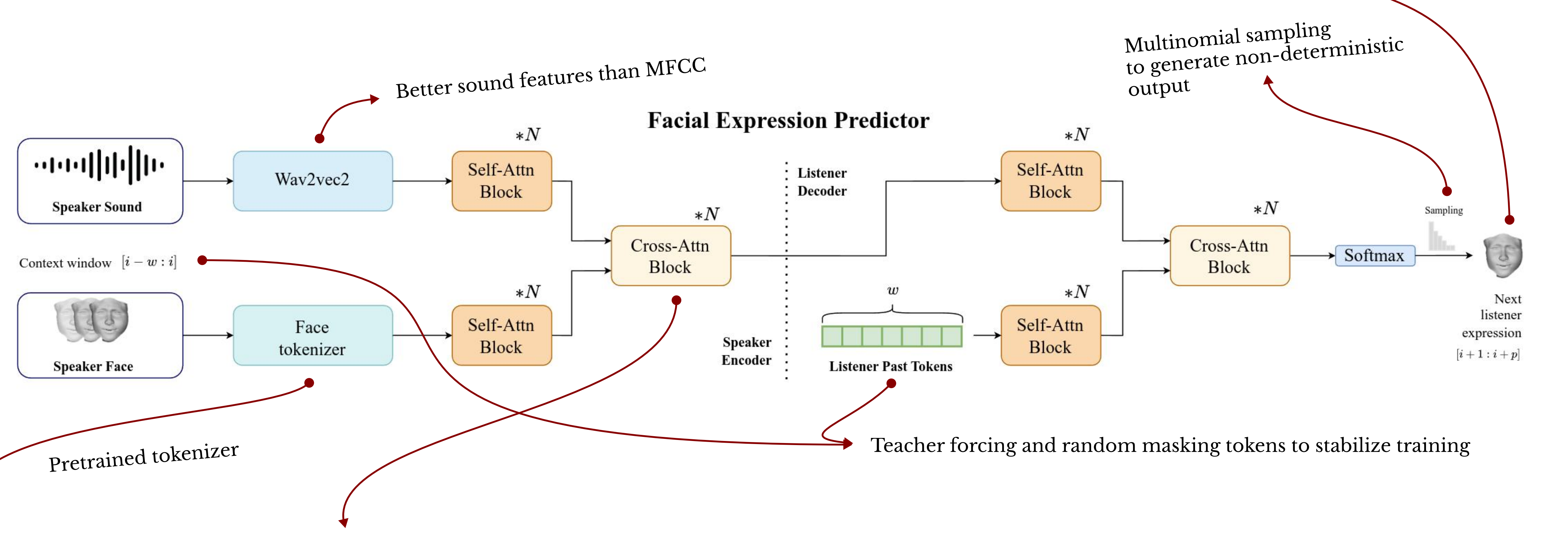


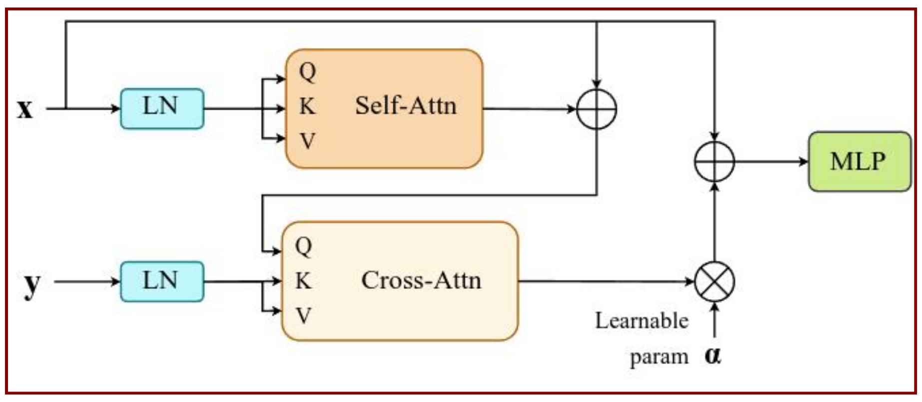
$$\hat{z}_i = f(z_i) := \text{round}\left(\left\lfloor \frac{L}{2} \right\rfloor \tanh(z_i)\right) \in \{-1, 0, 1\}.$$

FSQ enforce stronger regularization creating a more meaningful and compressed latent space.

$$L(D_{emotion}) = \|x_{emotion} - \hat{x}_{emotion}\| .$$

Autoregressively predict next reaction





This predictor contains two parts:

- Speaker feature encoder to process sound and facial expression of the human.
- Listener decoder to align next feature to its past facial expression.

Evaluation and results

TABLE I

BASELINES. COMPARISON OF OUR APPROACH WITH BASELINE MODELS [14] ON THE TEST SET.

	Appropriateness		Diversity			Realism	Synchrony
	FRCorr (↑)	**FRDist** (↓)	**FRDiv** (↑)	**FRVar** (↑)	**FRDvs** (↑)	**FRRea** (↓)	**FRSyn** (·)
Ground truth	8.73	0.00	0.0000	0.0724	0.2483	-	47.69
Random	0.05	237.23	0.1667	0.0833	0.1667	-	44.10
Mime	0.38	92.94	0.0000	0.0724	0.2483	-	38.54
MeanSeq	0.01	97.13	0.0000	0.0000	0.0000	-	45.28
MeanFr	0.00	97.86	0.0000	0.0000	0.0000	-	49.00
Trans-VAE	0.07	90.31	0.0064	0.0012	0.0009	69.19	44.65
BeLFusion(k=10)+BinarizedAUs	0.12	94.09	0.0379	0.0248	0.0397	-	49.00
Ours	**0.31**	**84.93**	**0.1164**	**0.0348**	**0.1166**	**34.66**	**47.42**

(·) means the closer to the ground truth, the better.
indicates the best average performance among the heuristic baselines for the groups of metrics.

Use NoXi and RECOLA internet conference dataset, we evaluated our method using metrics and baselines proposed in the REACT Competition 2024.

priority when designing the model

	Appropriateness		Diversity			Realism	Synchrony
	FRC	FRD	FRDvs	FRVar	FRDiv	FRRea	FRSyn
FSQ-val	**0.2737**	**86.6145**	0.1162	0.0345	0.1163	81.2801	45.7206
LFQ-val	0.2625	99.8672	0.1213	0.0434	0.1213	73.2092	45.8896
VQ-val	0.2693	91.5249	0.0943	0.0370	0.0943	96.1280	46.2099

References:

[NoXI database] A. Cafaro, et al. The noxi database: multimodal recordings of mediated novice-expert interactions. ICMI '17.

[RECOLA database] F. Ringeval, et al. Introducing the recola multimodal corpus of remote collaborative and affective inter actions. FG '13.

[REACT Competition 2024] Song, et al. "REACT 2024: The Second Multiple Appropriate Facial Reaction Generation Challenge." arXiv, January 10, 2024.

Please check qualitative results in our presentation laptop.

Key takeaway

1. We turn **facial expressions into a finite meaningful vocabulary** using Finite Scalar Quantization.
2. Then, we use an **autoregressive cross-modality transformer-based** model to generate multiple appropriate facial responses in dyadic conversation context.
3. The method achieves the best performance in the REACT2024 challenge.