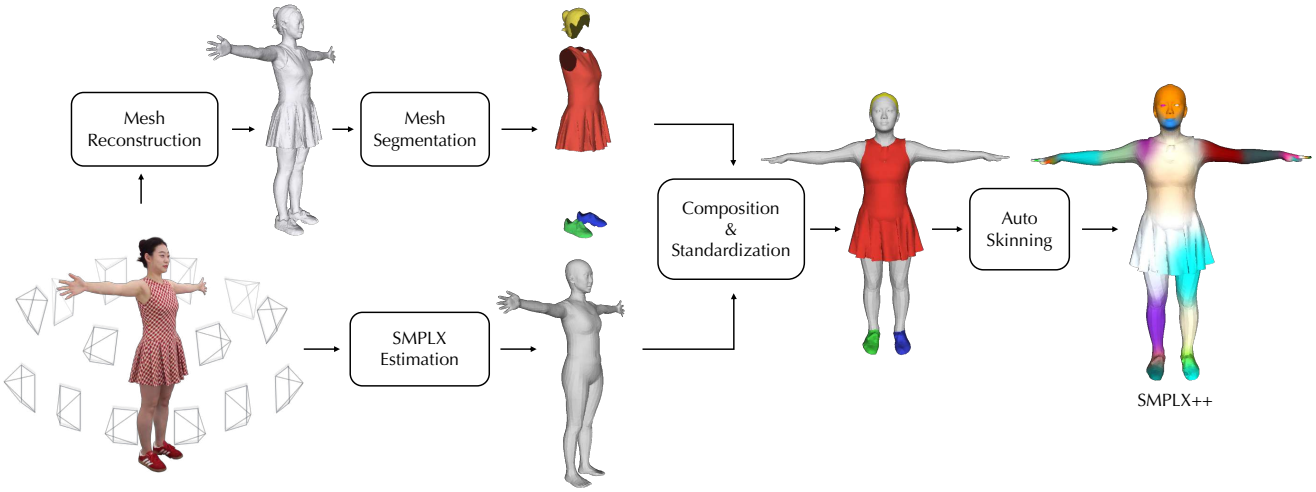

Figure 7. **The Pipeline of the Template SMPLX++ Reconstruction.**

A. Implementation Details

Template Reconstruction. The clothed template SMPLX++ plays a vital role in our approach. As shown in Fig. 7, we create a pipeline to obtain the personalized parametric template SMPLX++ from multi-view images. We choose a frame close to T-pose as a reference, providing more visible details and less sticky geometry and making obtaining accurate SMPLX parameters easier. First, we reconstruct the complete geometry from the multi-view images using NeuS2 [54]. Then we segment and simplify the non-body components such as skirt, shoes, and hair, according to the method proposed in 4D-Dress [53]. However these components are not under the standard T-pose space, we estimate the SMPLX parameters for the reference frame using existing tools [1, 2], and transform them back to T-pose space according to the inverse rigid transformation. Specifically, these skinning weights for non-body parts can be automatically generated by Robust Skinning Transfer [3]. Finally, we combine the naked SMPLX with segmented non-body components to create a personalized complete model SMPLX++. The parametric template

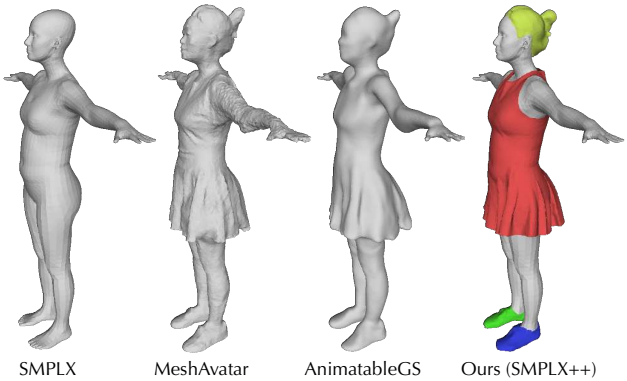

Figure 8. **Template Comparison.**

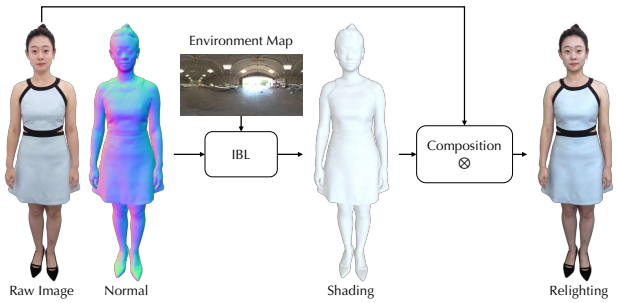

Figure 9. **Relighting Visualisation.**

SMPLX++ can be driven by expression and pose parameters same as the naive SMPLX, which is more expressive for loose clothing geometry. The template contains roughly $23k$ vertices and $45k$ faces, including $20k$ faces for clothes, $2k$ faces for hair, and $2k$ faces for shoes. In contrast to MeshAvatar [10] and AnimatableGS [34], which learn an implicit template from scratch, our template preserves a priori facial expressions and hand gestures as shown in Fig. 8, which are essential for achieving natural and expressive animations.

Network Architecture. We employ a compact MLP-based student network to learn the pose-dependent non-rigid deformation of mesh:

$$\begin{aligned} \mathbf{g}_i &= \varphi\left(\bar{\mathbf{v}}_i\right) \oplus \theta \oplus \mathbf{z}_t \\ \Delta\bar{\mathbf{v}}_i &= \mathcal{S}_c\left(\mathbf{g}_i\right) \cdot m_i + \mathcal{S}_b\left(\mathbf{g}_i\right) \end{aligned} \tag{10}$$

where $\bar{\mathbf{v}}_i \in \mathbb{R}^3$ is the i-th vertex coordinate in the canonical space, $\theta \in \mathbb{R}^{63}$ is the pose parameter, and $\mathbf{z}_t \in \mathbb{R}^{32}$ is a learnable embedding for each frame to compensate for inaccurate pose estimation. The positional encoding function $\varphi(\cdot)$ introduced in NeRF [40], is applied with a frequency band of $L = 6$ in our experiments. The architecture of the student network comprises two specialized MLPs. The first MLP, $\mathcal{S}_b$, models the body's non-rigid deformations, while

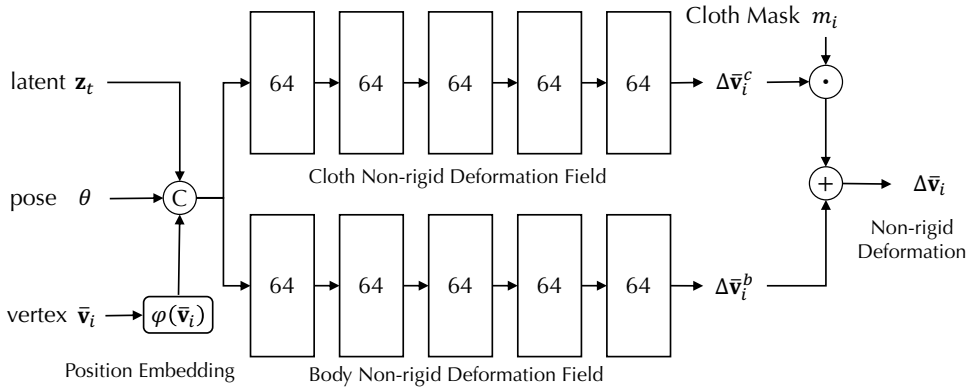

Figure 10. **Network Architecture of Mesh Nonrigid Deformation Field.**

Model	Template	Non-rigid	Gaussian/Face Num.		Quality		Controllability		Speed
			Head	Body	Head	Body	Head	Body	(Inference)
3DGS-Avatar [46]	SMPLX	MLP	19k	181k	low	low	low	low	54
GaussianAvatar [20]	SMPLX	Unet	45k	146k	low	low	low	medium	55
MeshAvatar [10]	Mesh (Implicit)	StyleUnet	5k	50k	low	medium	low	medium	22
AnimatableGS [34]	Mesh (Implicit)	StyleUnet	18k	246k	low	high	low	high	16
Ours (Teacher)	SMPLX++	StyleUnet	19k	250k	medium	high	medium	high	16
Ours (Student)	SMPLX++	MLP+BS	70k	200k	high	high	high	high	156

Table 4. **Summary about these State-of-the-art Methods of Full-body Avatars**.

the second MLP, $\mathcal{S}_c$, captures additional deformations arising from clothing dynamics. To ensure that clothing deformations are applied exclusively to vertices associated with clothing, we introduce a mask $m_i \in \{0, 1\}$, where $m_i = 1$ for the vertices belonging to clothing.

Relighting. We ensure that ambient lighting around the performer is as uniform and white as possible during capture. We use the raw rendered image as the base color and apply shading with new environment light based on the rendered normal map as shown in Fig. 9. Although this approach is not physically accurate, it results in better integration with the environment.

Deployment and 3D Digital Human Agent Pipeline. We make some efforts for mobile deployment, primarily including: a) FP16 quantization for the MLP; b) UInt16 quantization for Gaussian sorting; and c) asynchronous inference techniques, where the animation system operates at 20 FPS (capture frame rate of training data) while the rendering system interpolates animations to render at 90 FPS (maximum screen refresh rate) on the Apple Vision Pro. Please note that all these strategies are not applied on RTX4090 in Tab. 1, which can fundamentally demonstrate the performance of our method. We develop a 3D digital human agent on the Apple Vision Pro, which interacts with users through an ASR-LLM-TTS-Audio2BS pipeline [14, 16, 31, 62] as shown in Fig. 13. Notably, all models run locally after deployment. Please stay tuned for future work with more technical details.

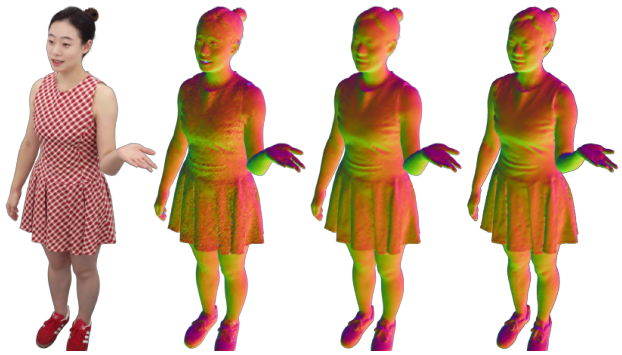

Figure 11. **The Impact of Normal Loss.**

Reanimation. We introduce a learnable embedding $\mathbf{z}_t$ for each frame to better compensate for misalignment issues caused by the inaccurate SMPLX [43] estimation and dynamic changes that cannot be captured by body pose θ (e.g., clothing inertia and swing, changes in hand muscles, etc.). For offline reanimation, we can utilize the Nonrigid Deformation Baking method introduced in the paper to obtain the corresponding $\mathbf{z}_t$ under novel poses from the teacher network. For online real-time body driving, we use the $\mathbf{z}_0$ from the first training frame, which is practically acceptable, al-

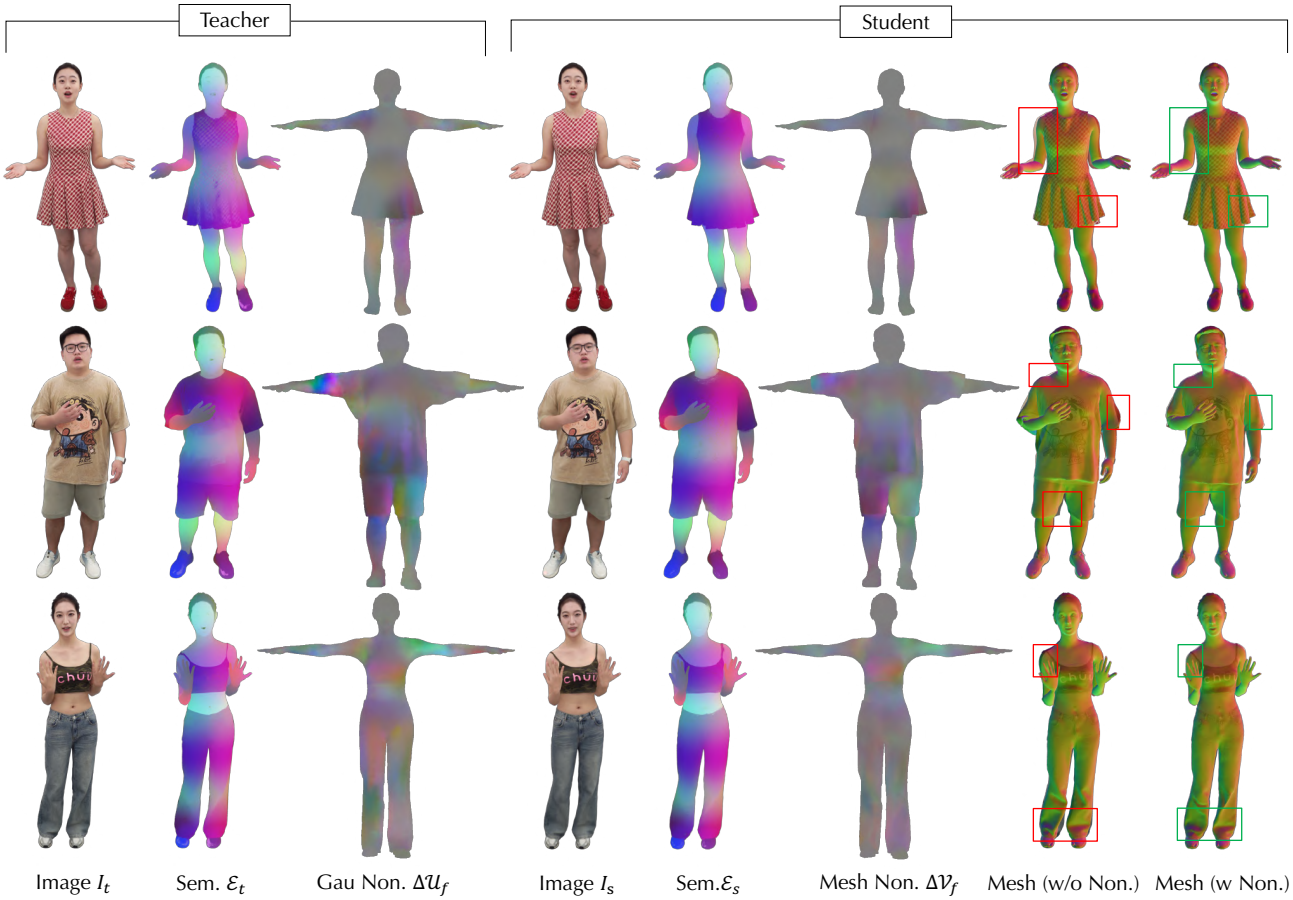

Figure 12. **Qualitative Visualization of Baking.**

though it compromises the accuracy of non-rigid deformations.

B. Experiment Details

Metric Evaluation. To quantitatively evaluate the quality of the rendered images, we choose Peak Signal-to-Noise Ratio (PSNR), Structure Similarity Index Measure (SSIM) [55], and Learned Perceptual Image Patch Similarity (LPIPS) [64]. In our experiments, we evaluate masked images at a resolution of 1500×2000, where the masks are provided by BiRefNet [67]. It is important to note that while PSNR and SSIM are highly sensitive to pixel misalignment, LPIPS demonstrates greater robustness by computing differences in deep feature maps. As illustrated in Fig. 15, the teacher network delivers superior full-body clothing details (better LPIPS scores), while the student network excels in the face region due to a more plausible Gaussian distribution. This discrepancy primarily arises from background residuals introduced by the segmentation network [67] and the teacher network's propensity to omit fine details, such as fragmented hair strands. Additionally, we crop the face region for evaluation based on the projection of the head bounding box. We also adopt point-to-surface distance (P2S) and Chamfer distance (Chamfer) to evaluate the geometry, while the ground truth mesh is generated from NeuS2 [54].

C. Additional Discussion

Discussion w.r.t AnimatableGS. In our teacher-student framework, we utilize AnimatableGS [34] as the teacher network due to its robust capability to model complex pose-dependent non-rigid deformations. Unlike the original AnimatableGS [34], which learns an implicit template from scratch, we adopt the SMPLX++ model as our predefined template. We input semantic positional maps into StyleUnet [52] by assigning distinct color labels to each component (e.g., red for clothing, yellow for hair) and combining these labels with the posed coordinates to generate the final vertex colors. This strategy can provide better semantic information about segmentation for the teacher network. Additionally, we incorporate a normal loss $\mathcal{L}_{nor}$ during the training of the teacher network, which contributes to the

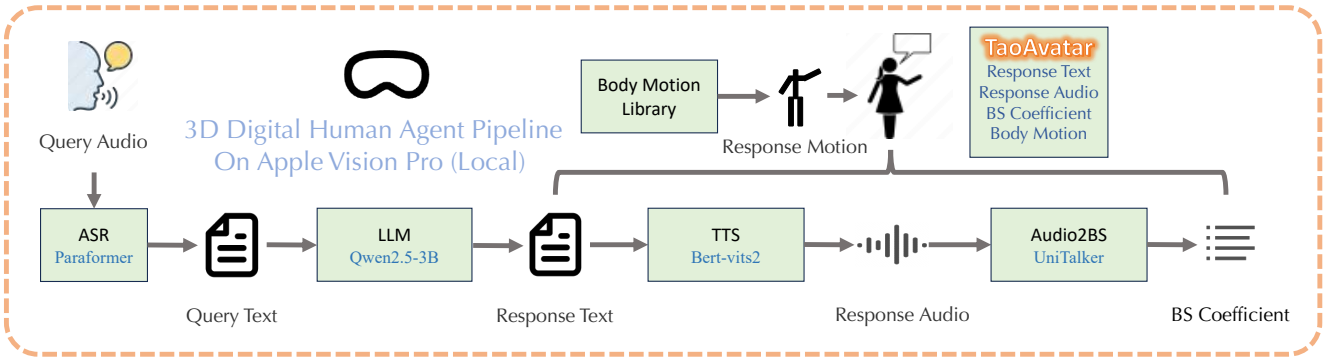

Figure 13. **3D Digital Human Agent Pipeline.**

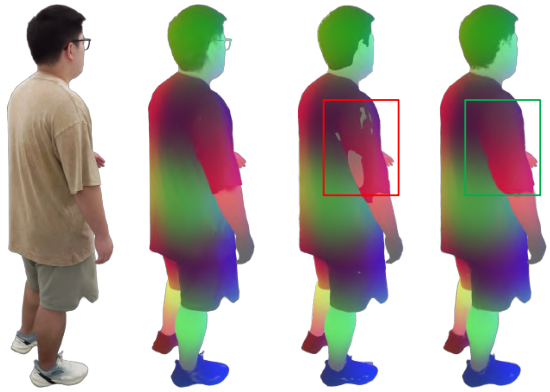

Figure 14. **Ablation Study on Semantic Loss during Baking.**

learning of smoother geometries.

Summary about Full-body Avatars. We present a comparative summary of full-body avatar methods in Tab. 4. 3DGS-Avatar [46] and GaussianAvatar [20] utilize the basic naked SMPLX model as their template, resulting in poor rendering quality for loose clothing. MeshAvatar [10] and AnimatableGS [34] develop implicit clothed templates from scratch which compromising the control over facial expressions and hand gestures. Regarding non-rigid deformation modeling, StyleUnet exhibits more robust expressive capabilities than MLP and Unet, as discussed in AnimatableGS [34]. Our method employs an MLP-based student network baked from the teacher network and two lightweight learnable blend shape compensations. This design enables high-performance rendering with minimal quality degradation. Notably, while maintaining the same number of Gaussians as the teacher model, we allocate more Gaussians to the face to achieve higher facial sharpness as shown in Fig. 17. In contrast, AnimatableGS [34] limits the number of Gaussians on the face due to the resolution constraints of the rendered positional maps.

The Non-rigid Loss and Semantic Loss during Baking. During the non-rigid deformation baking process, both the non-rigid loss $\mathcal{L}_{non}$ and the semantic loss $\mathcal{L}_{sem}$ play crucial roles. For the non-rigid deformation loss $\mathcal{L}_{non}$, we directly use the Gaussian non-rigid deformation maps $\{\Delta\mathcal{U}_f, \Delta\mathcal{U}_b\}$ generated by the teacher network to directly supervise the mesh non-rigid deformation maps $\{\Delta\mathcal{V}_f, \Delta\mathcal{V}_b\}$ of the student network under T-pose in the canonical space. Regarding the semantic loss $\mathcal{L}_{sem}$, we construct a semantic label $\mathbf{e}_i = \mathbf{c}_i + \sin(\tau\bar{\mathbf{v}}_i)$ for each vertex of the template, where τ is a scale factor is designed to increase the frequency of positional changes, inspired by the position embedding in NeRF [40]. As illustrated in Fig. 14, the semantic loss $\mathcal{L}_{sem}$ helps mitigate the intersection between clothing and the body. We provide visualizations of the products from the baking process of different identities in Fig. 12. The teacher network effectively guides learning mesh non-rigid deformation of the student network, resulting in geometry that is well-aligned with the performer's surface as shown in Fig. 12 (Mesh (w/o Non.) vs. Mesh (w Non.)). Without the help of the baking process, it isn't easy to decouple geometry and appearance.

The Impact of Normal Loss. Similar to most 3D Gaussian-based methods [28, 35], we define the direction of the Gaussian's shortest axis as its normal. In contrast to other approaches that dynamically determine the normal orientation based on the camera position, we assign a fixed normal $n = [1, 0, 0]$ and scaling $s = [\epsilon, 1, 1]$ in the local space for each Gaussian, where $\epsilon = 0.01$ is a minimum to make the Gaussian as thin as possible. Upon transforming its parent triangle, the Gaussian's normal in world space aligns with the triangle's normal. To promote smoother rendered normal maps, we introduce a normal loss $\mathcal{L}_{nor}$ as illustrated in Fig. 11. The ground truth normal maps are obtained from NeuS2 [54]. Additionally, the rendered normal maps facilitate image-based relighting, as demonstrated in the provided video demo.

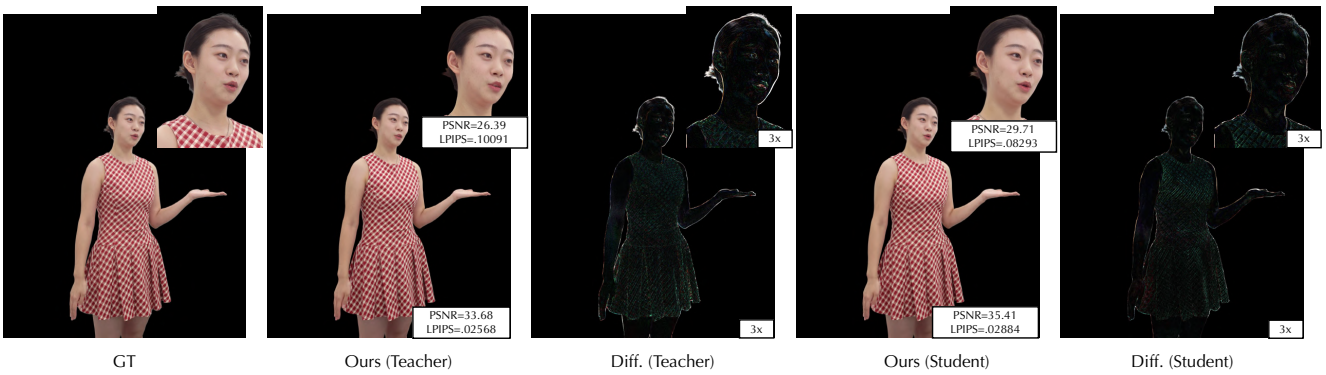

Figure 15. **Qualitative Comparison of Details.**

D. Failure Cases

Although TaoAvatar demonstrates remarkable performance in full-body talking tasks, it still faces challenges in handling complex motions and exaggerated outfits. Specifically, when the teacher network struggles to accurately model loose garments under intricate motions (e.g., dancing-induced skirt motions), the task becomes increasingly difficult for the student network as shown in Fig. 16. Moreover, TaoAvatar is highly reliant on the precision of SMPLX parameters and is susceptible to artifacts when the estimated SMPLX fails to align with the image properly.

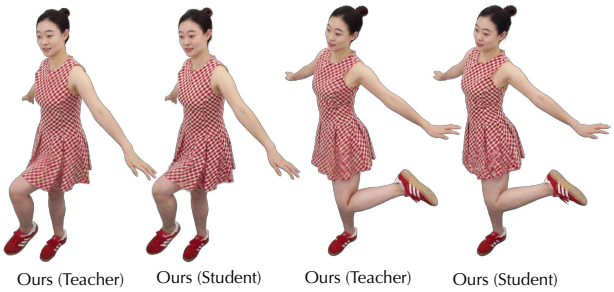

Figure 16. **Failure Cases.**

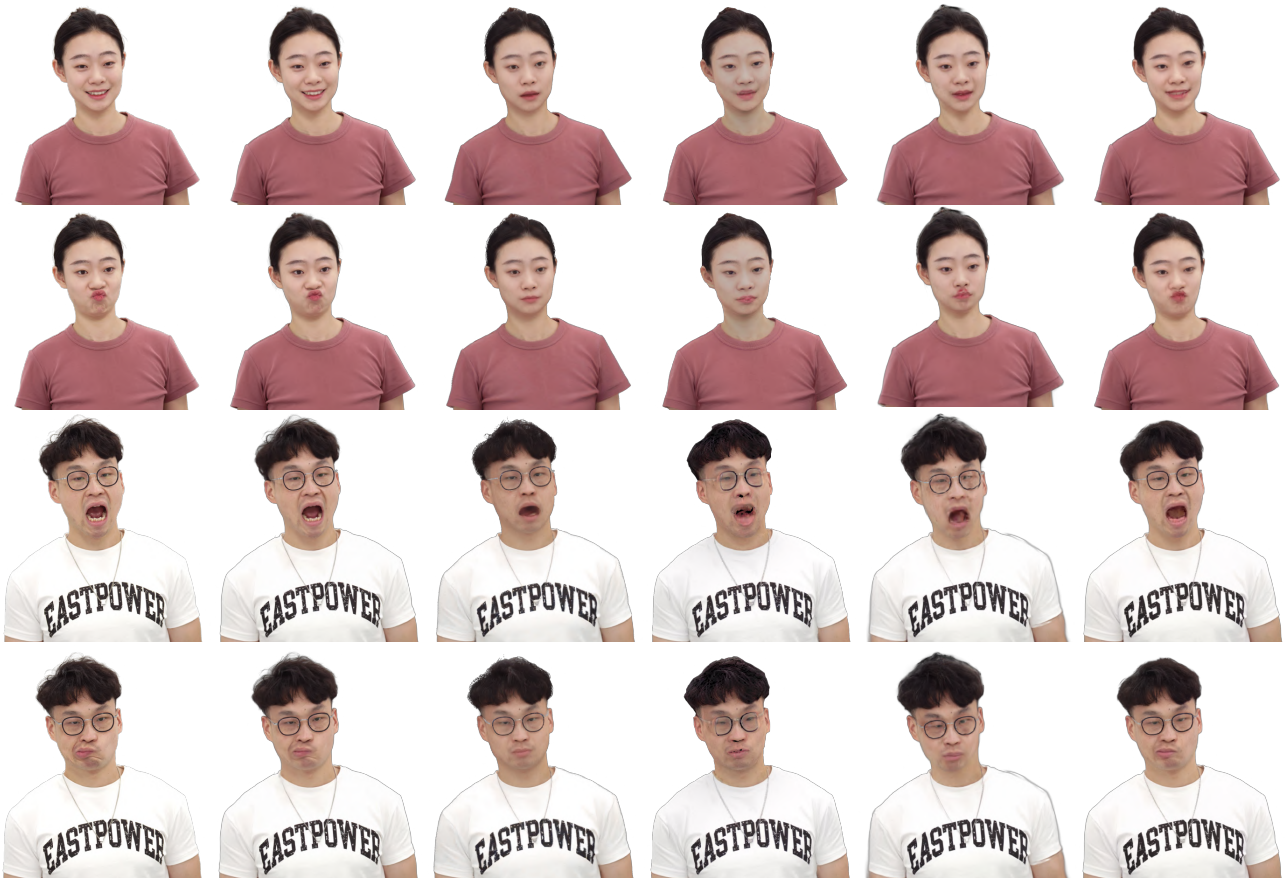

Figure 17. **Qualitative Comparisons on Exaggerated Expression.**