

# GaussianHeadTalk: Wobble-Free 3D Talking Heads with Audio Driven Gaussian Splatting

## Supplementary Material

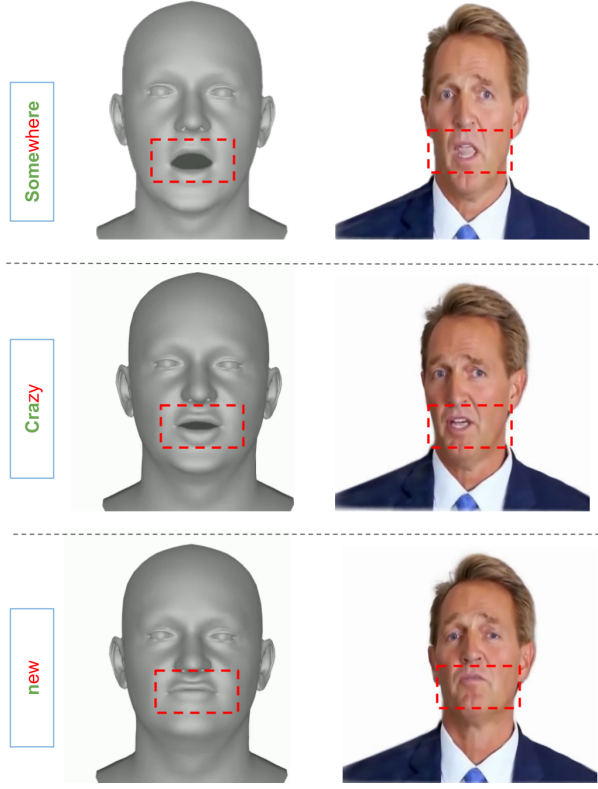


Figure A. For a given audio signal, GaussianHeadTalk generates a lip-sync 3D mesh and use the generated FLAME parameters to transfer lip motion on a trained GaussianAvatar with optimized FLAME parameters.

### A. Qualitative Ablation Study

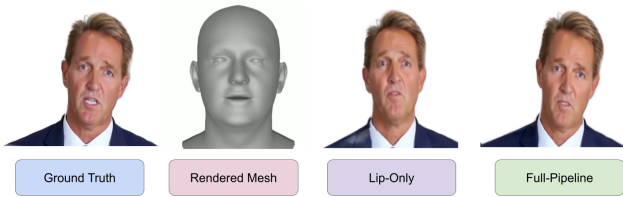


Figure B. Ablation Study: Effect of transferring the lip motion and keeping other parameters static (w/o Full Motion Transfer). The results shows visible artifacts in the generated avatar, as the FLAME parameters are not fully independent.

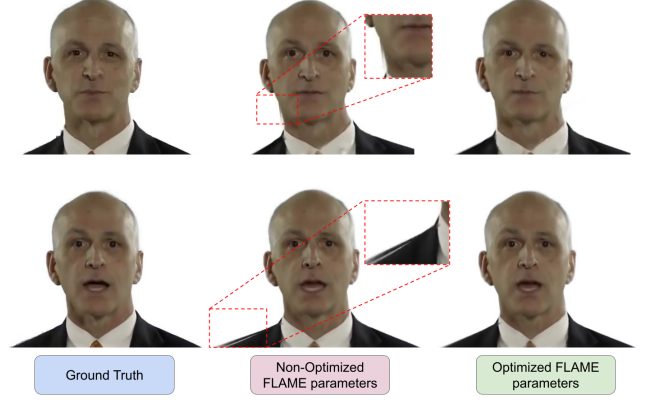


Figure C. Ablation Study: Using Non-Optimized FLAME parameters (w/o Parameter Optimization). This leads to artifacts around the torso region, and wobbling issues.

### B. Temporal Analysis of Keypoints

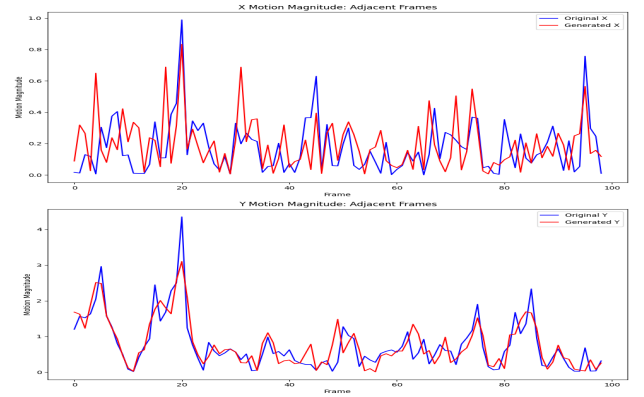


Figure D. Comparison of keypoint movement across time between an original video and a video generated using GaussianTalker. The overlay graph shows that there is flickering in the rendered video. In ideal case, these two graphs should be perfectly overlapping. We report  $x$ -axis,  $y$ -axis motion magnitude over time in upper and lower plots, respectively.

### C. User Study

Category	Method	Final Score	‘Best’ (3)	‘Average’ (2)	‘Worst’ (1)	Total Ratings
Naturalness	<b>GaussianHeadTalk (ours)</b>	<b>9.8</b>	284	14	2	300
	TalkingGaussian [29]	6.2	40	178	82	300
	GaussianTalker [9]	4.0	5	50	245	300
LipSync	<b>GaussianHeadTalk (ours)</b>	<b>7.8</b>	142	118	40	300
	TalkingGaussian [29]	7.2	128	92	80	300
	GaussianTalker [9]	5.0	25	100	175	300
Quality	<b>GaussianHeadTalk (ours)</b>	<b>9.5</b>	260	35	5	300
	TalkingGaussian [29]	6.5	80	125	95	300
	GaussianTalker [9]	4.0	1	58	241	300

Table A. Detailed breakdown of User Study Ratings. 30 participants evaluate 10 videos of each method, generating 300 ratings in total. For each triplet, a participant assigns the ranks 1, 2, and 3 once each, so the raw sum across the three methods is  $1+2+3=6$ . Over 10 triplets, this gives a total of  $10 \times 6=60$ . After dividing each method’s total by 3 for normalization, the overall sum across all methods is fixed at  $60/3 = 20$ .

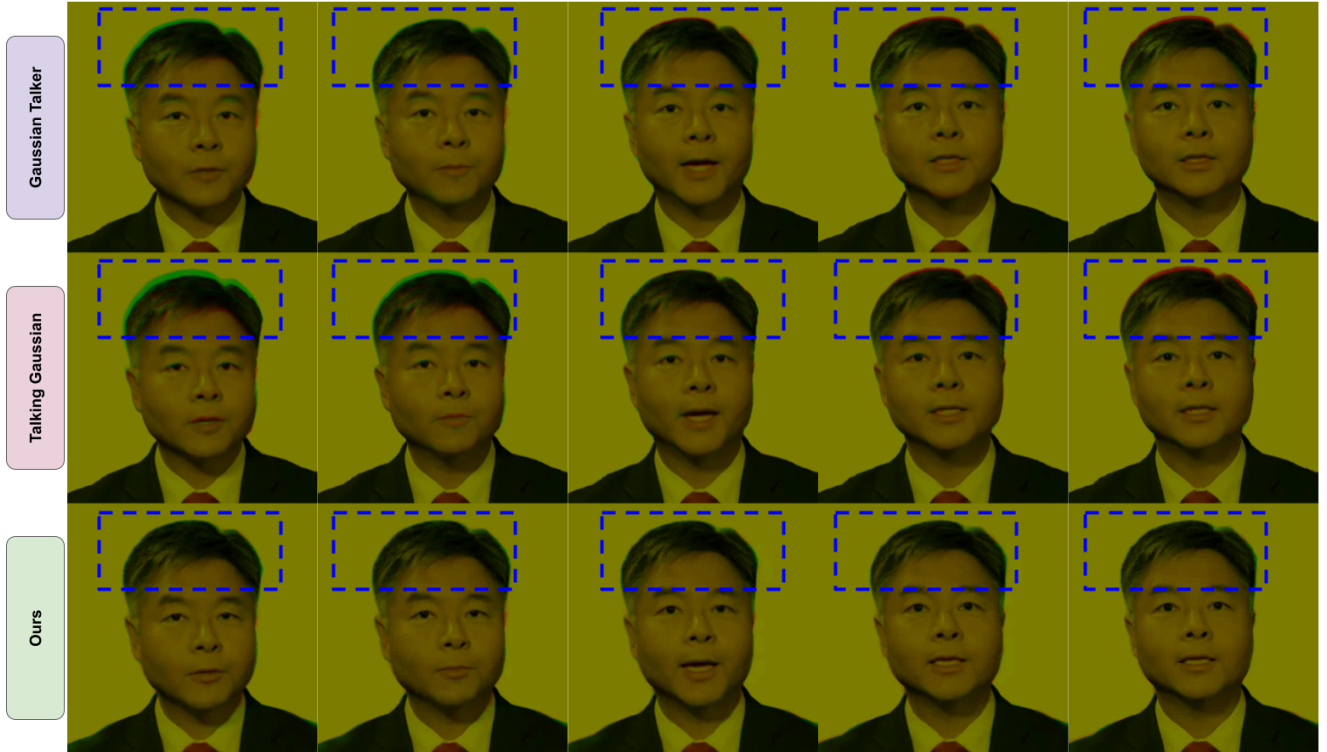


Figure E. Visualization of frame stability through color channel overlay with ground-truth video over 10 consecutive frames. The significant displacement (wobbling) observed in the GaussianTalker and TalkingGaussian methods contrasts with the high overlap and stability achieved by our proposed method. Green and red channels highlight the differences within the blue dashed boxes.

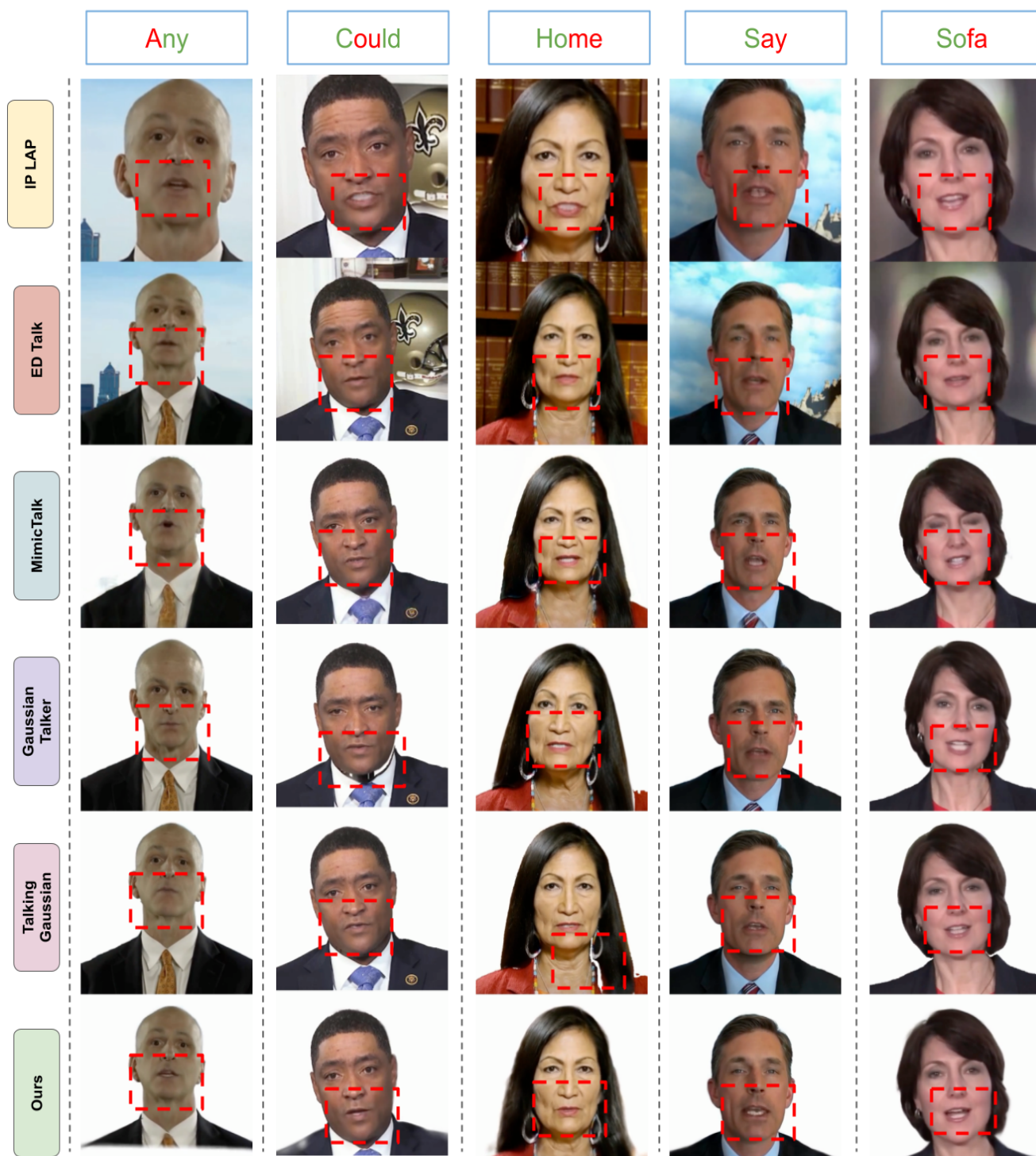


Figure F. Cross-Reenactment Results: We show the visual results by reenacting various methods using a different audio, from a different speaker. The top row shows the word from the audio, with red part highlighting the exact phoneme. GaussianHeadTalk provides the best possible lip movement for these new audio samples. Other methods struggle to have proper lip motion, generate high-quality videos and no artifacts.



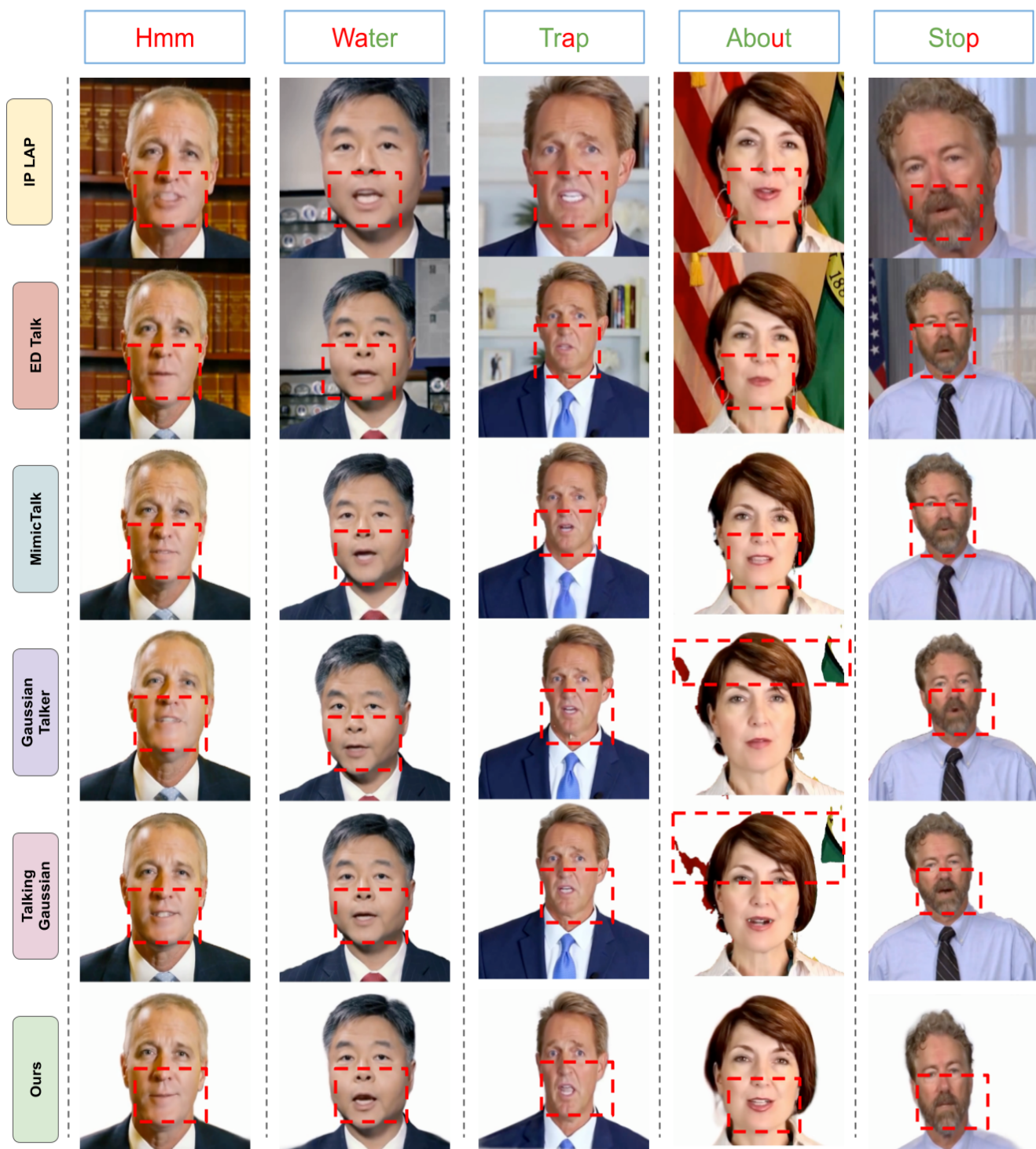


Figure G. Cross-Reenactment Results: We show the visual results by reenacting various methods using a different audio, from a different speaker. The top row shows the word from the audio, with red part highlighting the exact phoneme. GaussianHeadTalk provides the best possible lip movement for these new audio samples. Other methods struggle to have proper lip motion, generate high-quality videos and no artifacts.