OmniHuman-1: Rethinking the Scaling-Up of One-Stage Conditioned Human Animation Models

Gaojie Lin*   Jianwen Jiang*†   Jiaqi Yang*   Zerong Zheng*   Chao Liang

ByteDance

https://omnihuman-lab.github.io/

Figure 1. **The video frames generated by OmniHuman based on input audio and image.** The generated results feature head and gesture movements, as well as facial expressions, that match the audio. OmniHuman generates highly realistic videos with any aspect ratio and body proportion, and significantly improves gesture generation and object interaction over existing methods, due to the data scaling up enabled by omni-conditions training.

Abstract

End-to-end human animation, such as audio-driven talking human generation, has undergone notable advancements in the recent few years. However, existing methods still struggle to scale up as large general video generation models, limiting their potential in real applications. In this paper, we propose OmniHuman, a Diffusion Transformer-based framework that scales up data by mixing motion-related conditions into the training phase. To this end, we introduce two training principles for these mixed conditions, along with the corresponding model architecture and inference strategy. These designs enable OmniHuman to fully leverage data-driven motion generation, ultimately achieving highly realistic human video generation. More importantly, OmniHuman supports various portrait contents (face close-up, portrait, half-body, full-body), supports both talking and singing, handles human-object interactions and challenging body poses,

---
*Equal contributions
†Project lead

and accommodates different image styles. Compared to existing end-to-end audio-driven methods, OmniHuman not only produces more realistic videos, but also offers greater flexibility in inputs. It also supports multiple driving modalities (audio-driven, video-driven and combined driving signals). Video samples are provided on the `project page`.

1. Introduction

Since the emergence of the Diffusion Transformer-based (DiT) video diffusion models, the field of general video generation, including Text-to-Video and Image-to-Video [3–6, 16, 17, 22, 33, 35, 49, 60, 63, 66, 82] has made significant progress in producing highly realistic video content. A key factor driving this advancement is the large-scale training data, typically formatted as video-text pairs. Expanding the training dataset enables DiT networks to learn motion priors for various objects and scenes, resulting in strong generalization capabilities during inference.

Building upon these pretrained video diffusion networks, end-to-end human animation models, either for pose-driven human animation or audio-driven talking human generation, have developed rapidly since last year [8, 18, 26, 34, 52, 54, 62, 70, 71]. Despite achieving realistic results, these models are trained on highly filtered datasets to simplify the learning process, restricting their applicability to limited scenarios. For instance, most existing end-to-end audio-conditioned models are limited to facial or portrait animation, while most pose-conditioned models can only handle full-body images captured from a front-facing perspective with a static background. To date, no prior work has attempted to scale up training data for more generalizable human animation.

Scaling up human animation data may seem straightforward, but unfortunately it is not. Directly adding more data is not always beneficial for network training. Take audio-conditioned models as an example: audio is primarily associated with facial expressions and has little correlation with body poses, background motion, camera movement, or lighting changes. As a result, raw training data must be filtered and cropped to minimize the influence of these unrelated factors. Additionally, audio-conditioned models often undergo further data cleaning based on lip-sync accuracy, which is also important to stabilize training. Similarly, pose-conditioned models require extensive filtering, cropping, and cleaning. Unfortunately, these processes discard a substantial amount of data, making dataset scaling a futile effort, despite the fact that much of the discarded data contains valuable motion patterns essential for training data expansion.

In this paper, we address the challenges of scaling up human animation data and models. Our key insight is that incorporating multiple conditioning signals, such as text, audio, and pose, during training can significantly reduce data wastage. This approach offers two main advantages. On one hand, data that would otherwise be discarded for single-condition models (e.g., audio- or pose-conditioned) can be leveraged in tasks with weaker or more general conditions, such as text conditioning. Training on such data allows the model to learn more diverse motion patterns, mitigating the limitations imposed by data filtering. On the other hand, different conditioning signals can complement each other. For example, while audio alone cannot precisely control body poses, stronger conditions such as pose inputs can provide additional guidance. By integrating stronger conditioning signals alongside audio data during training, we aim to reduce overfitting and improve the generalization of generated results.

Based on the above considerations, we designed the omni-conditions training strategy, which follows two proposed training principles: (1) stronger conditioned tasks can leverage weaker conditioned tasks and their corresponding data to achieve data scaling up during the model training process, and (2) the stronger the condition, the lower the training ratio that should be used. To implement this strategy, we built a mixed conditioned human video generation model named OmniHuman, based on the advanced video generation model architecture, DiT [14, 42]. OmniHuman can train with three motion-related conditions (text, audio, and pose) from weak to strong. This approach addresses the data scaling up challenge in end-to-end frameworks, allowing the model to benefit from large-scale data training, learn natural motion patterns, and support various input forms.

Overall, our contributions can be summarized as follows:
1. We propose the OmniHuman model, a mixed-conditioned human video generation model. It leverages our omni-conditions training strategy to integrate various motion-related conditions and their corresponding data. Unlike existing methods that reduce data due to stringent filtering, our approach benefits from large-scale mixed conditioned data.
2. OmniHuman generates highly realistic and vivid human motion videos, supporting multiple modalities simultaneously. It performs well with different portrait and input aspect ratios. OmniHuman significantly improves gesture generation, a challenge for previous end-to-end models, and supports various image styles, significantly outperforming existing audio-conditioned human video generation methods.

2. Related Works

2.1. Video Generation

In recent years, the advent of technologies such as diffusion models [21, 29, 38, 50, 51] has propelled the capabilities of generative models to a practically usable level. The latest advancements in image generation [7, 14] produce results

that are almost indistinguishable from reality. Consequently, a growing number of studies [24, 31, 43, 57, 73, 76, 82] are shifting their focus toward the field of video generation. Early text-to-video works primarily centered on training-free adaptations of pre-trained text-to-image models [44, 49, 68] or integrated temporal layers with fine-tuning on limited video datasets [16, 63, 82]. However, due to the lack of extensive data, the video generation quality of these methods often remains unsatisfactory. To better exploit scaling laws and push the boundaries of video generation models, recent works [31, 43, 57, 73] have optimized in three major areas. First, they have collected larger-scale, high-quality video datasets, with the data volume increasing to (O(100M)) clips of high-resolution videos. Second, they employ 3D Causal VAE [75] to compress both spatial and temporal features of video data, thereby enhancing video modeling efficiency. Third, the foundational model structure has transitioned from UNet to Transformer, improving the model's scalability. Additionally, these works utilize meticulously designed progressive training recipes and datasets to maximize the model's potential. For example, [31, 43] first pre-train on a large volume of low-resolution images and videos, leveraging data diversity to enhance the model's generalization capabilities. They then perform fine-tuning on a subset of high-resolution, high-quality data to improve the visual quality of generated videos. Large-scale data has significantly improved the effectiveness of general video generation. However, progress in the field of human animation synthesis remains relatively slow.

2.2. Human Animation

As an important task of video generation, Human Animation synthesizes human videos using human images and driving conditions such as audios or videos. Early GAN-based methods [27, 47, 48, 65, 79] typically employ small datasets [40, 47, 69, 83] consisting of tens of thousands of videos to achieve video-driven in a self-supervised manner. With the advancement of Diffusion models, several related works [25, 46, 64, 78, 85] have surpassed GAN-based methods in performance while using datasets of similar scale. Instead of using pixel-level videos, these methods employ 2D skeleton, 3D depth, or 3D mesh sequences as driving conditions. Audio-driven methods used to focus on portrait [11, 15, 26, 56, 74, 77, 81]. Despite some efforts [10, 23, 34, 39, 55] to extend the frame to the full body, there are still challanges especially in hand quality. To bypass it, most approaches [10, 23, 39, 55] adopt a two-stage hybrid driving strategy, utilizing gesture sequences as a strong condition to assist hand generation. CyberHost [34] attempts to achieve one-stage audio-driven talking body generation through codebook design. Most notably, existing Human Animation methods typically focus on limited-scale datasets and limited-complexity structure, generally less than

a thousand hours and 2B. Although FADA [81] employs a semi-supervised data strategy to utilize 1.4K hours of portrait videos, VLogger [10] meticulously collects 2.2K hours of half-body videos, and Hallo3 [11] initializes its weights derived from CogVideoX5B-I2V [72], their performance does not exhibit the scaling law trends observed in other tasks such as LLMs [41, 58], VLMs [2, 37], and T2I/T2V [13, 30, 32]. Scaling effects in Human Animation haven't been investigated effectively yet.

3. Method

In this section, we introduce our framework, OmniHuman, which employs motion-related condition mixing during network training to scale up the training data. First, we provide an overview of the framework, including its inputs, outputs and key design elements. Next, we focus on the omni-conditions design, covering audio, pose, and reference conditions. We then detail the training strategy of OmniHuman, which leverages these omni-conditions for mixed data training, enabling the model to learn natural motion from large-scale datasets. Finally, we describe the implementation details for the inference phases of the OmniHuman model.

3.1. Overview

As illustrated in Figure 2, our approach consists of two primary parts: the OmniHuman model, a multi-condition diffusion model and the Omni-Conditions Training Strategy. For model, The OmniHuman model begins with a pretrained Seaweed model [35], which uses MMDiT [14, 42] and is initially trained on general text-video pairs for text-to-video and text-to-image tasks. Given a reference image, the OmniHuman model aims to generate human videos using one or more driving signals including text, audio and pose. To achieve this, we employ various strategies to integrate frame-level audio features and pose heatmap features into the OmniHuman model. The detailed procedure is explained in the following subsections. OmniHuman model utilizes a causal 3DVAE [80] to project videos at their native size [12] into a latent space and employs flow matching [36] as the training objective to learn the video denoising process. We employ a three-stage mixed condition post-training approach to progressively transform the diffusion model from a general text-to-video model to a multi-condition human video generation model. As depicted on the left of Figure 2, these stages sequentially introduce the driving modalities of text, audio, and pose according to their motion correlation strength, from weak to strong, and balance their training ratios.

3.2. Omni-Conditions Designs

Driving Conditions. We adopted different approaches for injecting audio and pose conditions. Regarding audio condition, the wav2vec [1, 45] model is employed to extract acoustic features, which are subsequently compressed using
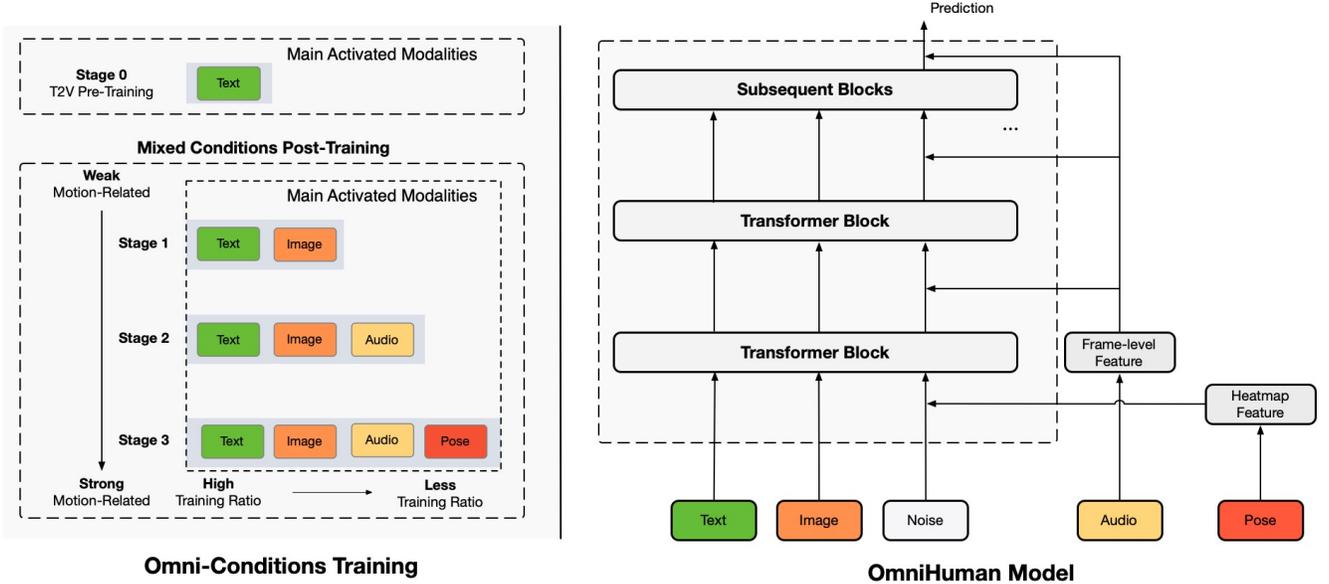
Figure 2. **The framework of OmniHuman.** It consists of two parts: (1) the OmniHuman model, which is based on the DiT architecture and supports simultaneous conditioning with multiple modalities including text, image, audio, and pose; (2) the omni-conditions training strategy, which employs progressive, multi-stage training based on the motion-related extent of the conditions. The mixed condition training allows the OmniHuman model to benefit from the scaling up of mixed data.

a MLP to align with the hidden size of MMDiT. The features of each frame are concatenated with the audio features from adjacent timestamps to generate audio tokens for the current frame. As depicted in Figure 2, these audio tokens are injected into each block of MMDiT through cross-attention, enabling interaction between the audio tokens and the noisy latent representations. To incorporate pose condition, we use a pose guider to encode the driving pose heatmap sequence. The resulting pose features are concatenated with those of adjacent frames to acquire pose tokens. These pose tokens are then stacked with the noise latent along the channel dimension and fed into the unified multi-condition diffusion model for visual alignment and dynamic modeling. The text condition is retained as in the MMDiT text branch.

Appearance Conditions. The goal of OmniHuman is to generate video outputs that preserve both the subject's identity and the background details from a reference image. To achieve this, previous research has proposed various strategies for injecting appearance representations into the denoising process. The most widely adopted approach involves using a reference network [26, 34, 54], a parallel, trainable copy of the entire diffusion UNet or DiT that integrates with the self-attention layers of the original denoising Net. While effective at transferring appearance features to the denoising process, this method requires duplicating a full set of trainable parameters, which presents scalability challenges as model size increases. To overcome this challenge, OmniHuman introduces a simple yet effective strategy for reference conditioning. Instead of constructing additional network modules, we reuse the original denoising DiT backbone to encode the reference image. Specifically, the reference image is first encoded into a latent representation using a VAE, and both the reference and noisy video latents are flattened into token sequences. These sequences are then packed together and simultaneously fed into the DiT, enabling the reference and video tokens to interact via self-attention across the entire network. To help the network distinguish between reference and video tokens, we modify the 3D Rotational Position Embeddings (RoPE) [53] in the DiT by zeroing the temporal component for reference tokens, while leaving the RoPE for video tokens unchanged. This approach effectively incorporates appearance conditioning without adding extra parameters. In addition to the reference image, to support long video generation, we draw on previous methods by using motion frames [52], concatenating their features with the noise features.

After introducing these conditions, the motion-related conditions now include text, reference image, audio, and pose. Text describes the current event, the reference image defines the range of motion, audio determines the rhythm of co-speech gestures, and pose specifies the exact motion. Their correlation strength with human motions can be considered to decrease in this order.

4

3.3. Scaling up with Omni-Conditions Training

Thanks to the multi-condition design, we can divide the model training into multiple tasks, including image and text to video, image and text, audio to video, and image and text, audio, pose to video. During training, different modalities are activated for different data, allowing a broader range of data to participate in the training process and enhancing the model's generation capabilities. After the conventional text-to-video pretraining phase, we follow two training principles for scaling up the conditioned human video generation task. **Principle 1**, stronger conditioned tasks can leverage weaker conditioned tasks and their corresponding data to achieve data scaling up during the model training process. Data excluded from audio and pose conditioned tasks due to filtering criteria like lip-sync accuracy, pose visibility, and stability can be used in text and image conditioned tasks, as they meet the standards for weaker conditions. Therefore, in the first stage 1, we drop the audio and pose conditions. **Principle 2**, the stronger the condition, the lower the training ratio that should be used. During training, stronger motion-related conditions, such as pose, generally train better than weaker conditions like audio due to less ambiguity. When both conditions are present, the model tends to rely on the stronger condition for motion generation, preventing the weaker condition from learning effectively. Therefore, we ensure that weaker conditions have a higher training ratio than stronger conditions. We construct stage 2 to drop only the pose condition, and in the final stage 3, use all conditions. Additionally, the training ratios for text, reference, audio, and pose are progressively halved. This approach assigns higher gradient weights to more challenging tasks and prevents overfitting to a single condition during overlapping condition training. Principle 1 allows us to significantly expand the training data, while Principle 2 ensures that the model fully utilizes the advantages of each motion-related condition during mixed conditions training and learns their motion generation capabilities. By combining Principles 1 and 2, OmniHuman can effectively train with mixed conditioned data, benefiting from data scaling up and achieving satisfactory results.

3.4. Inference Strategies

For audio-driven scenarios, all conditions except pose are activated. For pose-related combinations, all conditions are activated, but for pose-only driving, audio is disabled. Generally, when a condition is activated, all conditions with a lower motion-related influence are also activated unless unnecessary. During inference, to balance expressiveness and computational efficiency, we apply classifier-free guidance (CFG) [20] specifically to audio and text across multiple conditions. However, we observed that an increased CFG results in pronounced wrinkles on the characters, whereas a decreased CFG compromises lip synchronization and motion expressiveness. To mitigate these issues, we propose a CFG annealing strategy that progressively reduces the CFG magnitude throughout the inference process, thereby significantly minimizing the appearance of wrinkles while ensuring that expressiveness. OmniHuman is capable of producing video segments of arbitrary length within memory constraints based on the provided reference images and various driving signals. To ensure temporal coherence and identity consistency in long videos, the last five frames of the previous segment are utilized as motion frames.

4. Experiments

4.1. Implementation Details

Dataset. By filtering based on aesthetics, image quality, motion amplitude, etc. (common criteria for video generation), we obtained 18.7K hours of human-related data for training. Of this, 13% was selected using lipsync and pose visibility criteria, enabling audio and pose modalities. During training, the data composition was adjusted to fit the omni-condition training strategy. For testing, we conduct the evaluation following the portrait animation method Loopy [26] and the half-body animation method CyberHost [34]. We randomly sampled 100 videos from public portrait datasets, including CelebV-HQ [83] (a diverse dataset with mixed scenes) and RAVDESS [28] (an indoor dataset including speech and song) as the testset for portrait animation. For half-body animation, we used CyberHost's test set, which includes a total of 269 body videos with 119 identities, encompassing different races, ages, genders, and initial poses.

Baselines. To comprehensively evaluate OmniHuman's performance in different scenarios, we compare against portrait animation baselines including Sadtalker [77], Hallo [70], Vexpress [62], EchoMimic [8], Loopy [26], Hallo-3 [11], and body animation baselines including DiffTED [23], DiffGest [84] + Mimiction [78], CyberHost [34].

Metrics. For visual quality, FID [19] and FVD [59] are used to evaluate the distance between the generated and labeled images and videos. We also leverage q-align [67], a VLM to evaluate the no-reference IQA(image quality) and ASE(aesthetics). For lip synchronism, we employ the widely-used Sync-C [9] to calculate the confidence between visual and audio content. Besides, HKC (hand keypoint confidence) [34] and HKV (hand keypoint variance) [34] are employed, to represent hand quality and motion richness respectively.

4.2. Comparisons with Existing Methods

As shown in the Table 1 and 2, overall, OmniHuman demonstrates superior performance compared to leading specialized models in both portrait and body animation tasks using a single model. For audio-driven animation, the generated results cannot be identical to the original video, especially when the reference image contains only a head. The model's

Table 1. **Quantitative comparisons with audio-conditioned portrait animation baselines.**

Methods	CelebV-HQ					RAVDESS				
	IQA ↑	ASE↑	Sync-C↑	FID↓	FVD↓	IQA ↑	ASE↑	Sync-C↑	FID↓	FVD↓
SadTalker [77]	2.953	1.812	3.843	36.648	171.848	3.840	2.277	4.304	32.343	22.516
Hallo [70]	3.505	2.262	4.130	35.961	53.992	4.393	2.688	4.062	19.826	38.471
VExpress [61]	2.946	1.901	3.547	65.098	117.868	3.690	2.331	5.001	26.736	62.388
EchoMimic [8]	3.307	2.128	3.136	35.373	54.715	4.504	2.742	3.292	21.058	54.115
Loopy [26]	3.780	2.492	4.849	33.204	49.153	4.506	2.658	4.814	17.017	16.134
Hallo-3 [11]	3.451	2.257	3.933	38.481	42.125	4.006	2.462	4.448	28.840	26.029
OmniHuman	**3.875**	**2.656**	**5.199**	**31.435**	<u>46.393</u>	**4.564**	**2.815**	**5.255**	**16.970**	**15.906**

Table 2. **Quantitative comparisons with audio-conditioned body animation baselines.**

Methods	IQA ↑	ASE↑	Sync-C↑	FID↓	FVD↓	HKV ↑	HKC↑
DiffTED [23]	2.701	1.703	0.926	95.455	58.871	-	0.769
DiffGest. [84]+MomicMo. [78]	4.041	2.897	0.496	58.953	66.785	23.409	0.833
CyberHost [34]	3.990	2.884	6.627	32.972	28.003	24.733	0.884
OmniHuman	**4.142**	**3.024**	**7.443**	**31.641**	**27.031**	**47.561**	**0.898**

Table 3. **Subjective comparison of different training ratios for audio conditions.**

Methods	Identity Consistency	Lip-sync Accuracy	Visual Quality	Action Diversity	Overall
10% Audio Training Ratio	28.84	11.59	21.59	11.59	11.59
50% Audio Training Ratio	**50.87**	**53.62**	**44.93**	**40.58**	**69.57**
100% Audio Training Ratio	11.59	30.43	13.04	36.23	17.93

varying preferences for motion styles across different scenarios complicate performance measurement using a single metric. By averaging the metrics across the dataset, OmniHuman achieves the best results across all evaluated metrics, reflecting its overall effectiveness. Additionally, OmniHuman excels across almost all metrics in specific datasets. Notably, existing methods use a single model for specific body proportions (portrait, half-body) with fixed input sizes and ratios. In contrast, OmniHuman supports various input sizes, ratios and body proportions with a single model, achieving satisfactory results. This advantage stems from its omni-conditions training, which learns from a large scale of diverse content and varying sizes during mixed data training.

4.3. Ablation Studies on Omni-Conditions Training

Here, we primarily analyze and explain principles 1 and 2 of the omni-condition training in OmniHuman. For the first principle, we compare training using only data that meets the requirements for audio and pose animation (i.e., 100% audio training ratio) with training data for weaker conditions (i.e., text). Our experimental results demonstrate that the ratio of these two data parts significantly affects the final performance. From the visualizations in Figure 3, it is evident that a high proportion of audio condition-specific data training reduces dynamic range and can cause failures with complex input images. Including weaker condition data at a 50% ratio

yields satisfactory results (e.g., accurate lip-syncing and natural motion). However, excessive weaker condition data can hinder training, resulting in poorer correlation with the audio. We also conducted a subjective evaluation to determine the optimal mix of these two data types during training. Specifically, we conducted a blind evaluation with 20 subjects who compared the samples across various dimensions to select the most satisfactory one, with an option for abstention. In total, 50 samples depicting diverse scenarios were evaluated. The results in Table 3 were consistent with the conclusions drawn from the visualizations.

The second principle can also be simultaneously validated with the principle 1 experiment, but we additionally conduct another experiment using different ratios of pose conditions to study the effects of pose condition ratios. Visual comparisons are presented in Figure 4 and 5. When the model is trained with a low pose condition ratio and tested with only audio conditions, the model tends to generate intense, frequent co-speech gestures, as is proven by the motion blur effects in the top row of Figure 5 and the incorrect fingers in the top row of Figure 4. On the other hand, if we train the model with a high pose ratio, the model tends to rely on the pose condition to determine the human poses in the generated video. Consequently, given the input audio as the only driving signal, the generated results typically maintain a similar pose, as shown in the bottom rows of Figure 4 and 5.
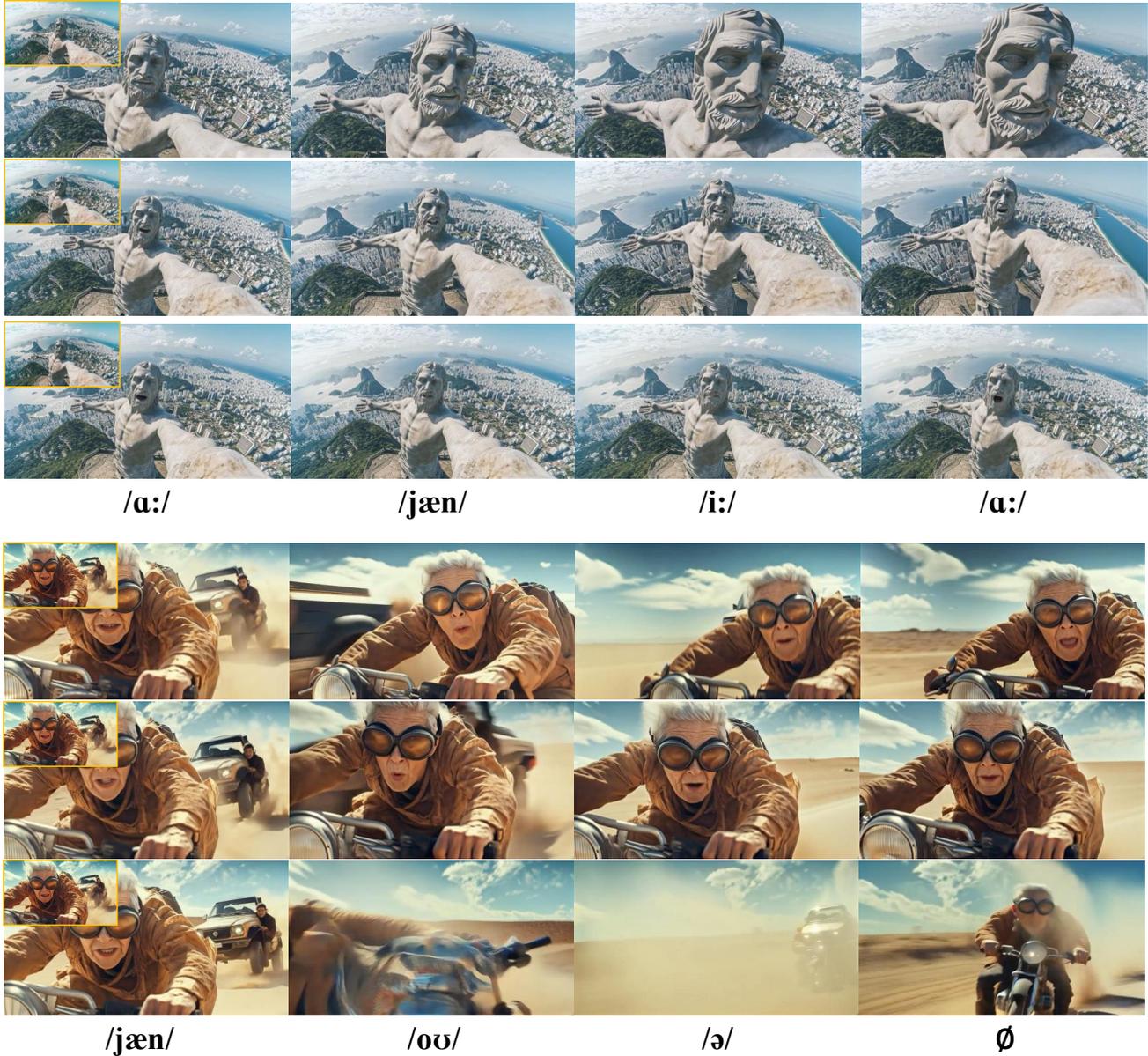
Figure 3. **Ablation study on different audio condition ratios.** The models are trained with different audio ratios (top: 10%, middle: 50%, bottom: 100%) and tested in an audio-driven setting with the same input image and audio.

Therefore, we set the pose ratio to 50% as our final training configuration.

Apart from analyzing the training ratios of new driving modalities in Stage 2 and Stage 3, the training ratio of the appearance condtion is equally important. We investigated the impact of reference image ratios on the generation of 30-second videos through two experiments: (1) setting the reference image ratio to 70%, lower than the text injection ratio but higher than audio; (2) setting the reference image ratio to 30%, lower than the injection ratios for both audio and

text. The comparative results are shown in Figure 6, revealing that a lower reference ratio leads to more pronounced error accumulation, characterized by increased noise and color shifts in the background, degrading performance. In contrast, a higher reference ratio ensures better alignment of the generated output with the quality and details of the original image. This can be explained by the fact that when the reference image training ratio is lower than that of audio, the audio dominates the video generation, making it difficult to maintain the ID information from the reference image.

Figure 4. **Ablation study on different pose condition ratios.** The models are trained with different pose ratios (top: 20%, middle: 50%, bottom: 80%) and tested in an audio-driven setting with the same input image and audio.
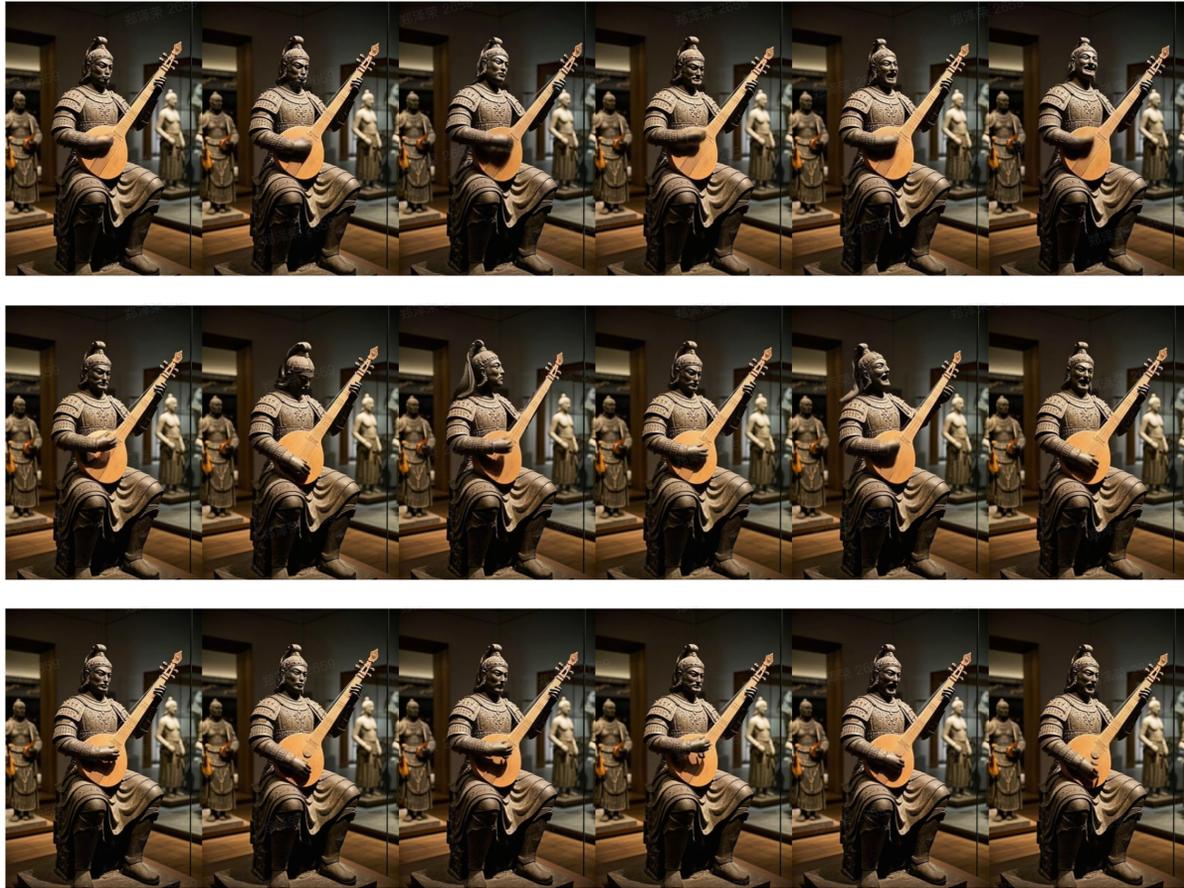

Figure 5. **Ablation study on different pose condition ratios.** The models are trained with different pose ratios (top: 20%, middle: 50%, bottom: 80%) and tested in an audio-driven setting with the same input image and audio.

Figure 6. **Ablation study on reference condition ratios.** Comparisons of visualization results for 30s videos at different reference ratios.
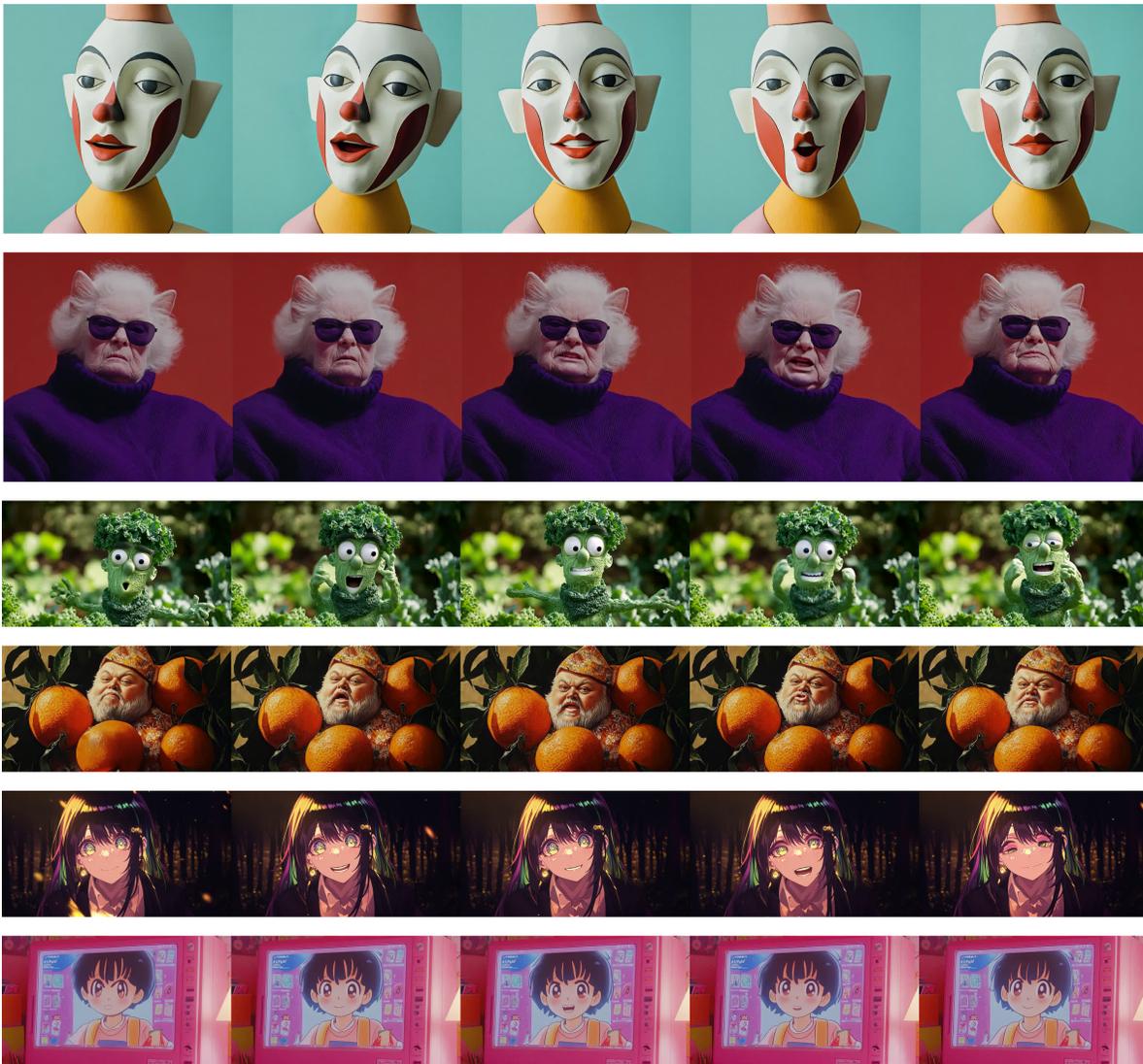

Figure 7. **The videos generated by OmniHuman based on input audio and images.** OmniHuman is compatible with stylized humanoid and 2D cartoon characters, and can even animate non-human images in an anthropomorphic manner.

4.4. Extended Visual Results

In the Figure 7, Figure 8 and Figure 9, we present more visual results to demonstrate OmniHuman's powerful capabilities in human animation, which are difficult to capture through metrics and comparisons with existing methods. OmniHuman is compatible with diverse input images and maintains the motion style of the input, such as preserving the characteristic mouth movements in anime. OmniHuman also excels in object interaction, generating videos of singing while playing different musical instruments and natural gestures while holding objects. Due to its compatibility with pose conditions during training, OmniHuman can perform pose-driven video generation or a combination of pose and audio-driven generation. More video samples can be seen on our project page (highly recommended).

5. Conclusion

We propose OmniHuman, an end-to-end multimodality-conditioned human video generation framework that generates human videos based on a single image and motion signals (e.g., audio, video, or both). OmniHuman employs a mixed data training strategy with multimodality motion conditioning, leveraging the scalability of mixed data to overcome the scarcity of high-quality data faced by previous methods. It significantly outperforms existing approaches, producing highly realistic human videos from weak signals, especially audio. OmniHuman supports images of any aspect ratio (portraits, half-body, or full-body) delivering lifelike, high-quality results across various scenarios.

Acknowledgments

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 3

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023. 3

[3] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024. 2

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.

[6] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei Efros, and Tero Karras. Generating long videos of dynamic scenes. *Advances in Neural Information Processing Systems*, 35:31769–31781, 2022. 2

[7] Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. Pixart-delta: Fast and controllable image generation with latent consistency models, 2024. 2

[8] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. *arXiv preprint arXiv:2407.08136*, 2024. 2, 5, 6

[9] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 251–263. Springer, 2017. 5

[10] Enric Corona, Andrei Zanfir, Eduard Gabriel Bazavan, Nikos Kolotouros, Thiemo Alldieck, and Cristian Sminchisescu. Vlogger: Multimodal diffusion for embodied avatar synthesis. *arXiv preprint arXiv:2403.08764*, 2024. 3

[11] Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with diffusion transformer networks. *arXiv preprint arXiv:2412.00733*, 2024. 3, 5, 6

[12] Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n' pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[13] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 3

[14] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 2, 3

[15] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5784–5794, 2021. 3

[16] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo

Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2, 3

[17] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. *arXiv preprint arXiv:2312.06662*, 2023. 2

[18] Tianyu He, Junliang Guo, Runyi Yu, Yuchi Wang, Jialiang Zhu, Kaikai An, Leyi Li, Xu Tan, Chunyu Wang, Han Hu, et al. Gaia: Zero-shot talking avatar generation. *arXiv preprint arXiv:2311.15230*, 2023. 2

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 2

[22] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2

[23] Steven Hogue, Chenxu Zhang, Hamza Daruger, Yapeng Tian, and Xiaohu Guo. Diffted: One-shot audio-driven ted talk video generation with diffusion-based co-speech gestures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1922–1931, 2024. 3, 5, 6

[24] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 3

[25] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 3

[26] Jianwen Jiang, Chao Liang, Jiaqi Yang, Gaojie Lin, Tianyun Zhong, and Yanbo Zheng. Loopy: Taming audio-driven portrait avatar with long-term motion dependency. *arXiv preprint arXiv:2409.02634*, 2024. 2, 3, 4, 5, 6

[27] Jianwen Jiang, Gaojie Lin, Zhengkun Rong, Chao Liang, Yongming Zhu, Jiaqi Yang, and Tianyun Zhong. Mobileportrait: Real-time one-shot neural head avatars on mobile devices. *arXiv preprint arXiv:2407.05712*, 2024. 3

[28] Kaggle. Ravdess emotional speech audio. https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio. 5

[29] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. 2

[30] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar,

Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023. 3

[31] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 3

[32] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2023. 3

[33] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 2

[34] Gaojie Lin, Jianwen Jiang, Chao Liang, Tianyun Zhong, Jiaqi Yang, and Yanbo Zheng. Cyberhost: Taming audio-driven avatar diffusion model with region codebook attention. *arXiv preprint arXiv:2409.01876*, 2024. 2, 3, 4, 5, 6

[35] Shanchuan Lin, Xin Xia, Yuxi Ren, Ceyuan Yang, Xuefeng Xiao, and Lu Jiang. Diffusion adversarial post-training for one-step video generation. *arXiv preprint arXiv:2501.08316*, 2025. 2, 3

[36] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3

[37] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 3

[38] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *ArXiv*, abs/2209.03003, 2022. 2

[39] Rang Meng, Xingyu Zhang, Yuming Li, and Chenguang Ma. Echomimicv2: Towards striking, simplified, and semi-body human animation. *arXiv preprint arXiv:2411.10061*, 2024. 3

[40] A Nagrani, J Chung, and A Zisserman. Voxceleb: a large-scale speaker identification dataset. *Interspeech 2017*, 2017. 3

[41] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 3

[42] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2, 3

[43] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 3

[44] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv:2303.09535*, 2023. 3

[45] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019. 3

[46] Ruizhi Shao, Youxin Pang, Zerong Zheng, Jingxiang Sun, and Yebin Liu. Human4dit: Free-view human video generation with 4d diffusion transformer. *arXiv preprint arXiv:2405.17405*, 2024. 3

[47] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019. 3

[48] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13653–13662, 2021. 3

[49] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2, 3

[50] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2

[51] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[52] Michal Stypulkowski, Konstantinos Vougioukas, Sen He, Maciej Zieba, Stavros Petridis, and Maja Pantic. Diffused heads: Diffusion models beat gans on talking-face generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5091–5100, 2024. 2, 4

[53] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 4

[54] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. *arXiv preprint arXiv:2402.17485*, 2024. 2, 4

[55] Linrui Tian, Siqi Hu, Qi Wang, Bang Zhang, and Liefeng Bo. Emo2: End-effector guided audio-driven avatar video generation. *arXiv preprint arXiv:2501.10687*, 2025. 3

[56] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, pages 244–260. Springer, 2025. 3

[57] Brooks Tim, Peebles Bill, Connorm Holmes, DePue Will, Yufeim Guo, Jing Li, Schnurr David, Taylor Joe, Luhman Troy, Luhman Eric, Ng Clarence, Wang Ricky, and Ramesh Aditya. Video generation models as world simulators. 2024. Accessed: 2024-02-15. 3

[58] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 3

[59] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 5

[60] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*, 2022. 2

[61] Cong Wang, Kuan Tian, Jun Zhang, Yonghang Guan, Feng Luo, Fei Shen, Zhiwei Jiang, Qing Gu, Xiao Han, and Wei Yang. V-express: Conditional dropout for progressive training of portrait video generation. *arXiv preprint arXiv:2406.02511*, 2024. 6

[62] Cong Wang, Kuan Tian, Jun Zhang, Yonghang Guan, Feng Luo, Fei Shen, Zhiwei Jiang, Qing Gu, Xiao Han, and Wei Yang. V-express: Conditional dropout for progressive training of portrait video generation. *arXiv preprint arXiv:2406.02511*, 2024. 2, 5

[63] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 2, 3

[64] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic human dance generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9326–9336, 2024. 3

[65] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021. 3

[66] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. Imaginator: Conditional spatio-temporal gan for video generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1160–1169, 2020. 2

[67] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 5

[68] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 3

[69] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 657–666, 2022. 3

[70] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Luc Van Gool, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024. 2, 5, 6

[71] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time. *arXiv preprint arXiv:2404.10667*, 2024. 2

[72] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiao-han Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3

[73] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiao-han Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3

[74] Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, Jinzheng He, and Zhou Zhao. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. In *The Eleventh International Conference on Learning Representations*, 2022. 3

[75] Lijun Yu, Jos Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023. 3

[76] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8850–8860, 2024. 3

[77] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8652–8661, 2023. 3, 5, 6

[78] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024. 3, 5, 6

[79] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3657–3666, 2022. 3

[80] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. 3

[81] Tianyun Zhong, Chao Liang, Jianwen Jiang, Gaojie Lin, Jiaqi Yang, and Zhou Zhao. Fada: Fast diffusion avatar synthesis with mixed-supervised multi-cfg distillation. *arXiv preprint arXiv:2412.16915*, 2024. 3

[82] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 2, 3

[83] 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 *European conference on computer vision*, pages 650–667. Springer, 2022. 3, 5

[84] Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. Taming diffusion models for audio-driven co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10544–10553, 2023. 5, 6

[85] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision*, pages 145–162. Springer, 2025. 3
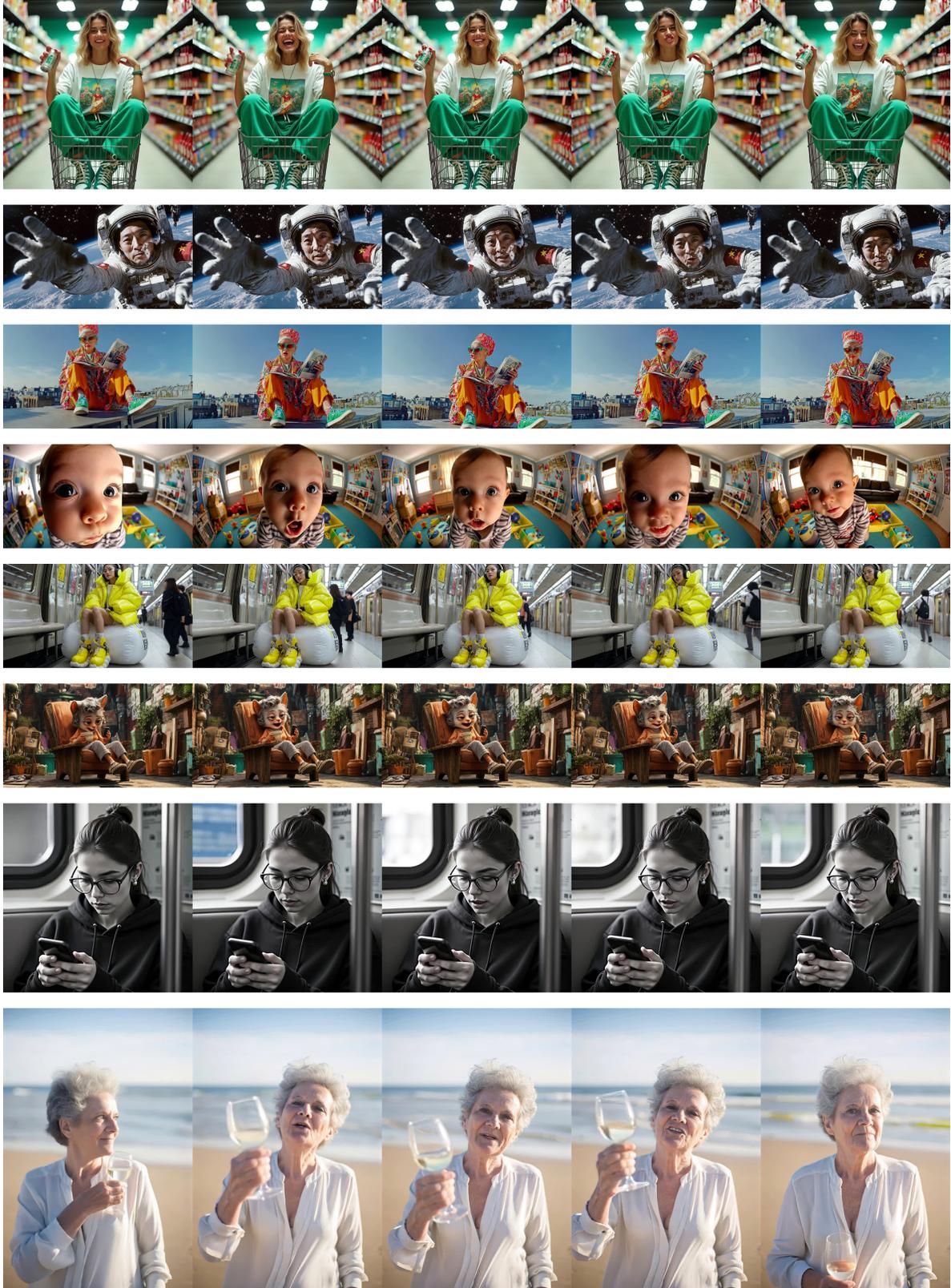
Figure 8. **The videos generated by OmniHuman based on input audio and images.** These demonstrates OmniHuman's compatibility with various environments, objects, and camera angles, producing satisfactory results.
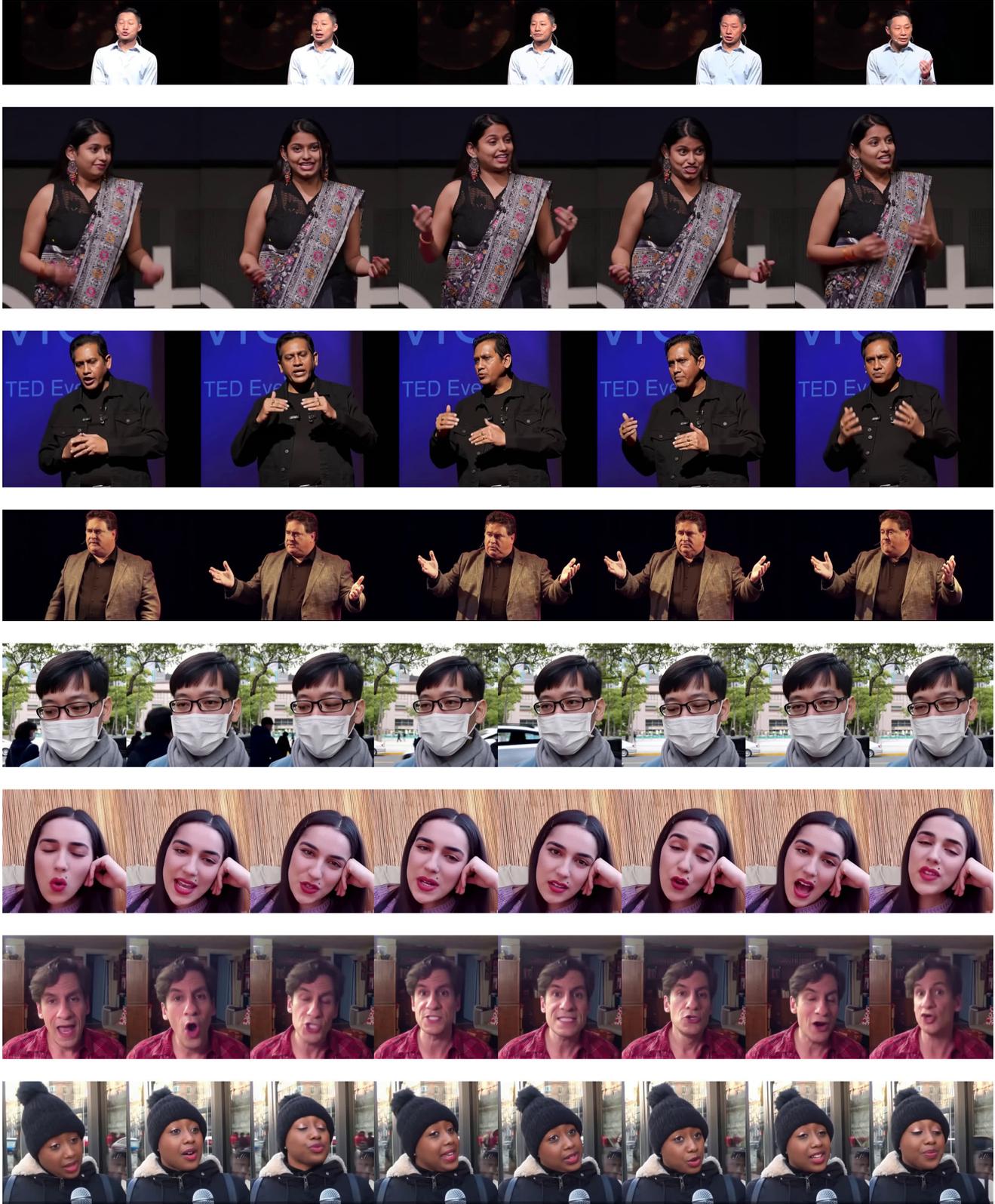
Figure 9. **The videos generated by OmniHuman based on input audio and images.** OmniHuman can generate highly realistic human motion videos, whether portrait or full-body. In these relatively standard scenarios, OmniHuman still performs satisfactorily.