
OmniTalker: One-shot Real-time Text-Driven Talking Audio-Video Generation With Multimodal Style Mimicking

Anonymous Author(s)

Affiliation

Address

email

1 A Model Details

2 Our model includes 22 audio-visual fusion blocks with two parallel branches (4 single-modality DiT
3 blocks each), 512-dim embeddings (audio/visual via linear layers, text via 4 ConvNeXt V2 blocks),
totaling 0.8B parameters. We provide detailed configurations in Table 1

Table 1: Model Configuration

Module	Hyper Parameter	Value
TextEmbedding	ConvNeXt-V2 blocks	4
	Embedding dimension	512
	FFN dimension	1024
AudioEmbedding	Linear Layer	1
	Embedding dimension	1024
VisualEmbedding	Linear Layer	1
	Embedding dimension	1024
Audio-visual Fusion	Transformer blocks	22
	Attention heads	16
	Embedding dimension	1024
	FFN dimension	2048
Audio DiT branch	Transformer blocks	4
	AdaLayerNorm	1
	Linear Layer	1
	Attention heads	16
	Embedding dimension	1024
	FFN dimension	2048
Visual DiT branch	Transformer blocks	4
	AdaLayerNorm	1
	Linear Layer	1
	Attention heads	16
	Embedding dimension	1024
	FFN dimension	2048

4

5 A.1 Preliminaries on Flow Matching

6 Flow matching, evolved from Continuous Normalizing Flows (CNFs)[6], aims to learn a model that
7 transforms a simple distribution p_0 into a more complicated one p_1 . This objective aligns closely
8 with the fundamental goal of diffusion models. The learning process is achieved by minimizing the

9 difference between the flow of the data and the flow predicted by the model, with a simple objective:

$$\mathcal{L}_{FM}(\theta) = \mathbb{E}_{t, p_t(x)} \|v_t(x) - u_t(x)\|^2 \quad (1)$$

10 where x denotes data points, $p_t(x)$ represents a time dependent probability density path, and $u_t(x) =$
 11 $\frac{dx_t}{dt}$ denotes the unknown time-dependent vector field governing the trajectory of the data distribution
 12 from from p_0 to p_1 .

13 Specifically, given a specific sample x_1 from some unknown data distribution $q(x_1)$, $p_t(x|x_1)$ refers
 14 to the conditional probability path. The marginal probability path can be obtained by taking the
 15 expectation of the conditional probability path over all samples in the data distribution:

$$p_t(x) = \int p_t(x|x_1)q(x_1)dx_1 = \mathbb{E}_{q(x_1)}(p_t(x|x_1)) \quad (2)$$

16 Assuming $p_t(x|x_1)$ is derived from the conditional vector field $u_t(x|x_t)$, it follows that $u_t(x)$ and
 17 $u_t(x|x_t)$ satisfy the following relationship:

$$u_t(x) = \int u_t(x|x_1) \frac{p_t(x|x_1)q(x_1)}{p_t(x)} dx_1 = \mathbb{E}_{p(x_1|x)}(u_t(x|x_1)) \quad (3)$$

18 Therefore, the original $\mathcal{L}_{FM}(\theta)$ can be reformulated as:

$$\mathcal{L}_{CFM}(\theta) = \mathbb{E}_{t, q(x_1), p_t(x|x_1)} \|v_t(x; \theta) - u_t(x|x_1)\|^2 \quad (4)$$

19 We consider the case of Gaussian distributions and adopt a simple yet effective approach: the optimal
 20 transport (OT) path. Given the target data point x_1 and the current data point x_t , the most efficient way
 21 is to go with a straight line. The conditional vector field is then defined as $u_t(x|x_1) = (x_1 - x)/(1 - t)$.
 22 In this case, the CFM loss takes the form:

$$\mathcal{L}_{CFM}(\theta) = \mathbb{E}_{t, q(x_1), p(x_0)} \|v_t((1 - t)x_0 + tx_1; \theta) - (x_1 - x_0)\|^2 \quad (5)$$

23 as described in Section 3.3

24 A.2 Render

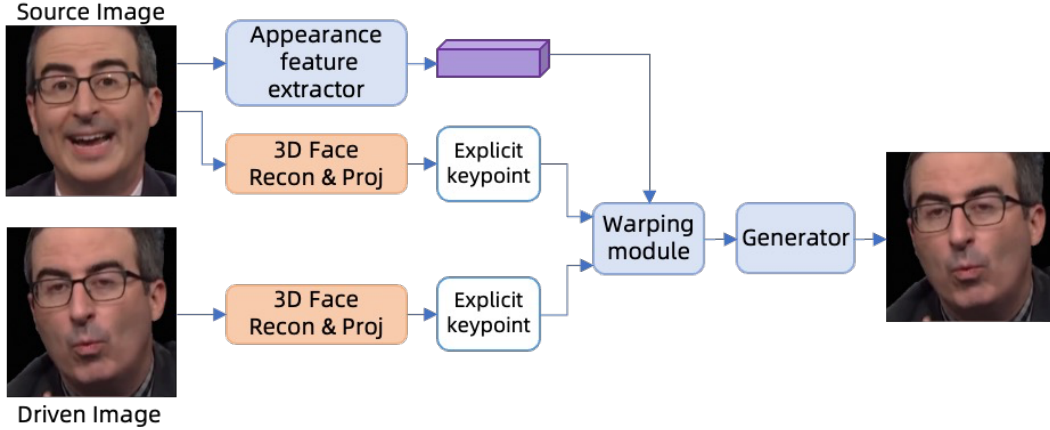


Figure 1: Overview of visual renderer.

25 To balance the quality of generated videos and model inference performance, we employ a warping-
 26 based GAN framework inspired by existing works [12, 4]. The general framework consists of four
 27 key components as shown in Figure 1: 1) *Appearance feature extraction* to get visual features from
 28 source image; 2) *Motion representation extraction* that extracts explicit facial keypoint to represent
 29 face movement; 3) *Warping field estimation* to calculate the transformation from source to target; and
 30 4) *Generator* to synthesize final images from warped appearance features. Specifically, face motion
 31 is represented by 3D keypoints projected from a 3DMM head template [11], and 3D keypoints on
 32 key face regions are projected from the head template, including eyes, mouth, eyebrows and face

contours, driven by 3DMM coefficients, which are capable of representing a diverse range of facial expressions.

We follow LivePortrait [4] using perceptual loss \mathcal{L}_{Per} , GAN loss \mathcal{L}_{GAN} , reconstruction loss \mathcal{L}_{Recon} for render network training. To enhance the quality of mouth region, we obtain the mouth mask and introduce a mouth region perceptual loss $\mathcal{L}_{Per}^{Mouth}$. The overall training objective is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{Per} + \mathcal{L}_{GAN} + \mathcal{L}_{Recon} + \mathcal{L}_{Per}^{Mouth} \quad (6)$$

During inference, the renderer generates the final video by projecting the visual codes predicted by OmniTalker onto 3D keypoints through 3DMM.

B Experiment Details

B.1 Datasets

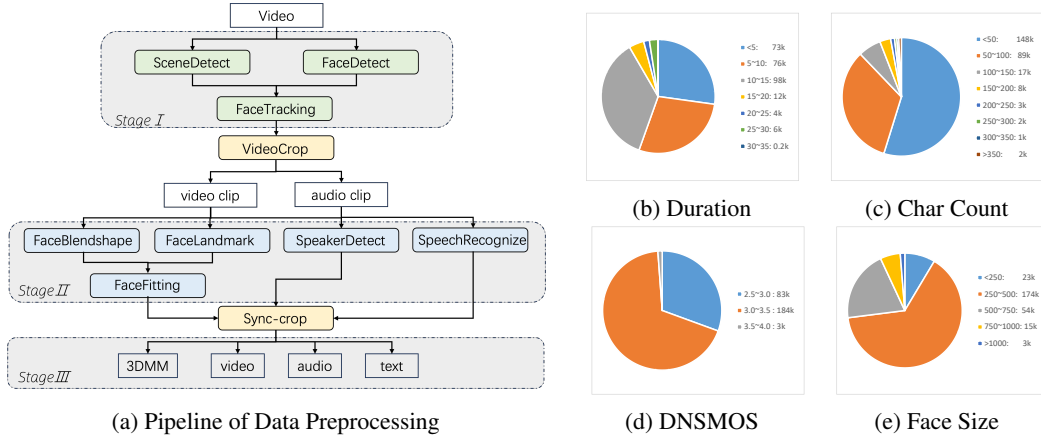


Figure 2: Overview of our proposed pipeline and dataset.

We employ TalkingHead-1KH [13], VoxCeleb [8], and CelebV-HQ [14] as pre-training datasets, while constructing a high-quality custom dataset for fine-tuning. This study proposes an automated preprocessing pipeline for large-scale multimodal data curation, as illustrated in Figure 2a. The pipeline comprises three sequential stages, containing eight specialized processing modules strategically designed to enable parallel execution and optimize computational efficiency through modular architecture.

- Stage 1: Coarse Segmentation
 - Scene Detection: PySceneDetect [1] with adaptive threshold ($\sigma = 1.5$)
 - Face Detection & tracking: Insightface [3] with IoU continuity > 0.5
- Stage 2: Multimodal Feature Extraction
 - Facial motion: FaceVerse [11] (52 blendshapes + 6DOF pose + 4 eye gaze)
 - Speaker Detection: LightASD [5] (confidence > 0.8)
 - Speech Recognition: Whisper-V3-Large [9]
- Stage 3: Fine Segmentation
 - Temporal constraints: Clip duration: $1s \leq t \leq 30s$ & Phoneme rate: $< 1s/\text{character}$
 - Spatial constraints: Face bounding box $> 15\%$ frame area
 - Others: Audio quality: DNSMOS P.835 OVRL [10] (cutoff > 2.5)

Figures 2b to 2e shows the distributions across duration, text length, quality score, and average face size in custom dataset. It comprises approximately 300,000 clips, totaling 690 hours of high-quality multimodal data with synchronized text, audio, and video components.

63 B.2 Training

64 **Strategy.** We begin by performing large-scale pre-training on open-source multimodal datasets
 65 as described in Appendix B.1 to establish foundational capabilities in text comprehension and
 66 multimodal generation. During the training process, we implement a masking strategy that randomly
 67 masks 70-100% of the audio-visual sequences, compelling the model to learn sequence reconstruction
 68 abilities. Concurrently applying random dropout with a 0.2 probability across text, audio, and video
 69 conditioning inputs facilitates classifier-free guidance (CFG) training. In the final stage, we conduct
 70 fine-tuning on custom dataset, which significantly enhances the model’s performance.

71 **Sampling.** We modified the timestep sampling distribution from a uniform distribution to a logit-
 72 norm distribution (mean=0.0, std=1.0) according to [2], and observed that employing logit-norm
 73 bias to prioritize the selection of training timesteps significantly enhances model performance than
 74 uniform distributions.

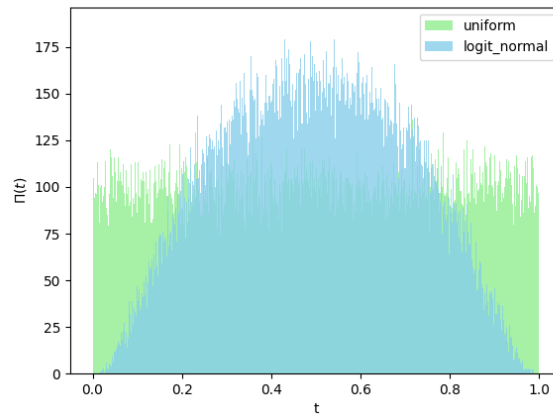


Figure 3: logit-norm and uniform distributions that bias the sampling of training timesteps.

75 B.3 More Qualitative Evaluation

76 **Eye Movements.** Figures 4 and 5 demonstrate the enhanced performance of our proposed method
 77 in generating eye movements. As indicated by the highlighted red regions in the visualizations, our
 78 approach produces complex gaze trajectories with dynamic variations that exhibit significant temporal
 79 correlations with head pose changes. By contrast, the baseline method maintains a fixed gaze direction
 80 determined at initialization, resulting in nearly static eye movements. This coordinated gaze-head
 81 motion mechanism plays a critical role in enhancing the biological plausibility of generated videos.
 82 Notably, Figure 6 further validates our method’s capability to inherit eye movement characteristics
 83 from reference videos. Experimental results show that the generated gaze trajectories demonstrate
 84 heritable features in both frequency and amplitude parameters compared to reference videos, while
 85 maintaining necessary difference to avoid mechanical repetition while preserving consistency.

86 **Emotional Generation.** As illustrated in Figure 7, we employ reference videos from the
 87 RAVDESS[7] emotional dataset. Leveraging in-context learning capabilities, our methods enables to
 88 generate results highly aligned with target emotions. Experimental results demonstrate that the pro-
 89 posed method accurately captures expressive facial micro-expressions. This capability of generating
 90 multimodal emotional representations effectively validates the model’s capacity for analyzing and
 91 reconstructing complex emotional states.

92 B.4 User Study Settings

93 We conducted user studies for both audio and video generation following our quantitative evaluation.
 94 We selected 25 reference videos each from the VoxCeleb2 and Custom datasets, totaling 50 references.

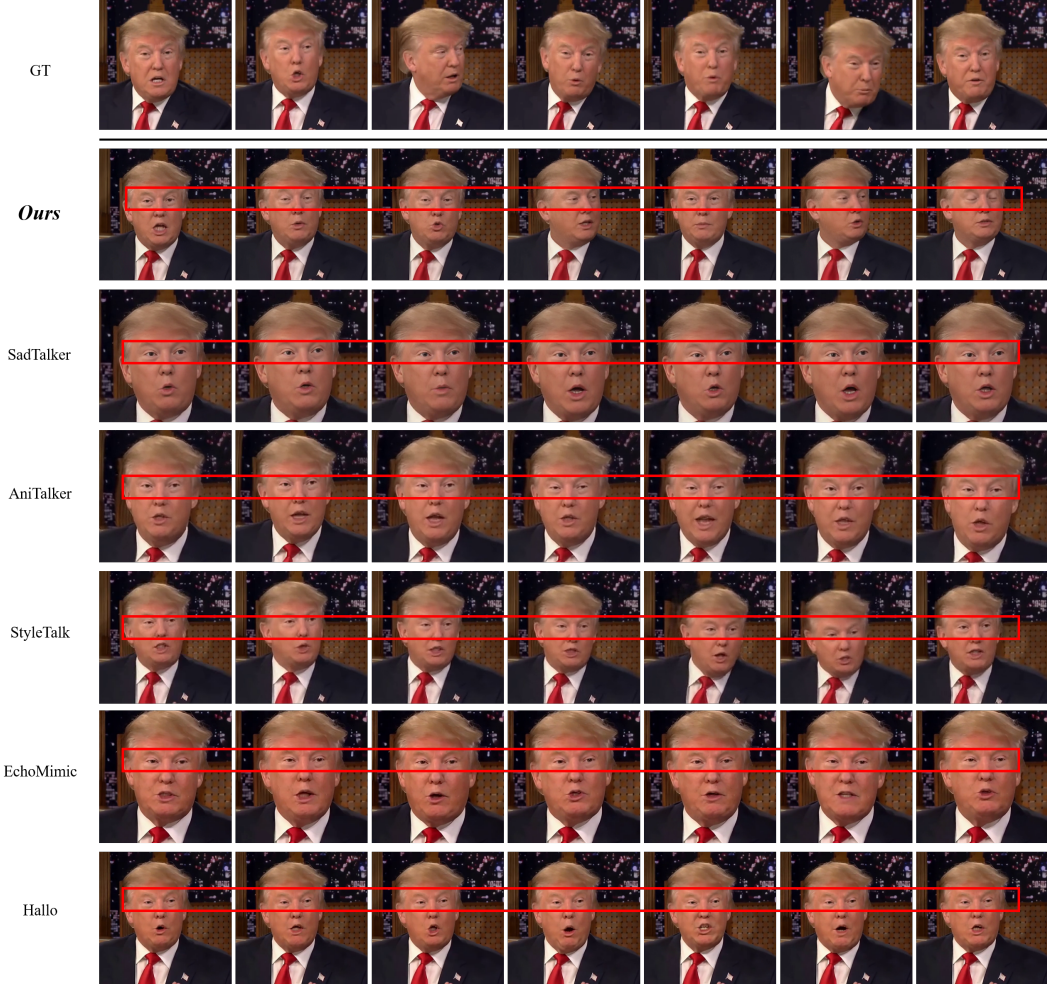


Figure 4: Comparative Analysis of Eye Movement Visualization on ID1.

For the audio evaluation, we generated 50 audio samples per TTS method using the provided text, then employed our synthesized audios to generate corresponding videos for each THG method. Each participant was asked to evaluate the results across six dimensions: for audio outputs, speech similarity (assessing vocal timbre resemblance) and speech rhythm (evaluating prosody alignment with text punctuation); for video outputs, visual quality (focusing on image clarity, artifacts, and geometric distortion), speaking style (comparing emotional expression, head motion, and gaze patterns), lip-sync, and pose-sync (verifying semantic accuracy of head gestures like nodding for affirmation). Figure 8 displays the interface used in our user study. 50 participants rated each generation result on a 5-point scale (1-5), with average scores presented in Table 3. This comprehensive evaluation framework ensures multi-dimensional assessment of both audio-visual quality and semantic fidelity in human-like generation tasks.

C Limitations and Future Work

Here we discuss the limitations of this work and potential directions for future improvement. Firstly, the scope of our method focuses primarily on dynamic generation of the head region without incorporating hand pose modeling or full-body motion control. This constraint limits the completeness and interactivity of generated content to some extent. The second challenge relates to generation quality: GAN-based approaches still face technical barriers when handling large-scale dynamic transformations. Notably, when motion amplitude exceeds critical thresholds, artifacts such as

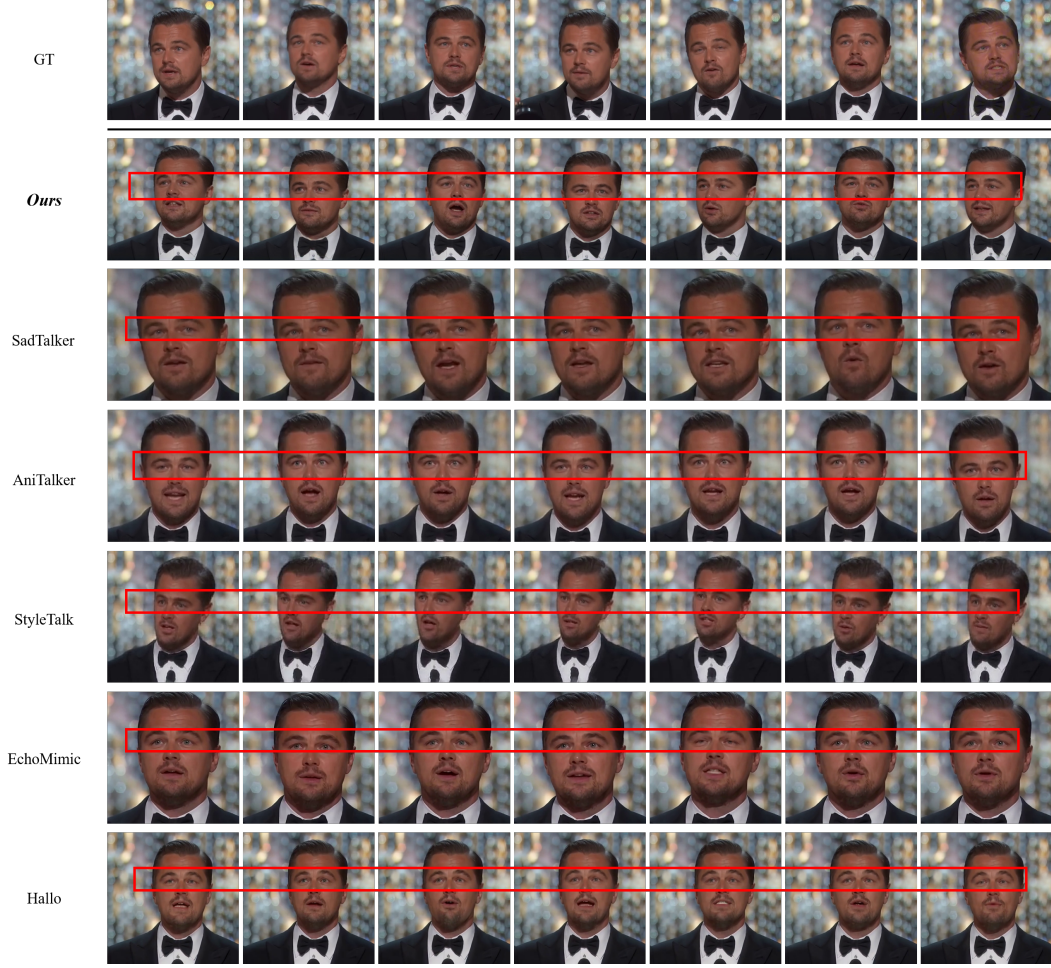


Figure 5: Comparative Analysis of Eye Movement Visualization on ID2.

113 texture blurring and boundary discontinuities frequently emerge, compromising the visual realism of
 114 generated results.

115 To address these limitations, we propose optimizations along two dimensions. First, in the motion
 116 control dimension, we can implement a coordinated generation framework by integrating hand skeletal
 117 tracking data with full-body pose estimation information. Second, for generation quality enhancement,
 118 we recommend adopting a video diffusion model-based generation paradigm. Compared to traditional
 119 GAN architectures, video diffusion models demonstrate superior spatiotemporal continuity through
 120 their progressive denoising process. The progressive sampling mechanism effectively mitigates
 121 quality degradation caused by large-scale movements. Furthermore, combining attention mechanisms
 122 with spatiotemporal feature fusion strategies promises to significantly enhance both detail fidelity and
 123 dynamic smoothness in generated outputs.

124 D Broader Impacts

125 Our work focuses on the efficient generation of realistic and expressive digital human videos, aiming
 126 to drive technological advancement with positive societal impact. By leveraging automation, real-time
 127 interaction, and multimodal perception capabilities, this technology enhances productivity across
 128 industries and accelerates innovation in the integration of text, audio, and video processing systems.
 129 It also fosters the emergence of new industries such as virtual idols and digital avatars, while enabling
 130 cross-lingual content generation that promotes cultural exchange across regions. However, the same
 131 technology carries risks: hyper-realistic deepfake algorithms could be exploited for disinformation

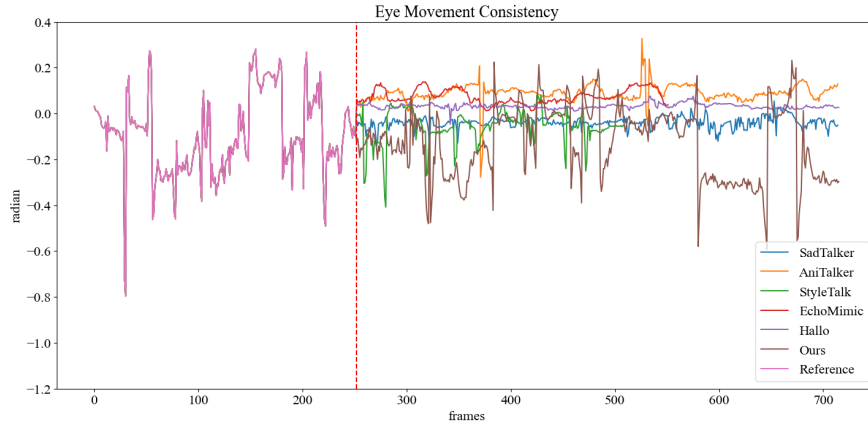


Figure 6: Visualization of eye movements over time for the reference video and videos generated by different methods. The red dashed line separates the values of the reference video on the left from those of the generated results on the right.

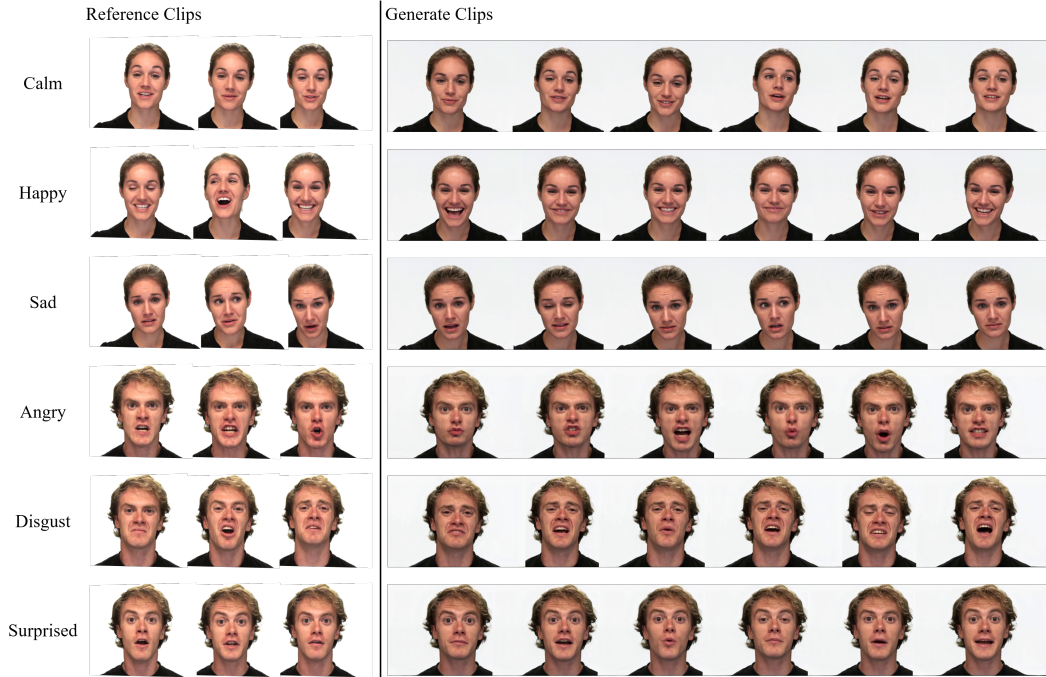


Figure 7: Visualization of Generation via different emotional reference.

132 campaigns, identity fraud, or synthetic media manipulation; sensitive biometric data collection
 133 required for digital humans may lead to large-scale privacy breaches if not properly secured; and
 134 prolonged engagement with virtual companions or personalized avatars might inadvertently create
 135 emotional dependencies or psychological impacts. To address these challenges,

- 136 • We incorporate visible watermarks into generated videos to proactively alert users that the
 137 content is synthetic in nature.
- 138 • We embed imperceptible digital watermarks within both video and audio streams to enable
 139 traceability and track the origin of generated content, thereby requiring creators to consider
 140 potential legal and ethical risks associated with synthetic media production.

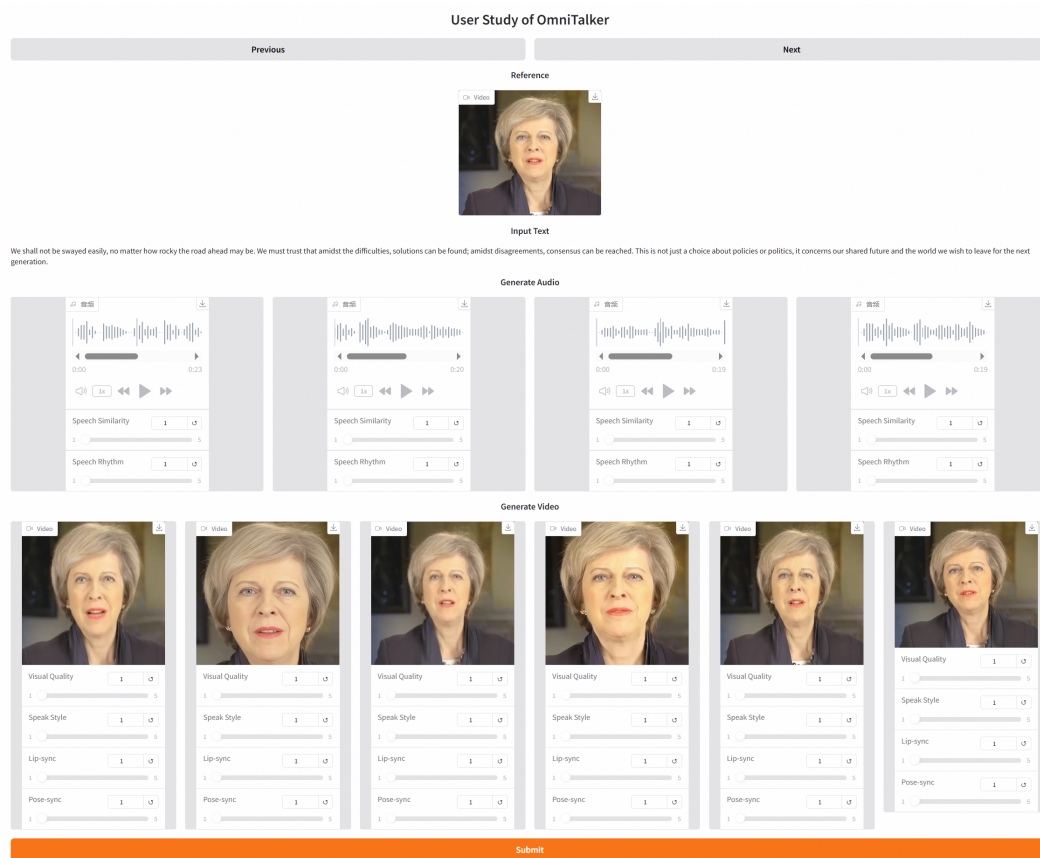


Figure 8: The interface of user study.

- We are committed to advancing our methodology to enhance deepfake detection techniques, aiming to improve the accuracy and reliability of automated detection systems through continuous algorithmic refinement.

References

- [1] PySceneDetect Developers. Pyscenedetect. <https://www.scenedetect.com/>, 2025.
- [2] Patrick Esser, Sumith Kulal, A. Blattmann, Rahim Entezari, Jonas Muller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. *ArXiv*, abs/2403.03206, 2024.
- [3] Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection, 2021.
- [4] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint:2407.03168*, 2024.
- [5] Junhua Liao, Haihan Duan, Kanghui Feng, Wanbing Zhao, Yanbing Yang, and Liangyin Chen. A light weight model for active speaker detection. In *CVPR*, pages 22932–22941, 2023.
- [6] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.
- [7] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.
- [8] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.
- [9] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *ICML*, pages 28492–28518. PMLR, 2023.
- [10] Chandan KA Reddy, Vishak Gopal, and Ross Cutler. Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP*, pages 886–890. IEEE, 2022.
- [11] Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *CVPR*, pages 20301–20310, 2022.
- [12] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, pages 10039–10049, 2021.
- [13] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, 2021.
- [14] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. CelebV-HQ: A large-scale video facial attributes dataset. In *ECCV*, 2022.