APPENDIX

A. Qualitative Results and Analysis

In Figure 4, we present dynamic emotional changes across frames within a single video at the first three frames from neutral to disgust. While MAVFlow effectively captures the emotional change of Ground Truth (GT) video from 15th frame, reflecting the shift starting from the 14th frame. AV2AV fails to reflect the emotion until around the 36th frame. Additionally, overall, MAVFlow better expresses emotions as well as the arousal level, which is indicated by the distance of a red dot from the center. Cascaded systems have been excluded from the comparison due to poor temporal alignment and their inability to embed emotional cues into the audio, which results in TFG output cannot reflect emotional expressions in the video. The DTW and DTW-SL metrics in Table 1 and Table 2, further confirm the notably poor temporal alignment of the cascaded systems.

B. Class-Wise Emotional Analysis

B.1. Audio Emotional Results

In Table 9, we evaluate the class-wise emotion recognition accuracy of the generated audio using the pretrained emotion2vec [40]. Compared to AV2AV, MAVFlow shows slightly lower performance for the Sad, Disgust, and Fear classes, while demonstrating comparable or superior results for Happy, Neutral, and Angry. Notably, MAVFlow exhibits a significant advantage in the Angry class, ultimately achieving better overall performance than AV2AV in both Emo-Acc and ES metrics (as shown in Table 7). Furthermore, the MAVFlow + model, trained with additional emotional datasets, achieves improved performance across most emotion classes, with a substantial gain in overall Emo-Acc.

Table 9. Class-wise emotion accuracy (%) of generated audio (+: additional training on CREMA-D).

Method	Happy	Sad	Neutral	Angry	Disgust	Fear	Emo-Acc ↑
GT	89.29	85.00	89.17	89.29	77.86	62.14	81.95
AV2AV	30.00	**22.86**	**80.00**	28.57	30.71	16.43	33.66
MAVFlow	36.43	11.43	**80.00**	62.86	20.00	14.29	36.46
MAVFlow +	**69.29**	**22.86**	66.67	**80.71**	**32.86**	**38.57**	**51.46**

B.2. Visual Emotional Results

In Table 10, we evaluated class-wise visual emotion accuracy using pretrained MAE-DFER [59]. Also, follow MAE-DFER, we report both Unweighted Average Recall (UAR) and Weighted Average Recall (WAR) as evaluation metrics. UAR calculates the average recall by treating each class equally, which helps account for class imbalance, while WAR weights the recall by the number of samples per class, reflecting the actual class distribution in the dataset. As a result, MAVFlow achieved strong performance in terms of both UAR and WAR, particularly excelling in the angry, disgust, and fear emotion classes.

C. Inference Time Comparison

MAVFlow does not rely on intermediate text representations, resulting in faster inference compared to the cascaded system. Furthermore, it is more efficient by applying the speed-friendly CFM module compare to diffusion model. We compared the inference speed using one A6000 GPU, observing processing times of 1.66s for MAVFlow, 1.22s for AV2AV, and 1.75s for the 4-cascaded model to handle a 2.35s audio-visual input through the complete pipeline.

D. Limitation

MAVFlow currently leverages emotional embeddings only from face and speaker embeddings from audio. However, we believe that incorporating emotional cues from audio (*e.g.*, prosody, timbre, and other paralinguistic features) into the guidance of CFM could further enhance performance. Furthermore, since we directly adopt the unit extractor and unit-to-unit translation modules from previous work [13], improving semantic translation quality remains an open challenge.

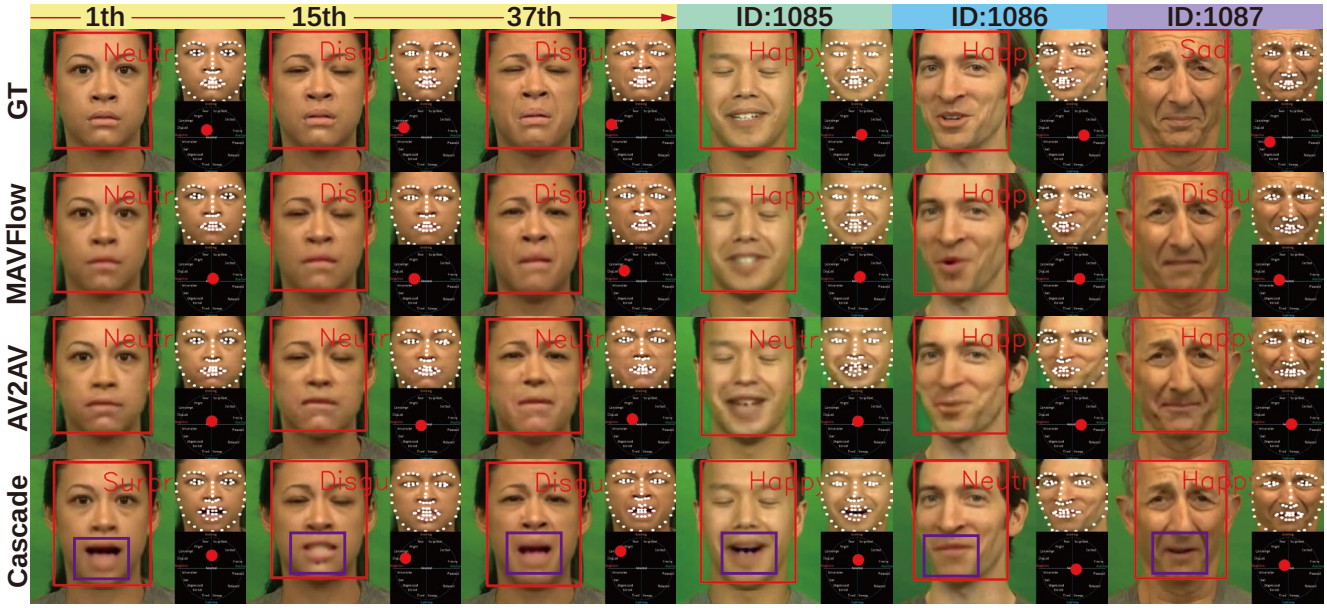

Figure 4. Additional qualitative comparison for frame-level analysis. Each row shows GT, MAVFlow, AV2AV, and Cascade (ASR+XTTS+TFG), respectively.

Table 10. Class-wise emotion accuracy, unweighted and weighted average recall (UAR%, WAR%), and ES of the generated visuals, all measured with MAE-DFER (+: additional training on CREMA-D).

Method	Happy	Sad	Neutral	Angry	Disgust	Fear	UAR	WAR	ES
GT	97.14	67.86	76.67	78.57	87.86	52.86	76.83	76.83	1.00
ASR+YourTTS+TFG	89.86	60.71	72.88	50.71	83.57	40.00	66.29	66.05	0.85
ASR+XTTS+TFG	94.93	55.00	72.03	72.86	85.71	31.43	68.66	68.50	0.91
AV2AV	**95.00**	**64.29**	**79.17**	62.14	77.14	27.14	67.48	67.20	0.87
MAVFlow	**95.00**	53.57	75.00	**80.71**	**88.57**	**43.57**	72.74	72.68	0.92
MAVFlow +	**95.00**	63.57	78.33	76.43	87.14	37.86	**73.06**	**72.93**	**0.93**