

## APPENDIX

### A USER STUDY

Since users’ visual perception of video is more tangible and reliable, we invite 20 scholars to evaluate the performance of our method and some other representative methods subjectively. The novel evaluation task we designed is to provide the interviewees with several pairs of videos synthesized by different methods, with every method appearing in turn at the same times, and request them to pick what they think is the real video from each pair. To ensure rigor, we will conduct the same task on the different datasets and record the results respectively. At last, we count the percentage of synthesized video that is thought to be real for every method. The percentage above is defined by us in terms of False Positive Rate (FPR). Consequently, the higher FPR is, the better performance the corresponding method carries out. The FPRs of all the methods are shown in Tab. 1, respectively.

From the results, we can clearly see that our proposal gets the highest FPR on both the GRID and LRS2 dataset, surpassing the second place Wav2Lip by approximately 10% on average, which means that more users consider the videos generated by FluentLip are more realistic and natural. In summary, with the help of phoneme driving, optical flow consistency loss and diffusion chain, the facial videos synthesized by FluentLip are more comprehensible and therefore achieve a good FPR performance on user study.

Table 1: Subjective comparisons of the comprehensive performance of different methods by users.

Methods	FPR (%) $\uparrow$		
	GRID	LRS2	Average
Wav2Lip	53.75	66.25	60.00
SadTalker	46.25	12.50	29.38
TalkLip	45.00	65.00	55.50
Diff2Lip	38.75	37.50	38.13
<b>FluentLip</b>	<b>66.25</b>	<b>68.75</b>	<b>67.50</b>