Supplementary Material of OmniHuman-1

1. Additional Dataset Details

To explore the scaling potential of DiT, we curated an in-house dataset with 18.7K hours. Here, we report more dataset details.

1.1. Dataset curation.

We first filter the T2V data using rule 1, 2, 3, 4, 5, 6. Subsequently, we further annotate the data to obtain audio and pose information using rule 7, 8.

1. **Video Clip**. We first use PySceneDetect (https://github.com/Breakthrough/PySceneDetect) to detect and trim shot transitions and fades in video clips. After this phase, all clips are constrained to a duration of 5 to 30 seconds.
2. **Human**. Based on the annotated video captions, we apply a rule-based approach to identify keywords such as "people", "human", "men", "women", "girl", "boy". If any of these keywords are present, the video is classified as human-related. This method ensures that the dataset is effectively filtered to include only videos that are relevant to human activities and interactions.
3. **Subtitles**. We employ PaddleOCR (https://github.com/PaddlePaddle/PaddleOCR) to detect subtitles in the videos and filter out clips where the subtitles change. This step ensures that the dataset focuses on continuous and consistent visual content, minimizing distractions caused by textual variations and enhancing the quality of the data for downstream tasks.
4. **Visual Quality**. We employ Q-align [5] to assess the visual quality of the videos and filter out clips that fall below a predefined threshold. This step ensures that the dataset maintains a high standard of visual clarity, which is crucial for accurate analysis and robust model performance. By removing low-quality segments, we enhance the overall reliability and effectiveness of the dataset.
5. **Aesthetics**. We employ Q-align [5] to assess the aesthetics of the videos and filter out clips that fall below a predefined threshold. By applying aesthetic quality assessment, we can filter out videos that contain post-production packaging, thereby enhancing the quality of the training dataset. This step ensures that the dataset consists of natural and unaltered video content, which is more representative of real-world scenarios and improves the robustness and generalizability of the trained models.
6. **Motion**. We employ Raft [4] to compute the optical flow of the videos and filter out clips with excessively intense motion. This step ensures that the dataset retains only those video clips with moderate and meaningful motion, which are more suitable for analysis and model training. By removing clips with extreme motion, we improve the stability and quality of the dataset, leading to more reliable results.
7. **Pose**. We employ RTMPose [2] to detect human poses in video frames and filter out videos that do not contain a single person.
8. **Syncnet**. For audio-driven, we utilize SyncNet [1] to assess whether the lip are synchronized with the audio. Videos that exhibit significant asynchronization are filtered out. This step ensures that the dataset contains only high-quality, synchronized audio-visual data. By removing out-of-sync segments, we enhance the overall quality and reliability of the dataset. Finally, 13% was selected to support audio and pose modalities.

1.2. Dataset Analysis

For better understanding of the dataset distribution, we conduct analysis of our dataset across three dimensions. Below, we provide detailed definitions for each dimension, and Figure 1 reports the results.

- **Human Size**. Human size refers to the extent of the human body visible within the frame. It can be categorized into the following levels: *Portrait* (head and shoulders), *Chest* (from head to chest), *Waist* (from head to waist), *Knees* (from head to knees), and *Full Body* (the entire body visible). To achieve this, we first detect the body keypoints using RTMpose [2], and then classify the human size based on their confidence scores. This approach ensures a robust and accurate categorization of the visible extent of the human body in each video.
- **Motion Amplitude**. After obtaining the body keypoints from the video, the amplitude of human motion can be calculated by measuring the displacement of the chest keypoint over time relative to the width of the shoulders. Based on the calculated results, we classify the motion

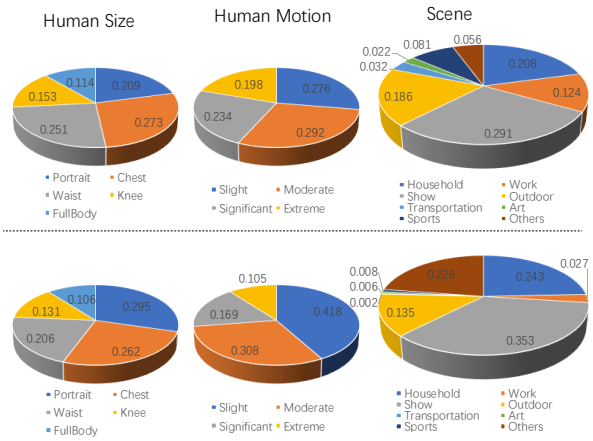

Figure 1. Analysis of our dataset. Top: full dataset, Bottom: audio part.

amplitude into four categories: *Slight*(< 0.1), *Moderate*($0.1-0.2$) , *Significant*($0.2-0.3$), and *Extreme*(> 0.3).

- **Scene**. Based on video captioning, we employ Doubao to categorize videos into the following scenarios: *Household*, *Work*, *Show*, *Outdoor Adventure*, *transportation*, *Arts and Crafts*, *Sports*, and *Others*.

2. Model Design and Training Details

Our approach is centered on a 7B-parameter MMDiT backbone, which serves as the core denoising model. This architecture features 32 Transformer blocks that alternate between 2D spatial and full 3D attention, with a hidden size of 3584. We utilize SwiGLU as the activation function and AdaSingle for timestep modulation. For multi-modal integration, we employ a minimalist fusion strategy. The denoising backbone is reused for reference image encoding to reduce parameters; text is injected consistently with the pre-trained base model; audio is incorporated via frame-wise cross-attention; and pose features are concatenated directly with the noise latents. This synergistic design allows each modality to address a specific challenge: audio ensures high-fidelity lip-sync, the reference image maintains identity consistency, text provides temporal stability, and pose conditioning mitigates motion exaggeration. The model operates on latents produced by a causal 3D VAE, which compresses video by a factor of 1024 (8x8x4 spatial-temporal compression followed by a patch convolution) into a 16-channel representation. We designed our model architecture based on the details provided in Seaweed [3].

The model is trained using a staged protocol to ensure stability and maximize the benefits of each modality. We progressively introduce text, audio, and pose conditions, with stronger signals like pose being sampled from smaller, high-quality data subsets. This strategy is crucial for avoiding instability and allows the model to effectively learn from large-scale, diverse datasets, including data often discarded by stricter pipelines. This superior data scaling capability not only enhances robustness in the core human animation task but also allows the model to retain strong generalization, leading to emergent abilities such as generating subjects in dynamic scenes with moving backgrounds or interacting with objects. During inference, our model can generate 109 frames at 720x1280 resolution with a Real-Time Factor (RTF) of 80 on 8 A100 GPUs.

3. More Experiment Results

In this section, we report more quantitative and qualitative results, as well as user studies to analyze the effectiveness of OmniHuman.

3.1. User Study

We also conduct user studies to report the human preference for our OmniHuman and baselines under audio-driven settings. 10 professional evaluators are invited to mark the MOS by their aesthetic. Four criteria are selected: identity preserving quality (ID.), emotional consistency (Emo.), visual quality (Vis.), and motion naturalness (Mot.). For half-body animation, we further report the hand quality(Han.). The results are shown in Tables 1 and 2. It can be concluded that OmniHuman beats over all s.

Table 1. User studies for portrait animation.

Methods	ID.↑	Emo.↑	Vis.↑	Mot.↑
EchoMimic	3.341	2.488	2.618	3.352
Loopy	3.872	3.335	4.236	3.127
Hallo3	3.432	2.214	2.939	3.086
OmniHuman-1	**4.015**	**3.961**	**4.754**	**3.804**

Table 2. User studies for half-body animation.

Methods	ID.↑	Emo.↑	Vis.↑	Mot.↑	Han.↑
DiffTED	0.480	0.711	0.919	1.140	0.884
DiffGest+MomiMo	2.479	2.743	3.315	2.167	2.642
CyberHost	2.925	3.073	3.862	3.527	2.855
OmniHuman-1	**3.748**	**3.835**	**4.783**	**4.121**	**3.376**

3.2. More Discussions

Effectiveness of Principle 2. Firstly, we provided more detailed experimental results to demonstrate the effectiveness of Principle 2 in Omni-Condition training. The results demonstrate that varying the audio-condition ratios significantly impacts the quality of the final video generation. From the visualizations in Figure 2, it is evident that a

Table 3. **Subjective comparison of different audio-condition ratios.**

Methods	Identity Consistency	Lip-sync Accuracy	Visual Quality	Action Diversity	Overall
10% Audio-Condition	28.84	11.59	21.59	11.59	11.59
50% Audio-Condition	**50.87**	**53.62**	**44.93**	**40.58**	**69.57**
90% Audio-Condition	11.59	30.43	13.04	36.23	17.93

Figure 2. **Ablation study on different audio-condition ratios.** The models are trained with different audio-condition ratios (top: 10%, middle: 50%, bottom: 90%) and tested in an audio-driven setting with the same input image and audio.

high proportion of audio-condition ratio training (**A=90%**) reduces dynamic range and can cause failures with complex input images. We observe that using **A=50%** for training yields satisfactory results, such as accurate lip-syncing and natural motion. However, an excessively low audio-condition ratio (**A=10%**) can hinder training, resulting in a poorer correlation with the audio. We also conducted a subjective evaluation to determine the optimal audio-condition ratio 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 training ratio of the appearance condtion. 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 3, 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.

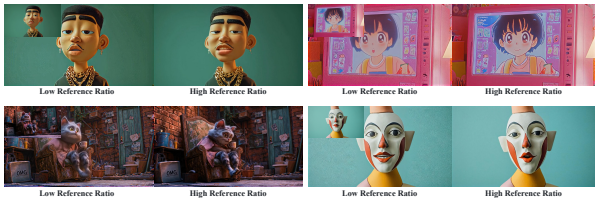

Figure 3. **Ablation study on reference condition ratios.** Comparisons of visualization results for 30s videos at different reference ratios.

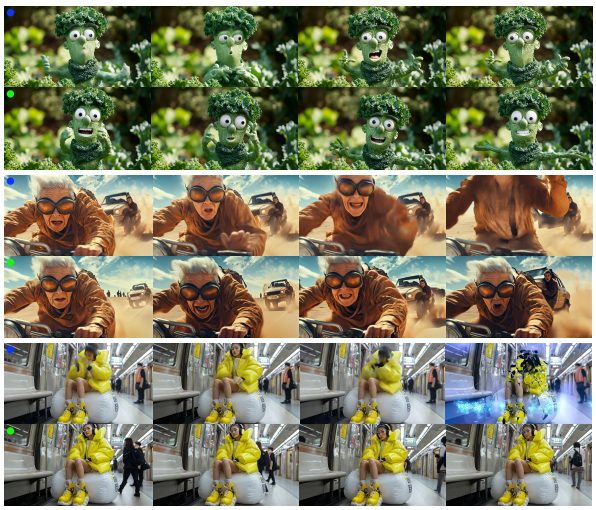

Figure 4. **Visualization analysis of inference with text-condition.** Blue circle indicates inference without text-condition and green circle denotes using text-condition.

Impact of text condition during inference. Because OmniHuman is trained with mixed-condition training, it can

Figure 5. **Visualization comparison of different audio CFG scales.** From top to bottom, the CFG scales are **1**, **3.5**, and **6.5**, respectively.

benefit from mixed-condition inference during deployment. Even in an audio-driven setting, we can obtain text input through image captioning. In Figure 4, we show the differences with and without text input. It is worth noting that these benefits all stem from the high extensibility brought by the omni-conditions training strategy.

Impact of CFG on the results. In Figure 5, we also validate the impact of different audio Classifier-Free-Guide (CFG) scales on performance. It can be observed that with CFG=1, the character is almost unable to speak and there are noticeable artifacts, especially for out-of-distribution (OOD) data. With CFG=3.5, the character can speak, but the expressiveness is insufficient and the image quality suffers from a loss of detail. However, with CFG=6.5, both the detail and expressiveness show significant improvement. Based on this, we attempted to have specific individuals in the frame speak by assigning different audio configurations to various regions during testing.

Long video generation. OmniHuman leverages motion frames to generate long videos in an autoregressive manner. In Figure 6, we present the results of a 90-second video. It can be observed that the character's identity and the video quality are well-preserved throughout.

Hybird-modality driving. Additionally, leveraging OmniHuman's support for both Audio and Pose-driven features, we can perform either standalone audio-driven or multimodal combination-driven animations. This flexibility makes it highly versatile for various use cases. The visual effects are illustrated in Figure 8.

Data Scaling. We also provide experimental results on data scaling, as shown in Figure 7. Compared to the baseline, which relies solely on audio conditioning, our OmniHuman demonstrates superior performance as the training data scales up.

4. Limitation of Current Model

Although OmniHuman reduces overfitting of motion to audio changes by introducing mixed-conditions training, especially with the inclusion of the pose condition, the weak correlation between audio and motion means that the same audio segment can correspond to different movements based on emotion and personal style. As a result, there can still be instances of uncoordinated motion. Specifically, for some audio inputs, the generated results may appear overly

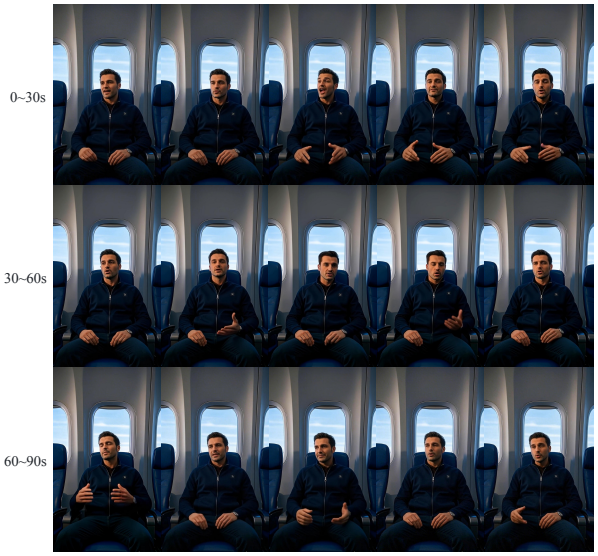

Figure 6. **Visualization of long video generation.**

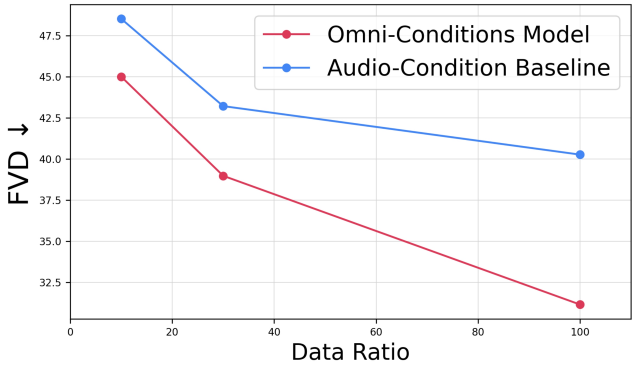

Figure 7. **Comparison of data scaling results for different training methods.**

Figure 8. Visualization of video results driven by different modality combinations.

forceful in articulation and excessively expressive, making them seem uncoordinated and unnatural. We illustrate this in Figure 9, and it is more easily noticeable in the provided video samples. In addition to cases of overly expressive articulation, OmniHuman can respond to actions involving object interaction. However, there are still instances where the results appear unrealistic, such as mismatched piano keys and music when playing the piano, or incorrect guitar strings when playing the guitar. For some rare combinations of human-object interactions, the generated results can also appear unnatural, such as hands remaining static and not interacting with the object.

In actuality, the current results generated by OmniHuman still exhibit differences from real human likenesses and can be identified as AI-generated content. This may also be due to the model's inability to accurately replicate variations in lighting, muscle movement details, and other subtle nuances, which remain challenging areas for the model to address.

The underlying causes are numerous, but the most fundamental issue may be the insufficiency of the model's training. This necessitates the use of techniques like classifier-free guidance (CFG) to achieve stable results, which, in turn, can lead to overfitting. In future work, we plan to address this by incorporating more comprehensive motion conditions. This could include explicit style embeddings, motion intensity parameters, or high-level intention descriptors to decouple the generation process from a simplistic reliance on audio cues and produce more natural and contextually appropriate human animations.

Figure 9. **Demonstrations of the limitations.** OmniHuman sometimes generates overly expressive results for audio with strong emotional content or emphasized articulation.

5. Discussion on Responsible AI

OmniHuman generates realistic AI portrait videos based on simple inputs, catering to the demands of entertainment and film production, such as AI music videos and AI movies. Although the current results are quite good, they still exhibit traces of being produced by AI models, which can help mitigate the risk of misuse to some extent. Nevertheless, we believe that this type of technology should be used with oversight to ensure it is not employed for malicious purposes, such as fraud. Specifically, during use, we can implement the following measures to prevent misuse: (1) Apply clear watermarks to all generated content to indicate that it is produced by AIGC algorithms. (2) Utilize filtering algorithms to review and block inappropriate, vulgar, or malicious input audio and generated video content. (3) Embed traceable watermarks within the content for accountability. By following these steps, we can help ensure that the technology is used responsibly and ethically.

References

[1] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016. 1

[2] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023. 1

[3] Team Seawead, Ceyuan Yang, Zhijie Lin, Yang Zhao, Shanchuan Lin, Zhibei Ma, Haoyuan Guo, Hao Chen, Lu Qi, Sen Wang, et al. Seaweed-7b: Cost-effective training of video generation foundation model. *arXiv preprint arXiv:2504.08685*, 2025. 2

[4] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 1

[5] 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. 1