

---

# Let Them Talk: Audio-Driven Multi-Person Conversational Video Generation Supplementary Material

---

Anonymous Author(s)

Affiliation

Address

email

## 1 Dataset and Implementation Details

### 1.1 Dataset Details

In this paper, we utilize three distinct testing datasets: the talking head dataset, the talking body dataset, and the dual-human talking body dataset with interactive scenarios. For the talking head and talking body datasets, we employ conventional evaluation techniques for comparison with other methods. However, for the dual-human talking body dataset, where each reference image contains two persons, we evaluate Sync-C, Sync-D, and E-FID by splitting the video into two segments: the left part and the right part. Each segment contains only one person and their corresponding audio. We then average the scores of these two segments to derive the final result for this dataset. Fig.8 showcases some examples of our dual-human dataset.



Figure 8: Some examples of our MTHM dataset.

### 1.2 Sample Details

In all the experiments and evaluations conducted within this paper, we utilize 40 sampling steps. To filter out undesired variations in diffusion models, we employ the following negative prompt during sampling: "bright tones, overexposed, static, blurred details, subtitles, style, works, paintings, images, static, overall gray, worst quality, low quality, JPEG compression residue, ugly, incomplete, extra fingers, poorly drawn hands, poorly drawn faces, deformed, disfigured, misshapen limbs, fused fingers, still picture, messy background, three legs, many people in the background, walking backwards." Additionally, we employ Qwen-VL for reference image captioning.

## 2 Analyses

### 2.1 Full Parameter Training vs Cross-attention Training

We compare full parameter training with fine-tuning only the audio cross-attention layer. Our findings indicate that network training parameters are crucial. When compute resources and data are limited, fully parameterized training can lead not only to degradation in the model’s instruction-following ability, especially for motion and interaction, but also to hand and object distortion. Conversely, training only the audio cross-attention does not result in these issues, and the instruction-following ability of the base model is well preserved. The comparison results between full parameter training and cross-attention training are shown in Fig. 9. It can be seen that full parameter training degrades the model’s instruction-following ability and causes hand distortion.

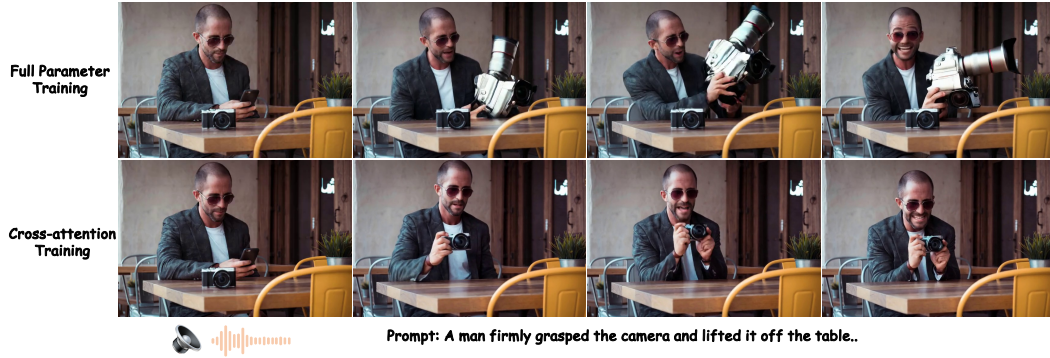


Figure 9: Comparison between full parameter training and cross-attention training.

### 2.2 Long Video Generation

Utilizing the autoregressive-based method facilitates the long video generation of our method. The experimental results for long video generation are shown in Fig.10. This example shows a generated result containing 305 frames.



Figure 10: The generation result of long videos.

## 3 Societal Impacts

This paper introduces an effective tool for audio-driven multi-person conversational video generation to the community. However, there exists a risk wherein malicious entities could exploit this framework to generate fake videos of celebrities, potentially misleading the public. This concern is not unique to our approach but is a shared consideration across various human animation methodologies.