AudioAvatar: Personalized Audio-driven Whole-body Talking Avatars

Supplementary Material

Appendix

In this supplementary material, we provide additional details, experiments, analyses, and discussions that were not included in the main manuscript due to space constraints. We also provide supplementary video for temporal results of our approach. As this supplementary materials contain important information that will help deepen your understanding of the paper, we kindly and respectfully ask you to read it carefully. In the Appendix, we provide the following:

A. Additional Related Work

Co-Speech Gesture Video Generation. Similar to large-scale human video diffusion models, prior work has studied human video generation from audio, skeleton data, or 2D/3D poses, often through a two-stage pipeline: mapping audio to poses and then using a pre-trained GAN-based pose2video model [15, 67]. More recently, diffusion models [20, 22, 41, 42, 67] have been applied. A study [26] has also been introduced that generates a style-specific anchor avatar video from only a one-minute video clip. With notable work [38] directly generating videos from audio, showing that bypassing the audio-to-pose step—long a performance bottleneck—can advance the task. While this task shares the common goal of generating talking human videos from audio signals, it differs significantly from our approach: such methods typically produce only 2D videos rather than 3D avatars, and the generated content is limited to the upper body. Furthermore, they have been validated only on constrained domain-specific datasets, such as TED talks [73].

Speech-driven Whole-body Motion Generation. This section focuses on methodologies that generate body, face, and hand parametric motions together. It is the task of automatically predicting natural, human-like body and hand gestures that align with spoken language. Unlike earlier works that focused on generating only facial expressions or body gestures in isolation, recent research has begun to explore the simultaneous generation of body, face, and hand gestures. These studies have introduced several methodological advances, including the use of VQ-VAE architectures [95], the adoption of large-scale datasets [43], diffusion-based generative models [9, 11, 54], and real-time generation [46] enabled by MAMBA [16] or flow matching [40] approaches. More recently, motion generation has been significantly improved through multi-task learning that incorporates diverse multimodal signals such as speech, text, and music, along with tasks including text-to-motion, audio-to-motion, and dance generation [6].

B. Implementation Details

For reproducibility, we provide additional implementation details of our model that were not fully described in the main paper. To perform knowledge distillation, we train a Gaussian deformation field model [89] using synthesized talking videos generated from large-scale video diffusion models as supervision. From this model, we extract the time-varying Gaussian deformation field $\Delta\mathcal{G}_t$ from the canonical Gaussians. This deformation field serves as guidance for aligning the audio sequence within a shared embedding space. To establish this shared embedding space, we adopt a CLIP-wise objective, where we train an encoder to minimize the cosine distance between the feature vectors of the audio representation and those of $\Delta\mathcal{G}_t$ (after being mapped to particle motion). The encoder consists of a pre-trained Perceiver [27] followed by a 2-layer MLP. To encourage the embeddings to capture local context, we apply a patchify strategy with a window size of 25. For audio-driven particle motion generation, we employ a diffusion transformer that synthesizes a fixed-length motion sequence of $T = 125$ frames. During training, we set the diffusion time step to $\tau = 200$ and make use of a cosine noise scheduler. We further adopt a curriculum learning scheme: the model is first trained using the DDPM-style objective L_{simple}, and subsequently fine-tuned with the full objectives. Our implementation is based on PyTorch [58]. We use the AdamW optimizer [48] with a learning rate of 1×10^{-5}.

C. Hybrid Data Construction

Audio-driven Human Video Diffusion Model. We describe the detail for a multimodal conversational talking human video dataset, to train our proposed method for

audio-driven whole-body 3d talking human avatar, by integrating text, audio, and video through a unified pipeline (Fig. 6). We first build an attribute dictionary covering gender, age, body type, hairstyle, and clothing to ensure diverse yet structurally consistent full-body human renderings. Using a diffusion-based text-to-image model [5], we synthesize photorealistic humans with clear gestures and coherent full-body appearance. In parallel, a curated text corpus is then designed to reflect natural conversational flow, serving as prompts for generating high-quality speech via TTS [13]. Finally, audio-driven human video diffusion models [10, 14] transform each static image into a temporally synchronized talking human clip, aligning lip motion, facial expressions, and subtle gestures with the audio. For each generated subject, we create 120 training motions to ensure sufficient diversity and reliable evaluation. This hybrid construction yields a diverse, reproducible resource suitable for training whole-body, speech-driven human avatar models.

Multimodal Video Foundation Model. However, the aforementioned models are trained on constrained-domain datasets with relatively small data volume, and thus often produce videos with insufficient visual quality, such as restricted resolution and temporally inconsistent motion. To mitigate these issues, we additionally leverage a large-scale, multi-modal video foundation model trained on an open-domain dataset. Given a reference image and a text prompt—"He/She is engaging in a conversation with someone positioned at the camera's viewpoint, accompanied by dynamic co-speech talking gestures."—the model synthesizes four high-quality talking human videos with diverse gestures aligned to arbitrary audio. It is generated for each subject and used as supplementary supervision to enhance the robustness and realism of our framework.

D. User Study

Evaluating the naturalness, expressivity, and conversational realism of talking motion—particularly when co-speech gestures are involved—is highly challenging using quantitative metrics alone, due to the stochastic nature of gesture production and the subjectivity inherent to gesture perception. Therefore, in addition to the quantitative evaluations, we conduct four different user studies for both comparative and ablative analysis.

To assess statistical significance among the existing comparative methods, we perform paired t-tests. Based on the results reported in Table 1, we observe that HunyuanVideo-Avatar [10] shows no statistically significant difference from OmniAvatar [14], while it has statistically significant difference in the remaining methods including the 3d avatar methods. Consequently, we perform an user preference study comparing our method against the baseline, randomly sampled results from OmniAvatar and HunyuanVideo-Avatar. We recruited a total of 41 participants for the survey. For each result, participants were asked to evaluate: (i) identity preservation, (ii) visual quality, (iii) synchronization, (iv) gesture naturalness, and (v) gesture diversity. Across all criteria, our method achieved better preference scores. Furthermore, a two-proportion z-test confirms that our approach yields statistically significant perceptual improvements over the baseline. Please refer to Fig. 11 and Fig. 14 for experimental results and a questionnaire sample.

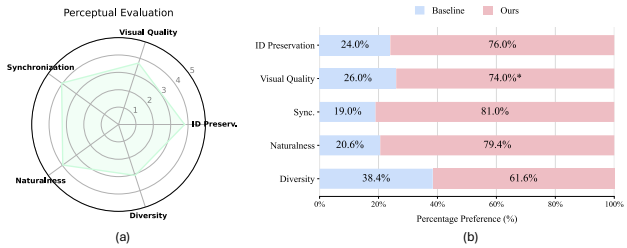

Figure 11. **User study.** (a) perceptual evaluation with preference score of our model, and (b) perceptual comparison between baseline and ours.

E. Generalization on Input Audio

We additionally evaluate the robustness of our method to variations in input audio sequences, specifically its generalization capability with respect to its semantics. This is measured by comparing the performance drop ratios between training and test audio sequences under in-distribution conditions, as well as performance drops on out-of-distribution (OOD) audio sequences. Please refer to Table 3 for detailed results.

We first analyze the performance drop ratio between the training and test sets under in-distribution conversational speech. Our method exhibits only a 2.7% deviation on average across all evaluation metrics, indicating strong generalization within the conversational speech domain. Next, we evaluate generalization to OOD audio by introducing 20 additional audio sequences per test subject. The OOD set consists of 5 monologue sequences, 5 lecture sequences, 5 narration sequences, and 5 highly emotional speech sequences, while the training audio comprises only conversational-style speech. The results show that our method yields less than 6.5% performance drop relative to in-distribution test sequences. This demonstrates that our approach generalizes well even to OOD audio inputs.

We clarify that our framework is a subject-specific personalized model. For each subject, the audio used during training corresponds to the same identity, and therefore we do not target generalization to entirely new voice identities. Furthermore, our primary application scenario assumes audio generated from personalized TTS models [13] conditioned on the user's own voice. Consequently, extreme

Data	IQA↑	ASE↑	SyncC↑	SyncD↓	HKC↑	CSIM↑	FID↓	FVD↓
Train	4.33	2.90	7.40	5.28	0.912	0.563	11.9	232
Test	4.22	2.83	7.20	5.42	0.897	0.551	12.4	240
OOD	4.15	2.77	7.05	5.55	0.884	0.542	13.2	252

Table 3. Quantitative table for generalization performance evaluation. We report quantitative measurement on in-distribution training/test, and out-of-distribution(OOD) data. We performed paired t-test after multi-comparison correction using false discovery ration (FDR) [4] at $q = 0.05$. There is no statistical significance on the OOD, which suggest that our method is well-generalized on unseen audio sequences.

noise corruption or mismatched voice identities are outside the scope of our evaluation.

F. Ablation Study on the Proposed Losses

In Fig. 9 and Table 2 of the main manuscript, we present a comprehensive ablation study on the proposed components, including the video score distillation loss $\mathcal{L}_{vsd}$ and the trajectory alignment loss $\mathcal{L}_{traj}$. To more delicately verify the effectiveness of $\mathcal{L}_{vsd}$ and $\mathcal{L}_{traj}$, we further evaluate their contributions to *motion realism* and *motion smoothness* (temporal consistency). These two aspects correspond directly to the primary motivations for introducing $\mathcal{L}_{vsd}$ and $\mathcal{L}_{traj}$, respectively.

To evaluate motion realism, we adopt the Fréchet Video Motion Distance (FVMD) [44]. This metric computes the feature distribution distance between motion sequences extracted from real videos and those rendered by our model.

To assess motion smoothness (temporal consistency), we measure the acceleration error [30]. Specifically, we first estimate human joints—including body, face, and hands—from both ground-truth videos and rendered videos using a pre-trained human pose estimator [8]. For each, accelerations are computed via second-order differentiation of the joints across adjacent frames. The acceleration error is then defined as the L2 norm between the accelerations of the rendered video and the corresponding ground-truth video.

Table 4 reports FVMD and Acc. Err. for variants where $\mathcal{L}_{vsd}$ and $\mathcal{L}_{traj}$ are individually removed (FVMD is scaled by 10^{-3} for readability). The results confirm that $\mathcal{L}_{vsd}$ substantially improves motion realism, while $\mathcal{L}_{traj}$ greatly enhances motion smoothness.

G. Ablation on the Hybrid Data Construction

To enhance both appearance fidelity and identity preservation, we propose a hybrid data construction strategy that leverages not only synthesized data generated by audio-driven video diffusion models specialized for talking human video generation, but also additional data synthesized by an open-domain large video generative model [88], regardless of category. To clearly isolate the effectiveness of this hybrid strategy, we provide visual comparisons between a variant trained without the open-domain synthesized data and our full model trained with it, as shown in Fig. 12. The result demonstrate that our hybrid data construction strategy effectively encourages the generation of rendered avatar with high-fidelity visual quality.

$\mathcal{L}_{vsd}$	$\mathcal{L}_{traj}$	FVMD ↓	Acc. Err. ↓
✗	✗	11.37	16.84
✗	✓	9.76	15.91
✓	✗	6.84	28.79
✓	✓	**5.63**	**14.34**

Table 4. **The Effectiveness of $\mathcal{L}_{vsd}$ and $\mathcal{L}_{traj}$.** The objective terms improve temporal consistency in terms of motion realism (FVMD) and smoothness (Acceleration Error(Acc. Err.)), respectively.

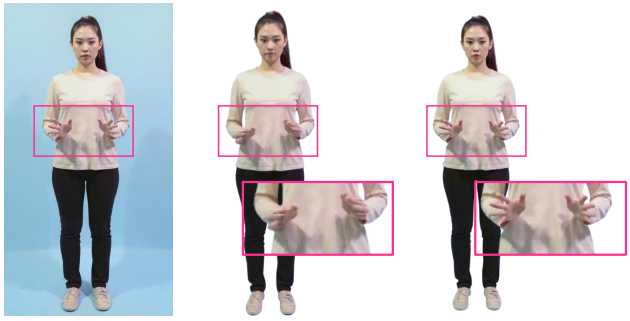

Figure 12. **The effectiveness of hybrid data strategy.** It effectively encourages the generation of rendered avatars with high-fidelity visual quality.

H. Additional Quantitative Comparisons

In addition to Table 1 in the main paper, we further report quantitative comparison results on the publicly available seamless-interaction dataset, as well as on our collected in-the-wild talking videos, summarized in Table 5 and Table 6. Consistent with the observations in the main table, our method demonstrates clear and stable improvements over existing animatable 3D avatar methods and audio-driven human video generation approaches. When compared with 3D avatar–based methods, the results highlight the effectiveness of our direct audio conditioning strategy and the successful distillation of motion priors from large-scale video diffusion models. Moreover, when compared with audio-driven human video generation models, the superior performance verifies the benefit of our hybrid data construction strategy in combination with a open-domain large-scale image/text-to-video diffusion model.

Methods	IQA↑	ASE↑	SyncC↑	SyncD↓	HKC↑	CSIM↑	FID↓	FVD↓
EchoMimicV2	2.86	1.68	3.50	11.73	0.711	0.389	26.2	483
OmniAvatar	3.39	2.24	5.44	8.74	0.729	0.446	21.4	403
HunyuanVid	3.47	2.30	5.87	8.19	0.744	0.458	19.8	368
LHM	3.23	2.12	5.19	8.05	0.731	0.425	0.595	420
PERSONA	3.30	2.19	5.36	7.82	0.737	0.434	21.7	397
Ours	**3.59**	**2.41**	**6.12**	**6.23**	**0.762**	**0.468**	**14.3**	**276**

Table 5. Quantitative comparison with existing animatable 3d avatar and audio-driven human video generation methods on Seamless Interaction [1].

Methods	IQA↑	ASE↑	SyncC↑	SyncD↓	HKC↑	CSIM↑	FID↓	FVD↓
EchoMimicV2	2.26	1.33	2.76	13.57	0.560	0.307	30.3	558
OmniAvatar	2.67	1.77	4.29	10.10	0.575	0.352	24.7	465
HunyuanVid	2.73	1.82	4.62	9.47	0.587	0.361	22.8	426
LHM	2.55	1.68	4.09	9.31	0.576	0.335	25.9	487
PERSONA	2.60	1.73	4.23	8.98	0.583	0.344	25.1	462
Ours	**2.83**	**1.90**	**4.82**	**7.20**	**0.601**	**0.369**	**16.5**	**319**

Table 6. Quantitative comparison with existing animatable 3d avatar and audio-driven human video generation methods on collected in-the-wild full-body talking videos.

I. Additional Qualitative Results

We provide additional qualitative comparisons in Fig. 15 and Fig. 16 to demonstrate the superior visual quality of the rendered avatar animations produced by our method. These figures compare our approach against existing animatable 3D avatar methods and audio-driven human video generation models, respectively. We also include sample qualitative results on both the Seamless Interaction dataset and in-the-wild talking human videos in Fig. 17 and Fig. 18.

J. Runtime Report

We measure the end-to-end processing time, including input data loading, preprocessing, the full forward pass of our audio-driven particle deformation module and differentiable Gaussian renderer, and final talking-video synthesis. On average, our model requires approximately 70 seconds to generate a 125-frame output sequence on a NVIDIA A100 GPU. We cautiously note that, based on our internal measurements rather than officially published numbers, existing audio-driven human video diffusion models [10, 14] typically require around one hour to synthesize a talking-video sequence of comparable length. This indicates that our approach—distilling video diffusion knowledge into sparse Gaussian primitives via 3D Gaussian splatting—offers approximately 50× faster inference, while still preserving high-fidelity visual quality. Although existing 3DGS-based human avatar approaches can achieve near–real-time rendering—for example, PERSONA [74] reports around 25 FPS—our method delivers substantially higher visual quality across all evaluation metrics. In particular, compared to PERSONA, we obtain notable improvements such as +8.8% IQA, +9.7% ASE, +14.3% SyncC, and a 20.3% reduction in temporal desynchronization in SyncD, demonstrating significant gains in both perceptual fidelity and audio–motion coherence, as reported in Table 1. Our method may represent a more suitable choice for generating high-fidelity, fully animatable 3D talking avatars.

K. Discussion

Direct Audio Conditioning. A central claim of this work is that animatable 3D avatars for talking animation can be effectively modeled without relying on parametric motion representations such as SMPL-X. Instead, we propose directly driving avatar motion from the raw audio signal. As shown qualitatively in Fig. 7, our approach produces more faithful appearance synthesis and richer expressiveness compared with existing 3D animatable avatar methods that require an auxiliary audio-to-SMPL-X module for driving. This finding is further corroborated by the quantitative results in Table 1, where our method outperforms prior parametric-motion–driven 3D avatar approaches.

Video Diffusion Distillation. Training an animatable 3D talking avatar typically requires video data containing the subject's talking motion. However, in the single-image scenario—where only one reference image is provided—acquiring such subject-specific supervision is challenging. To address this limitation, we distill knowledge from large-scale video diffusion models pretrained on diverse datasets. We introduce a hybrid data construction strategy that synthesizes person-specific talking videos guided by the reference image using multiple diffusion models. To further ensure that the avatar faithfully inherits the motion priors encoded in these models, we design the video score distillation loss and the trajectory alignment loss. The effectiveness of our strategy is quantitatively validated in Table 2 and Table 4, and visually demonstrated in Fig. 9 and Fig. 12. Together, these results indicate that our proposed framework enables efficient rendering while preserving high visual quality, outperforming not only existing animatable 3D avatar methods but also audio-driven human video generation models.

L. Limitation and Future Work

While our approach achieves compelling results on photo-realistic rendering and animation of whole-body 3d animatable avatars, directly drivable from input audio sequences, we acknowledge several limitations.

Novel-view Synthesis. Since our system constructs avatars given a single reference image and augments training/supervision talking video through multiple video diffusion models, the generated videos are primarily near-frontal views, which limits synthesize the appearances of avatar

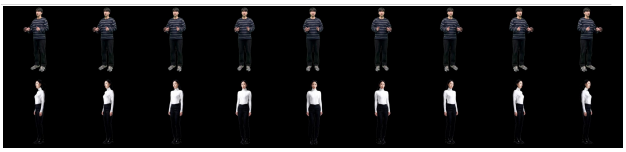

Figure 13. **Limitation on novel-view synthesis.** The rendering fidelity degrades as the viewpoint deviates from the frontal direction, primarily because the synthesized talking videos distilled from video diffusion models exhibit a strong bias toward frontal views.

under large viewpoint shifts. As a result, our method reliably supports synthesis only for near-frontal viewpoints (Fig. 13). Nevertheless, our framework offers two key advantages: (1) compared to existing animatable 3D avatar methods, it enables talking-avatar animation directly driven from raw audio while producing substantially improved visual quality; and (2) compared to large-scale audio-driven human video generation models, it supports efficient and fast inference/rendering, allowing high-quality talking video synthesis with significantly reduced computational cost.

Interactive Multi-Person 3D Talking Animation. Our framework focuses on generating talking animations for a single human subject. However, for real-world applications such as telepresence and VR/AR, scenarios involving multiple interactive speakers, each requiring coherent talking animation and rendering—naturally arise. As a future direction, leveraging datasets that capture dyadic conversations, behaviors, and interactions [1] could enable the modeling and animation of interactive dyadic 3D avatars. We believe that extending our framework in this direction represents an exciting avenue for research.

M. Broader Impacts

Potential Negative Societal Impacts. Our work advances high-fidelity, audio-driven 3D talking avatars but also carries risks. The technology could be misused to create deceptive or harmful media, such as deepfakes for misinformation, harassment, or identity fraud, raising ethical and legal concerns about trust in digital communication. Fairness and bias are also issues, as underrepresented groups in training data may experience degraded performance. Privacy risks emerge if avatars are generated without consent, and high computational demands may limit accessibility, reinforcing the digital divide.

Broader Impacts. At the same time, this technology offers significant benefits. Personalized 3D avatars can enhance telepresence, education, and remote collaboration, lowering communication barriers across diverse contexts. For individuals with disabilities, avatars may open new channels for expression and inclusion. The method also benefits entertainment, creative industries, and mixed reality applications. More broadly, it contributes to understanding the coupling of speech and gesture. To support responsible use, future work should incorporate safeguards such as watermarking, provenance tracking, and bias-aware evaluations.

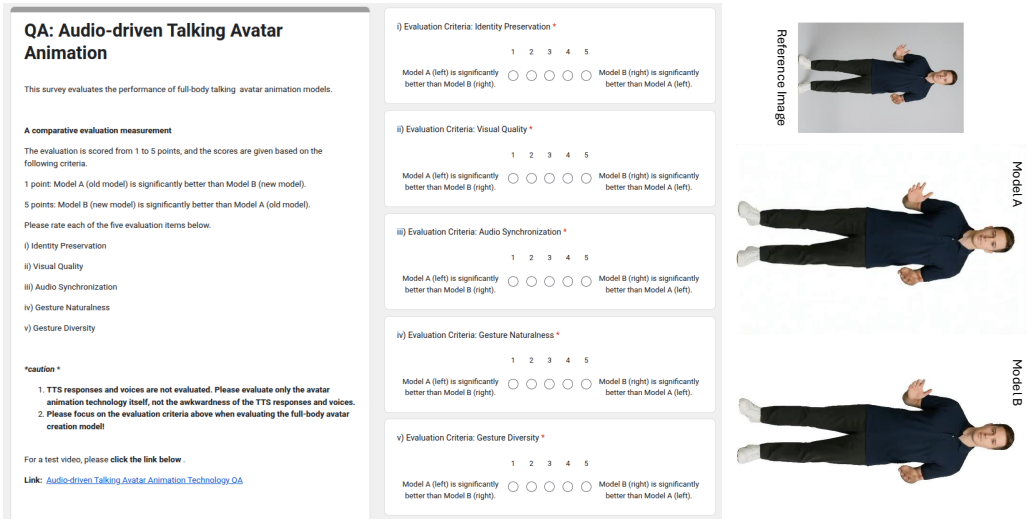

Figure 14. A survey sample for user study.

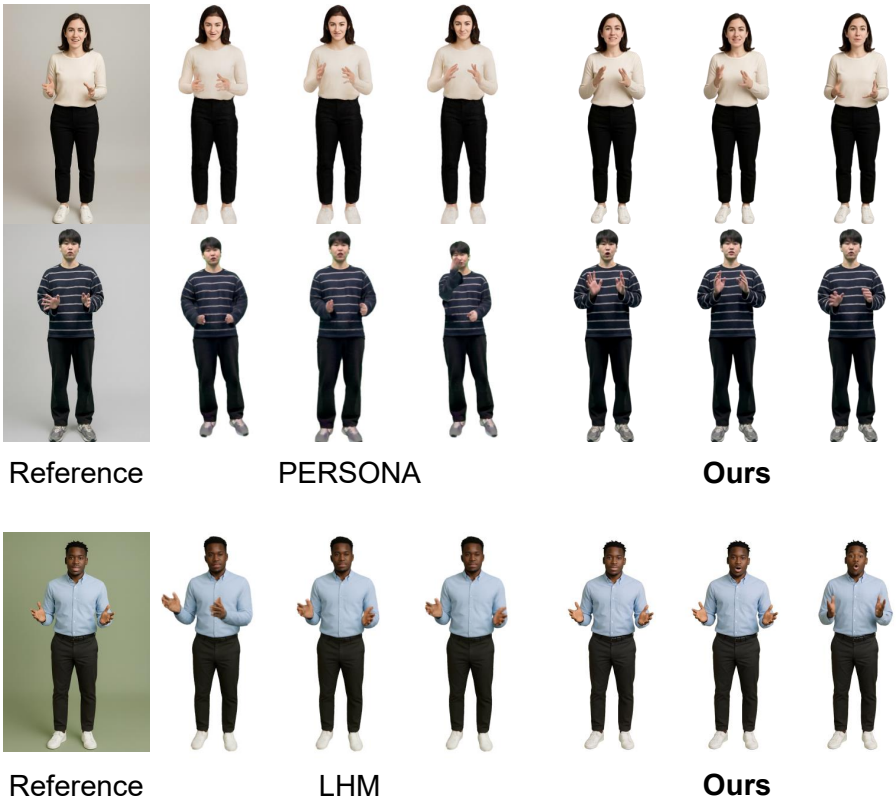

Figure 15. Visual comparison with existing single-image animatable 3D avatar methods.

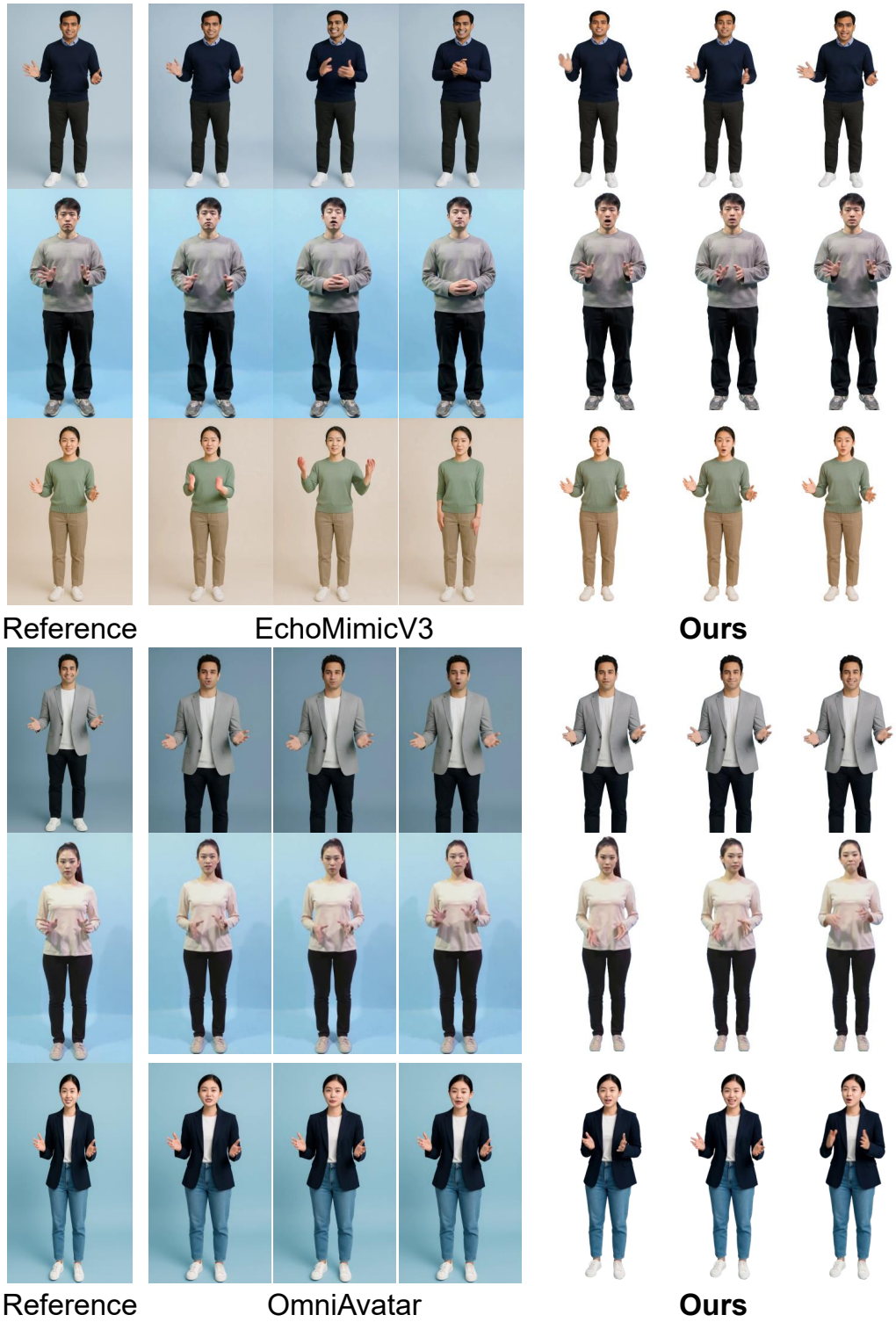

Figure 16. Visual comparison with existing audio-driven human video diffusion models.

Figure 17. Qualitative examples on Seamless Interaction dataset.

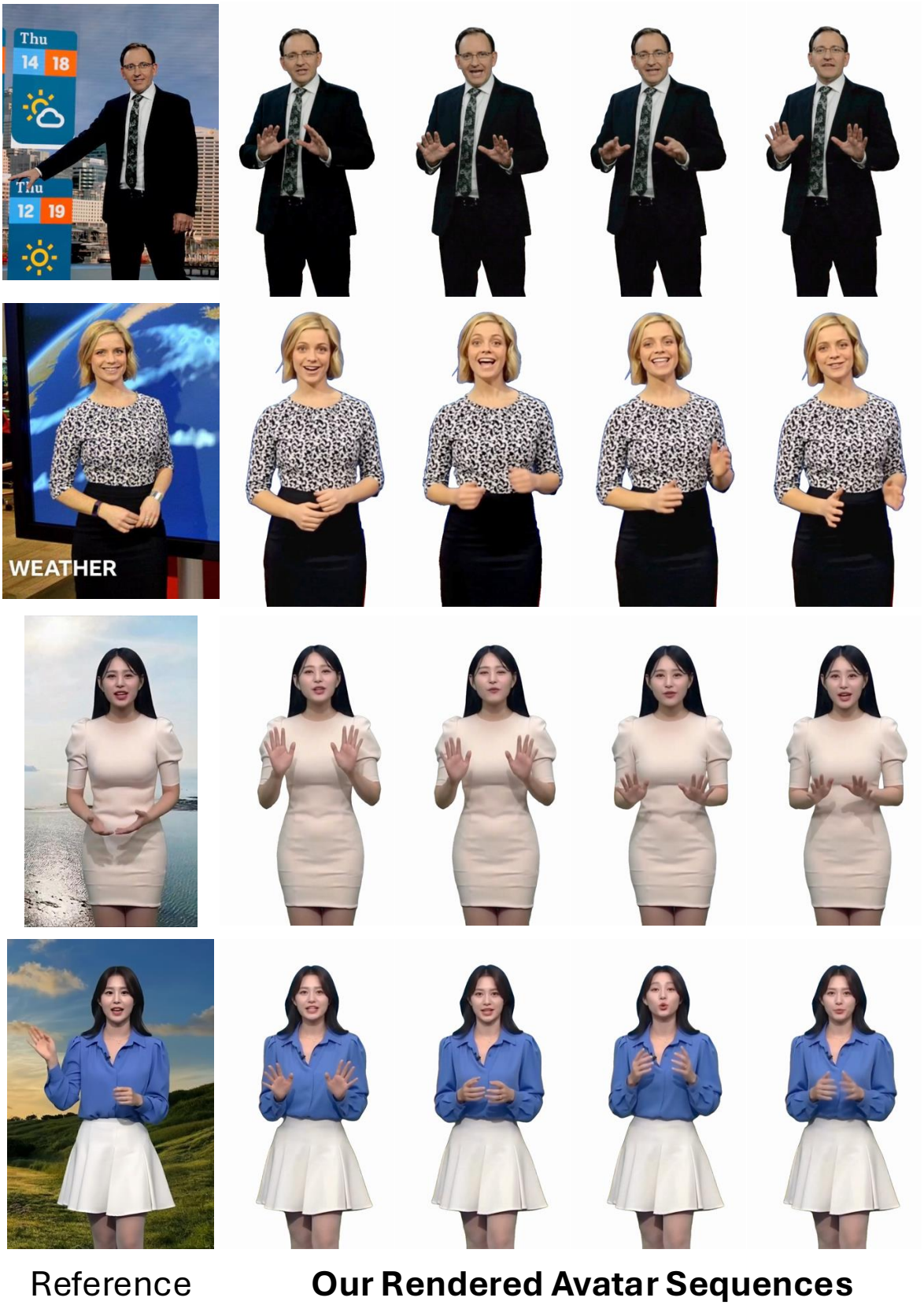

Figure 18. Qualitative examples on in-the-wild subjects.