# Expressive Talking Human from Single-Image with Imperfect Priors

## Supplementary Material

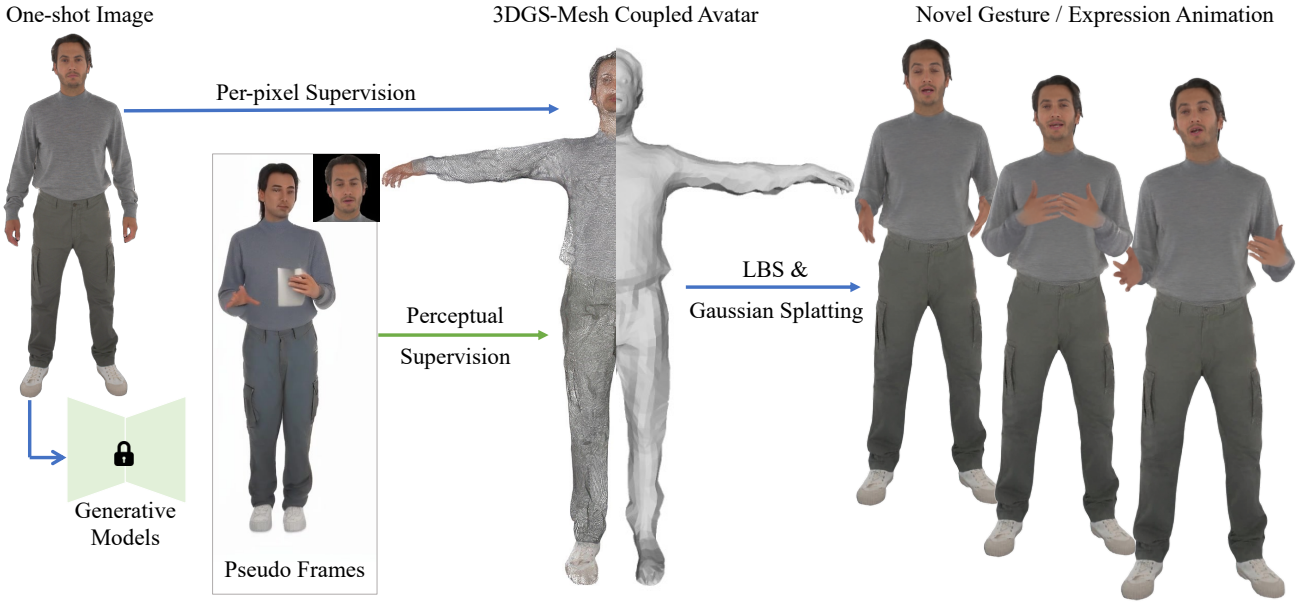


Figure 9. A simplified illustration of our pipeline.

## A. Pipeline

For a comprehensive understanding, we present a simplified illustration of our pipeline in Fig. 9. In Gaussian avatar representation, two deformation fields are defined in canonical space (SMPL-X T-pose), with LBS for animation and Gaussian Splatting for image rendering. Unlike these dense-input Gaussian avatars, we introduce mesh deformation and related soft constraints (Eq.7) to regularize Gaussians and stabilize training.

## B. Limitation

**Tracking.** Accurate SMPL-X tracking is essential for mesh-based avatar representation. Our method relies on precise registration between the input image or video and the parametric human mesh, which can be compromised by tracking inaccuracies. Additionally, self-intersection may occur when we conduct cross-identity animation, particularly observed in the finger area, as illustrated in Fig. 10.

## C. Additional Results

**Comparison with more 3D works.** SHERF is a concurrent work alongside ELICIT, sharing similarities such as SMPL and NeRF-based representation. Visual comparisons are provided in Fig. 11. The NeRF-based approach constrains their rendering resolution and efficiency, while the reliance on the SMPL proxy prevents them from accurately modeling fine hand and facial movements.

LHM proposes a large reconstruction mode to infer 3D

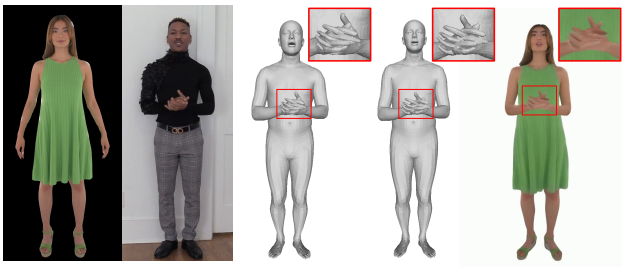


Figure 10. Inaccurate tracking and finger self-intersection during cross-identity animation.


Figure 11. Qualitative comparisons with SHERF. Our method effectively captures detailed facial and hand movements while ensuring better coherence and higher overall visual quality.


Figure 12. Qualitative comparison with LHM.

Gaussian avatars in a feed-forward pass. While efficient, it lacks expression control and fails to fully utilize input information, leading to texture loss and coarse outputs, as shown in Fig. 12. Our well-motivated designs are crucial for these improvements.

**Comparison with more 2D works.** We have already compared two representative 2D methods, StableAnimator and Make-Your-Anchor, in the main paper. Here, we further compare two additional 2D approaches: MimicMotion and Animate-X, with results shown in Fig. 14. MimicMotion enhances pose accuracy using a confidence-aware strategy, while Animate-X enables image-conditioned, pose-guided video generation with high generalizability. However, both methods still struggle with identity preservation and capturing fine texture details. Animate-X performs worse on real-human image inputs due to the lack of pose alignment. Moreover, like other pose-guided human video diffusion models such as StableAnimator, both methods fail to accurately control foot poses due to the limitation of sparse 2D landmark conditioning.

**Visualization on more scenarios.** We present dance-driven animations and long skirt results in Fig. 13. Our method is robust in most cases; however, simulating complex motions for long skirts remains challenging (bottom right).


Figure 13. More visualization on more scenarios including dance-driven animations and long skirt results. Our method demonstrates sufficient motion diversity and is robust in most cases, although our focus is on talking scenes.

**Subjective evaluation.** We also conducted a user study with 30 participants, showing the best percentage for each method in Tab. 3.

| Indicators | ExAvatar | StableAnimator | LHM | Ours |
|---|---|---|---|---|
| Identity Preservation | 9.9 | 6.0 | 11.3 | **72.8** |
| Motion Preservation | 2.6 | 31.1 | 6.0 | **60.3** |
| Result Consistency | 2.6 | 9.3 | 12.6 | **75.5** |
| Overall Quality | 2.6 | 9.9 | 8.6 | **78.9** |

Table 3. User study: Percentage of methods rated the best.

## D. Broader Impact

Our work enables the reconstruction of expressive whole-body talking avatars from a single photo, allowing for realistic animations with vivid body gestures and natural expression changes. We consider this a significant advancement in the research and practical applications of multimodal digital humans. However, this technology carries the risk of misuse, such as generating fake videos of individuals to spread false information or harm reputations. We strongly condemn such unethical applications. While it may not be possible to entirely prevent malicious use, we believe that conducting research in an open and transparent manner can help raise public awareness of potential risks. Additionally, we hope our work can inspire further advancements in forgery detection technologies.


Figure 14. Qualitative comparisons with MimicMotion and Animate-X. Compared to pose-guided 2D approaches, our method inherently ensures better 3D consistency while preserving fine details. Additionally, it offers more precise control over foot poses.