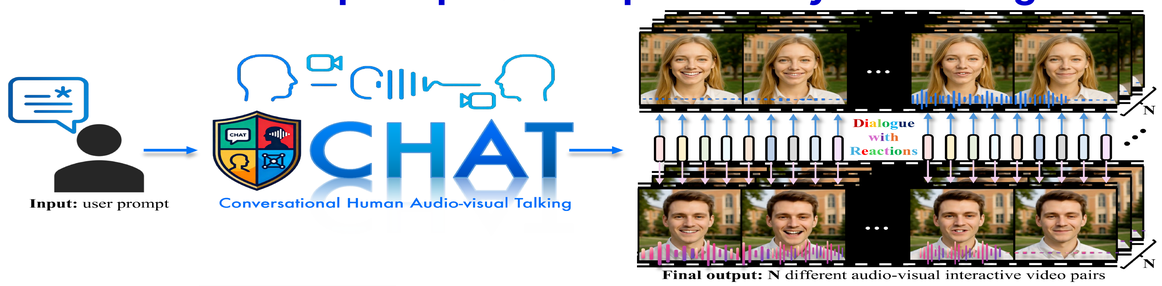
Conversational Human Audio-visual Talking Dialogue Generation

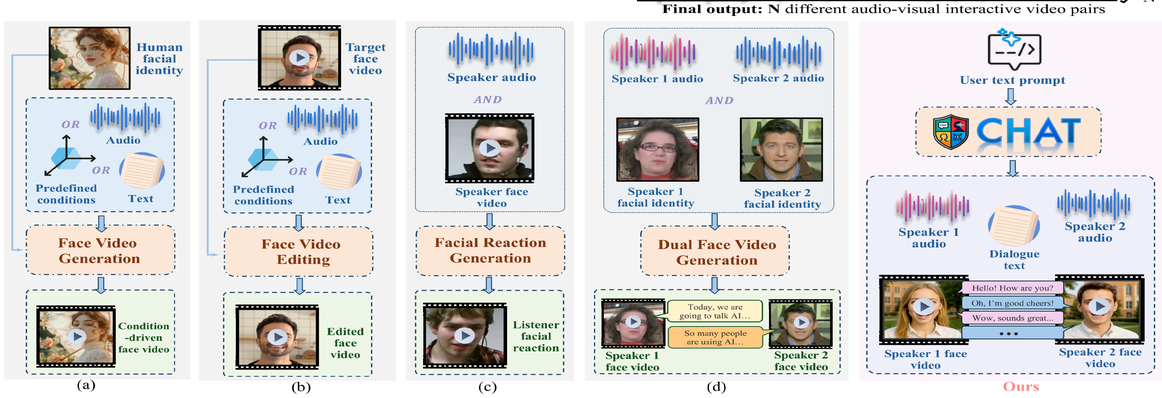
Junhao Song[1], Lluis Guasch[1], Xilin He[2], Zhongyu Yang[3], Yingfang Yuan[4],
Weicheng Xie[5], Linlin Shen[5], Haijun Lin[6], Shizhe Liu[7], Wei Pang[3], Siyang Song[8]

[1] Imperial College London, [2] Mohamed bin Zayed University of Artificial Intelligence, [3] Heriot-Watt University, [4] Northumbria University,
[5] Shenzhen University, [6] Hunan Normal University, [7] University of Oxford, [8] University of Exeter

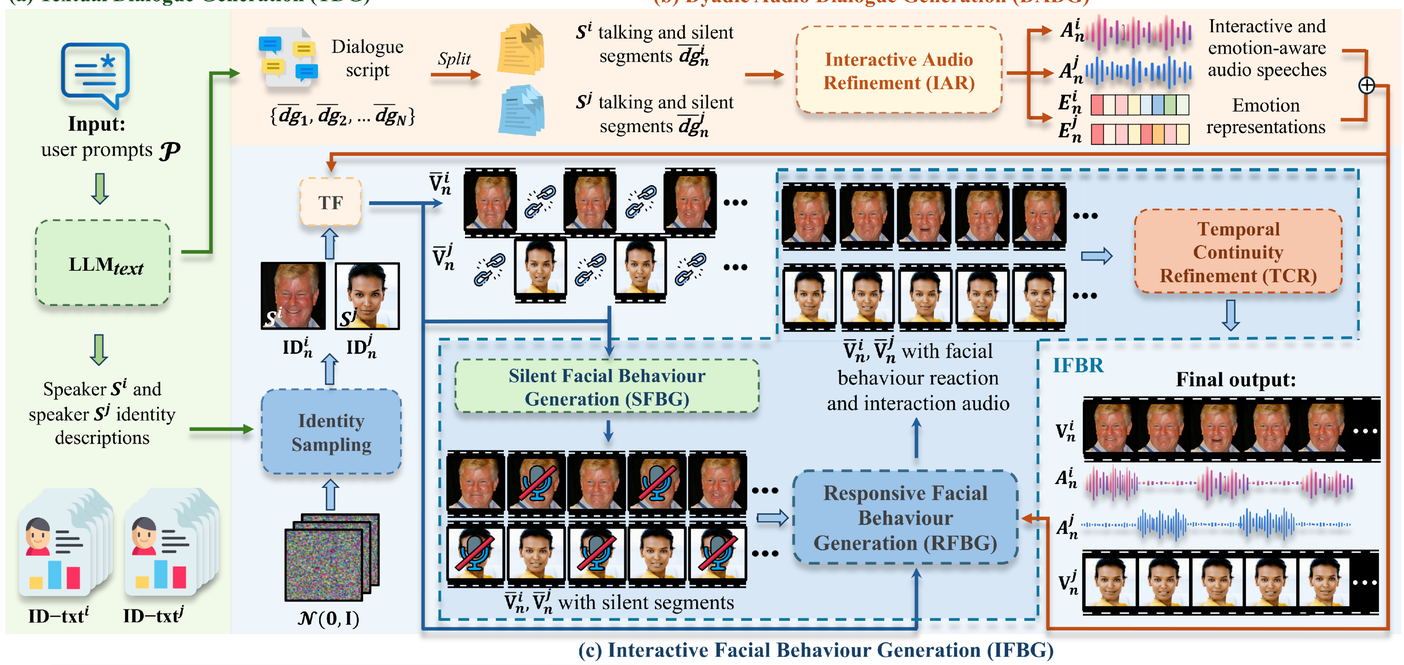
ECCV 2026 | MALMÖ | SEPT 8–12

1. Our task: one prompt → a responsive dyadic dialogue





2. CHAT framework: Text → {TDG → DADG → IFBG} → Audio-visual



3. Results of experiments

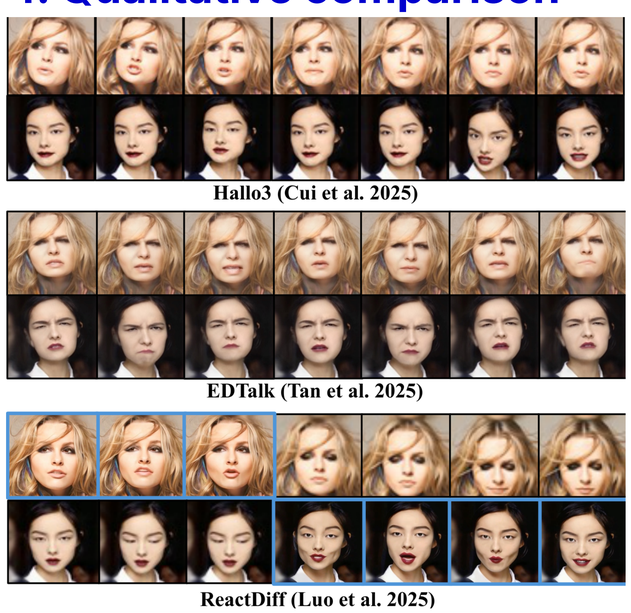
Method	Visual Quality		AV Sync		Reaction ($\times 10^{-2}$)		ID Preservation	
	FID ↓	FVD ↓	LSE-C ↑	LSE-D ↓	FRCorr ↑	FRDiv ↑	CSIM ↑	LPIPS ↓
Hallo3 [10]	20.56	**362.23**	4.61	10.41	N/A	N/A	0.79	0.55
SadTalker [74]	22.53	385.51	6.85	8.17	N/A	N/A	0.81	0.48
DIM [61]	36.52	460.35	6.82	8.89	31.02	12.55	0.83	0.52
EDTalk [60]	18.74	619.92	5.62	9.61	N/A	N/A	0.78	**0.36**
ReactDiff [40]	21.36	386.22	5.23	9.02	48.05	15.31	0.77	0.58
CHAT (Ours)	**17.33**	365.03	**6.89**	**8.07**	**50.02**	**15.34**	**0.85**	0.38

Quantitative comparison of CHAT and baselines on evaluation set (mean of 1000 samples). The best result is in **bold** and the second best in underline.

Method	Visual	Audio	Sync	Expr.	Interact.
Hallo3 [10]	4.5	3.5	3.9	3.8	2.5
SadTalker [74]	2.5	3.7	1.2	2.0	1.3
DIM [61]	3.7	3.1	3.4	2.8	3.6
EDTalk [60]	2.2	3.6	2.3	2.8	2.7
ReactDiff [40]	4.2	3.5	3.8	4.1	3.2
CHAT (Ours)	**4.6**	**4.8**	**4.6**	**4.6**	**4.8**

Qualitative user study results from 60 participants. Mean opinion scores for each method and evaluation dimension. The best result is in **bold** and the second best in underline.

4. Qualitative comparison



CHAT also used a method to enhance generated audio-visual coherence and quality in the IFBG: