PHOSA: Photorealistic 3D Sign Avatar Modeling and Benchmark

Haodong Wang[1], Hezhen Hu[2*], Wengang Zhou[1*], and Houqiang Li[1]

[1] University of Science and Technology of China
[2] University of Texas at Austin

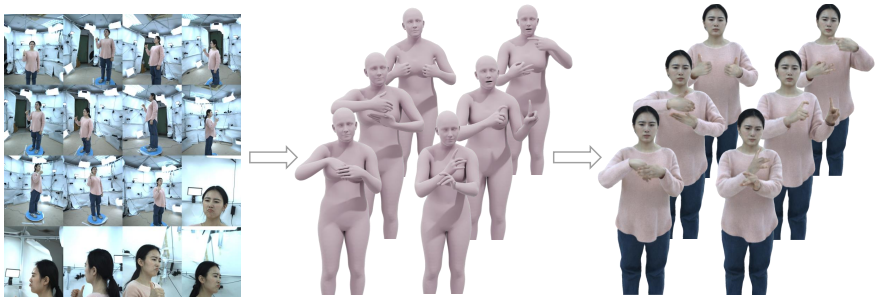

Fig. 1: Overview of **PHOSA** for photorealistic 3D sign avatar modeling. To deal with sign avatar modeling, we present: (1)**MVSign** (left), a multi-view Chinese sign language benchmark captured by 16 synchronized RGB cameras (2048×2448), together with precise hand and facial annotations (middle) produced by our hybrid SMPL-X fitting pipeline; (2): Photorealistic sign avatar (right) generated from MVSign via our decoupled sign avatar representation, showing high fidelity and generalization to complex sign articulations.

Abstract. In this work, we focus on photorealistic sign avatar modeling, which is crucial for effective communication with the Deaf community and is characterized by complex hand gestures and nuanced facial expressions. To this end, we introduce MVSign, the first multi-view Chinese sign language dataset co-designed with Deaf experts, featuring diverse gestures and rich annotations. For precise SMPL-X annotation, we develop a hybrid fitting pipeline that produces accurate body, hand, and facial parameters and can also be applied to the monocular setting. Building on MVSign, we propose a decoupled sign avatar representation that isolates body, head, and hand components to capture complex articulations, together with a motion-aware sampling strategy to handle motion blur and balance gesture diversity. Extensive experiments demonstrate that our method achieves high-fidelity visual results on MVSign, particularly in detailed hand and facial regions, and generalizes well to in-the-wild monocular sign language videos. Project page: https://naaapi.github.io/PHOSA.

Keywords: Sign Language · Human Avatar Modeling · Benchmark

* Corresponding authors.

1 Introduction

Sign Language Production (SLP), which converts spoken language into continuous sign sequences, is crucial for facilitating communication between hearing and Deaf community. While recent SLP works [2, 55, 65] have achieved progress in motion generation, they typically represent signers using parametric body meshes such as SMPL-X [39]. These mesh-based representations lack photorealism and fine-grained expressiveness, failing to deliver the natural, human-like signing experiences that Deaf community strongly prefer [44]. This highlights the need to develop photorealistic and animatable avatars with SMPL-X parameter-driven articulation capabilities.

In this work, we focus on photorealistic and drivable sign avatar modeling. Our goal is to build the dataset, annotation pipeline, and representation needed to faithfully capture expressive signing. Achieving this goal is challenging for two main reasons. First, suitable datasets are lacking. Existing multi-view human datasets [6, 22, 41, 53] mainly capture general body motion and overlook the detailed hand articulations and facial expressions essential for sign communication. Conversely, current sign language datasets [4, 9, 55] are designed for motion production, without the multi-view imagery or precise annotations needed for human avatar modeling. Second, modeling fine-grained articulations is difficult. Prior avatar modeling works [20, 29, 30, 47] target general human motions and struggle with modeling on complex hand gestures and facial expressions.

To overcome these challenges, we introduce MVSign (see Figure 1), the first multi-view Chinese sign language dataset co-designed with Deaf experts and collected under IRB approval. It covers diverse gestures spanning basic hand shapes and complex motion patterns, supporting generalization to unseen signs. Each sample includes rich annotations, *e.g.,* image matting, body-part segmentation, 3D keypoints, and SMPL-X parameters. To ensure accurate recovery of expressive details, we develop a hybrid SMPL-X fitting pipeline that integrates predictions from multiple state-of-the-art models, achieving precise hand articulation and nuanced facial motion. The pipeline can also serve as a plug-in for monocular RGB settings.

To deal with fine-grained articulation during sign avatar modeling, we further develop an efficient decoupled sign avatar representation that separates the whole-body Gaussian maps into body, head, and hands. By introducing partial kinematic decoupling, body joints are fixed while hand mobility is preserved, making hand pose maps independent of body motion and improving gesture generalization. Furthermore, a motion-aware frame sampling strategy filters motion-blurred frames and balances diverse motion types, enhancing visual fidelity and training stability. Extensive experiments demonstrate that our method achieves state-of-the-art performance on MVSign and in-the-wild sign videos from the Web.

In summary, our main contributions are threefold:

- We systematically study photorealistic sign avatar modeling and establish a multi-view benchmark to facilitate evaluation and comparison.
- We present MVSign, the first multi-view Chinese sign language dataset co-designed with Deaf experts, featuring diverse gestures and rich annotations.

To ensure annotation accuracy, we develop a hybrid SMPL-X fitting pipeline to achieve precise hand and facial parameters.
– We propose a decoupled avatar representation that separates body, head, and hands, and employs partial kinematic decoupling for improved gesture generalization and fine-grained articulation fidelity.

2 Related Work

2.1 Animatable Human Avatar Modeling

Model animatable human avatar from RGB video is a long-standing and challenging problem. Early work [16, 18, 22–24, 32, 49, 51, 60] uses implicit neural representations such as NeRF [34] or SDF to model human avatar. However, they struggle to balance high-quality results with fast rendering. Recently, 3D Gaussian Splatting [26] has made substantial progress in improving both training and rendering times over traditional NeRFs while preserving high quality, inspiring many works [19, 20, 29, 30, 36, 38, 63] to adopt it as a foundational representation. GART [29] uses a mixture of moving 3D Gaussians to explicitly approximate the geometry and appearance of deformable subjects, leveraging categorical template models with learnable forward skinning. EVA [19] introduce a plug-and-play module that significantly ameliorates SMPL-X misalignment issues, while a context-aware adaptive density control strategy is applied to accommodate the varied granularity across body parts. AnimatableGaussians [30] learns a parametric template from the input videos, and then parameterize the template on canonical Gaussian maps where each pixel represents a 3D Gaussian. Such template-guided 2D parameterization enables they to employ a powerful StyleGAN [25]-based CNN to learn the pose-dependent Gaussian maps for modeling detailed dynamic appearances.

Different from previous works, which mainly focus on general human motions, our network introduces an efficient decoupled sign avatar representation specifically designed to capture fine-grained hand and facial details essential for photorealistic sign avatar modeling.

2.2 Human-centric and Sign Language Benchmarks

Human-centric benchmarks provide the foundation for learning and evaluating photorealistic, animatable human representations. Recent multi-view human datasets, such as DNA-Rendering [6], HumanRF [22], ZJU-MoCap [41], and MVHumanNet [53], have advanced high-fidelity novel-view synthesis and animatable avatar modeling. However, they mainly focus on general human motions and do not specifically target the complex hand articulations and subtle facial expressions required for sign communication.

In parallel, existing sign language datasets mainly focus on recognition, translation, or motion production. For example, PHOENIX14T [4], CSL-Daily [62] and How2Sign [9] provide valuable video-language annotations for sign understanding, while SMILE [10] targets lexical-level sign recognition and assessment.

Table 1: Dataset statistics comparison with existing multi-view human-centric datasets and sign language datasets.

Dataset	Sign	Head	SMPL-X	#View	#ID	#Frames	Resolution
PHOENIX14T [4]	✔	✗	✗	1	9	0.94M	260P
CSL-Daily [62]	✔	✗	✗	1	10	1.5M	512P
How2Sign [9]	✔	✗	✗	2	11	5.7M	720P
SignAvatars [55]	✔	✗	✔	1	153	8.34M	-
SMILE [10]	✔	✗	✗	4	66	-	1080P
VSL*[42]	✔	✗	✗	16	2	50K	4096P
SGNify*[15]	✔	✗	✔	12	3	-	-
Human3.6M [21]	✗	✗	✔	4	11	3.6M	1000P
NHR [52]	✗	✗	✔	80	3	100K	2048P
ZJU-MoCap [41]	✗	✗	✔	24	10	180K	1024P
THuman 4.0 [61]	✗	✗	✔	24	3	10K	1150P
MVSign (Ours)	✔	✔	✔	16	5	115K	2048P

* The dataset is not publicly available.

SignAvatars [55] further introduces large-scale 3D motion annotations for sign language production. Nevertheless, these datasets are not designed for photorealistic sign avatar reconstruction. Most lack synchronized multi-view imagery and annotations required by human avatar modeling benchmarks.

This gap motivates our MVSign benchmark, which provides multi-view sign videos together with rich annotations and a dedicated annotation pipeline, enabling systematic evaluation of photorealistic sign avatar modeling.

3 MVSign Dataset

3.1 Data Capture

Data Collection Protocol. As described by the Hamburg Notation System (HamNoSys) [17], sign language consists of several fundamental components, including hand shapes, movements, orientations, locations, and facial expressions. To ensure that MVSign covers diverse sign articulations, we collaborated with Deaf experts to carefully design the signing actions. First, we adopt 109 basic hand shapes, consisting of 26 letters, 22 counting numbers and 61 commonly used words from the reference book Chinese Sign Language Tutorial [37], which is designed for daily sign-language learning and communication. These words represent the most frequently used signs in daily communication, forming a strong foundation for basic hand shapes. Next, to provide diverse hand motion patterns, facial expressions, and natural transitions between consecutive signs, we curated daily-use sign language sentences from the beginner and intermediate volumes of the Chinese Sign Language Tutorial. The books cover 20 common daily-life scenarios, including greetings, study, family, hobbies, weather, shopping, travel, and medical services. Each scenario contains approximately 10 sentences. For

each recorded signer, we randomly selected 5 scenarios from the 20 scenarios and used the corresponding sentences, resulting in around 50 distinct sentences per signer. This collection protocol ensures that MVSign captures both fundamental hand shapes and complex dynamic gestures, thereby enhancing the generalization ability of avatars to unseen sign poses.

System Setup. Our capture system consists of a high-fidelity array of 16 synchronized RGB cameras, each recording at a resolution of 2048×2448 pixels and 25 frames per second. The camera layout is carefully designed to optimize the modeling of sign avatars by balancing body coverage and fine-grained detail. Specifically, one camera captures the frontal view, ten cameras are distributed along the lateral arc to capture multi-angle side perspectives and detailed hand and arm articulations, while the remaining five focus on the head region to preserve fine facial expressions and mouthing. All cameras are temporally synchronized to ensure consistent multi-view capture, and the entire system is calibrated using VGGSfM [48].

Dataset Statistics and Comparison. MVSign includes five native Chinese Sign Language signers (two males, three females) with high fluency and expressivity. Each actor contributed approximately 15 minutes of recording, totaling around 23,000 frames, with 10 minutes on basic hand shapes and 5 minutes on diverse sentences. The detailed comparison between MVSign and other relevant datasets are shown in Table 1. Most existing sign language datasets [4, 9, 55, 62] used for production are monocular, short in duration, and relatively low in visual quality, which limits their suitability for benchmarking multi-view avatar modeling. In contrast, current multi-view human datasets [6, 22, 41, 53] primarily focus on general body motion and do not emphasize the fine-grained hand articulations and facial expressions that are critical for sign language. MVSign is the only dataset that simultaneously provides multi-view sign language data, dedicated head portrait views, and SMPL-X parameters annotation, while also maintaining high resolution and a comparable number of frames.

Ethics Considerations. This project received IRB approval, and all participants provided informed written consent for public release of anonymized data. The IRB documentation will be released alongside the dataset.

3.2 Data Annotation

To facilitate sign avatar modeling, our dataset contains comprehensive annotations along with the raw data, including image matting, body part segmentation, 3D keypoints, and SMPL-X parameters.

Matting and Body Part Segmentation. We utilize the state-of-the-art human foundation model Sapiens [27] to extract body-part segmentation on the raw images. However, we observe that the segmented hand and hair regions often contain extraneous background areas. To address this issue, we further apply the SAM model [28] to obtain refined human matting and remove unwanted background artifacts, ensuring clean and accurate body-part segmentation.

Keypoints and SMPL-X Parameters. Current whole-body pose estimation [5, 7, 54] and mesh recovery [3, 31, 35] methods struggle to accurately capture complex

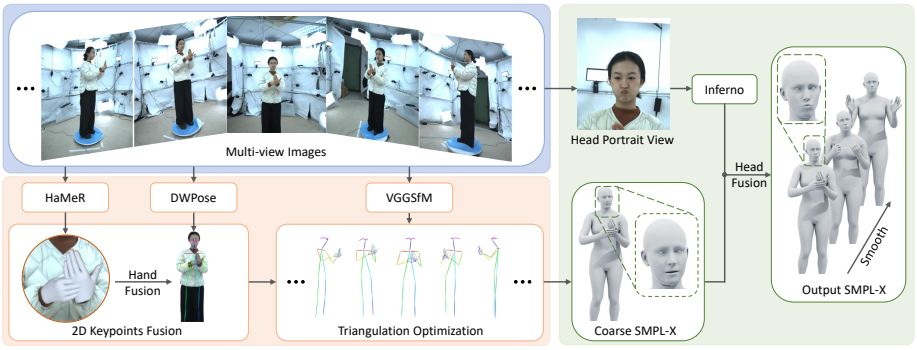

Fig. 2: Overview of the hybrid SMPL-X fitting pipeline. Our pipeline fuses outputs from multiple models to leverage their complementary strengths, achieving precise and robust SMPL-X parameter estimation.

sign language gestures. This motivates us to propose a *hybrid SMPL-X fitting procedure* that integrates outputs from multiple state-of-the-art models. By leveraging the strengths of each model, our approach achieves precise 3D keypoint localization and SMPL-X parameter estimation. The full pipeline is shown in Figure 2.

Given synchronized multi-view images $\mathcal{I} = \{I_1, ..., I_N\}$ with N calibrated cameras. We first apply DWPose [54] for each view to detect 2D keypoints J^{2D}, including keypoints of the body, hands and face. To improve hand mesh precision, we subsequently process each view through HaMeR [40], obtaining more robust MANO parameters [45] and corresponding hand keypoints J_h^{2D}, which replace the original DWPose hand estimations and produce optimized whole-body keypoints $\hat{J}^{2D}$. Leveraging calibrated camera intrinsics $\mathbf{K}$ and extrinsics $[\mathbf{R}|\mathbf{t}]$, we obtain precise 3D keypoints $\hat{J}^{3D}$ by applying multi-view geometric constraints via triangulation optimization [1]:

$$\hat{J}^{3D} = \arg\min_{J^{3D}} \sum_{i=1}^{N} \left\| (\mathbf{K}_i[\mathbf{R}_i|\mathbf{t}_i]J^{3D}) - \hat{J}_i^{2D} \right\|_2 . \tag{1}$$

Then we fit SMPL-X parameters by minimizing the following objective:

$$\mathcal{L}(\theta, \beta, \psi) = \sum_{i \in \mathcal{J}} \gamma_i \omega_i \rho \left(R_\theta(J(\beta)_i) - \hat{J}_i^{3D} \right) + \lambda_\beta \left\| \beta \right\|_2 + \lambda_\psi \left\| \psi \right\|_2 , \tag{2}$$

where $\theta \in \mathbb{R}^{165}$ denotes the full body pose parameters, $\beta \in \mathbb{R}^{10}$ refers to body shape parameters, $\psi \in \mathbb{R}^{10}$ is the facial expression parameters, $\rho(\cdot)$ denotes the Geman-McClure robust kernel which helps prevent the disturbance from noisy supervision signals, $R_\theta(\cdot)$ denotes the function which rotates the $J(\beta)$ given the pose θ, γ and ω are predefined joint weights and confidence weights from detection.

Since the face keypoints are too sparse to capture nuanced facial articulations, the fitted facial parameters often fail to accurately represent the true expression.

To address this, we utilize INFERNO [8, 13, 14, 64] to extract more precise and expressive facial expression parameters and replace the initial facial parameters derived from skeletal fitting.

We further enforces temporal consistency of the SMPL-X parameters through two strategies: (1) employing temporal continuity constraints by initializing each frame's optimization with the preceding frame's converged parameters, and (2) applying SmoothNet [56] to the resulting SMPL-X sequence to mitigate high-frequency artifacts.

4 Sign Avatar Modeling

To improve the quality of sign avatar modeling, we approach the problem from both data and methodological perspectives. From the data perspective, we design a motion-aware sampling strategy that filters out motion-blurred frames and balances the distribution of gestures. On the methodological side, as shown in Figure 3, we introduce a decoupled sign avatar representation that effectively addresses the topological complexity in hand articulation, leading to improved visual quality.

4.1 Motion-aware Data Sampling Strategy

Motion blur is common in sign language videos, where hand movements are often very fast. Despite increasing the recording frame rate, we observe that motion blur still persists in certain parts of the dataset. Additionally, we identify an imbalance in the types of sign articulations within the dataset. A large portion of the frames contain hand in a stationary, hanging position. This imbalance induces model convergence to suboptimal local minima, substantially compromising the fidelity of hand regions. To address these issues, we design a Motion-aware Data Sampling Strategy that filters out motion-blurred frames while also balancing the distribution of sign articulations. Please refer to the technical appendices for formalized algorithmic workflow.

We first detect motion-blurred frames using three indicators: 1) Hand confidence score from DWPose detection, providing prior information from the pose estimation model. 2) Laplacian gradient of the hand image, which quantifies local texture variations. 3) Hand motion velocity derived from SMPL-X parameter trajectories, serving as a physical constraint. We define thresholds for each indicator and filter out frames that do not meet the criteria set by these indicators.

After filtering out motion-blurred frames, we calculate the distance between each sign gesture based on the SMPL-X parameters. Gestures that are close in distance will be grouped into the same category. We then remove excessive gestures from each class, balancing the number of frames within each sign gesture category.

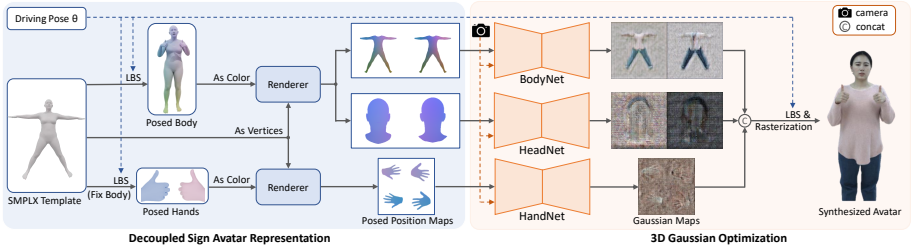

Fig. 3: Overview of sign avatar modeling pipeline. The posed vertices are taken as vertex color on the canonical SMPL-X template, and then render to generate posed position maps. For the hand part, we employ partial kinematic decoupling by fixing body joints, only preserving hand joint mobility. We then predict pose-dependent Gaussian maps through three specialized StyleUNet, deform the Gaussians by LBS, and render the synthesized avatar by differentiable rasterization.

4.2 Decoupled Sign Avatar Representation

In sign language, subtle hand gestures and facial expressions are crucial for conveying meaning. Previous avatar modeling methods [20, 30, 57] typically represent the entire body holistically, which forces the model to learn large-scale body movements together with fine-grained hand and facial motions. This joint representation often biases the network toward coarse motion patterns while overlooking the dynamics of the hands and face.

To address this limitation, we propose a decoupled representation that explicitly decompose the whole-body Gaussians into three anatomically distinct components: body, hands, and head. This decomposition enables the model to focus more effectively on each specific region, thereby improving the viusal quality. Specifically, following [30], we adopt StyleUNet [49] as our backbone to predict pose-dependent Gaussian attributes in canonical space. To ensure compatibility with 2D networks, the 3D representation of the human avatar needs to be parameterized in 2D space. Given a driving pose $\boldsymbol{\Theta}$, we deform the body and head part through linear blend skinning (LBS). Then we take the posed coordinate as the vertex color on the canonical SMPL-X template, and render it to both front and back views to generates posed position maps $\mathcal{P}$:

$$\begin{aligned} \mathcal{P}^{\mathrm{f}}_{\mathrm{body}}, \mathcal{P}^{\mathrm{b}}_{\mathrm{body}} &= \mathcal{R}(\mathcal{M}_{\mathrm{body}}, \Psi_{\mathrm{LBS}}(\boldsymbol{\Theta}, \mathcal{M}_{\mathrm{body}})), \\ \mathcal{P}^{\mathrm{f}}_{\mathrm{head}}, \mathcal{P}^{\mathrm{b}}_{\mathrm{head}} &= \mathcal{R}(\mathcal{M}_{\mathrm{head}}, \Psi_{\mathrm{LBS}}(\boldsymbol{\Theta}, \mathcal{M}_{\mathrm{head}})), \end{aligned} \tag{3}$$

where $\mathcal{R}(\mathcal{M}, \mathcal{C})$ denotes the render process of mesh $\mathcal{M}$ with color $\mathcal{C}$, Ψ_{LBS} denotes the LBS deformation operator, $\mathcal{M}_{\mathrm{body}}$ and $\mathcal{M}_{\mathrm{head}}$ represents the body and head part template mesh.

For hand parameterizations, we further employ partial kinematic decoupling by fixing body joints while preserving hand joint mobility. This isolation ensures that the hand pose maps are only related to hand gestures and are independent of body poses, thus improving the generalization of hand modeling. We render

the two hands to both up and down views and concatenate them together:

$$\mathcal{P}_{\text{hand}} = \mathcal{R}(\mathcal{M}_{\text{hand}}, \Psi_{\text{LBS}}(\mathbf{\Theta}_{\text{hand}}, \mathcal{M}_{\text{hand}})), \tag{4}$$

where $\mathbf{\Theta}_{\text{hand}}$ contains hand-specific pose parameters, with the body fixed in its canonical pose.

4.3 Optimization

Our multi-branch architecture employs three specialized StyleUNet modules $\mathcal{F}_{\text{body}}$, $\mathcal{F}_{\text{head}}$ and $\mathcal{F}_{\text{hand}}$ for independent Gaussian attributes prediction:

$$\begin{aligned} \mathcal{G}^{\text{f}}_{\text{body}}, \mathcal{G}^{\text{b}}_{\text{body}} &= \mathcal{F}_{\text{body}}(\mathcal{P}^{\text{f}}_{\text{body}}, \mathcal{P}^{\text{b}}_{\text{body}}, \mathcal{V}), \\ \mathcal{G}^{\text{f}}_{\text{head}}, \mathcal{G}^{\text{b}}_{\text{head}} &= \mathcal{F}_{\text{head}}(\mathcal{P}^{\text{f}}_{\text{head}}, \mathcal{P}^{\text{b}}_{\text{head}}, \mathcal{V}), \\ \mathcal{G}_{\text{hand}} &= \mathcal{F}_{\text{hand}}(\mathcal{P}_{\text{hand}}, \mathcal{V}), \end{aligned} \tag{5}$$

where $\mathcal{G}_*$ denotes predicted Gaussian attribute maps (position μ, rotation r, scale s, opacity α, color c) in canonical sapce, and $\mathcal{V}$ encodes view-dependent appearance variations. To ensure that the position attribute of the predicted Gaussian maps closely aligns with the SMPL-X human mesh, we predict position offsets $\Delta\mathcal{P}$ relative to SMPL-X mesh rather than absolute positions.

We then employ LBS to deform the whole body Gaussians from canonical space to observation space. A canonical 3D Gaussian's position μ_c and rotation μ_c are transformed as follows:

$$\mu_o = \mathbf{R} \cdot \mu_c + \mathbf{T}, \quad r_o = \mathbf{R} \cdot r_c, \tag{6}$$

where μ_o and r_o are 3D Gaussian's position and rotation in observation space, $\mathbf{R}$ is the rotation matrix and $\mathbf{T}$ is the translation vector, both of which are calculated with the skinning weights of each 3D Gaussian and driving pose. Finally, we render the posed 3D Gaussians to a desired camera view through splatting-based rasterization [26].

Optimization Objectives. The composite loss function integrates multiple constraints:

$$\begin{aligned} \mathcal{L} = &\underbrace{\mathcal{L}_1 + \lambda_{\text{SSIM}}\mathcal{L}_{\text{SSIM}} + \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}}}_{\text{Photometric}} \\ &+ \underbrace{\lambda_{\text{hand}}\mathcal{L}_{\text{hand}} + \lambda_{\text{head}}\mathcal{L}_{\text{head}}}_{\text{Anatomical Focus}} + \underbrace{\lambda_{\text{offset}}\|\Delta\mathcal{P}\|_2}_{\text{Regularization}}, \end{aligned} \tag{7}$$

where $\mathcal{L}_1$, $\mathcal{L}_{\text{SSIM}}$, and $\mathcal{L}_{\text{LPIPS}}$ are the L1, SSIM [50], and LPIPS [59] losses, respectively. $\mathcal{L}_{\text{hand}}$ is the L1 loss on segmented hand regions, encouraging the model to focus on hand details. Similarly, $\mathcal{L}_{\text{head}}$ applies L1 loss on the head region. $\mathcal{L}_{\text{offset}}$ is the L2 norm of position offsets, preventing Gaussians from deviating too far from the template.

5 Experiments

5.1 Experimental Setting

Datasets and Metrics. We conduct experiments on our MVSign dataset and a set of collected monocular in-the-wild sign language videos. The in-the-wild videos are collected from the Web with the Creative Commons (CC) license. They have four individuals (two males and two females) and are recorded at a resolution of 1920×1080. To extract SMPL-X parameters, we apply our hybrid fitting procedure, excluding the 3D triangulation step due to the monocular setting. To evaluate the results quantitatively, we adopt three commonly-used metrics, including PSNR, SSIM [50], and LPIPS [59].

Dataset Split. For the MVSign dataset split, we use the first 90% of the sampled frames as the training set and the remaining frames as the test set. We use all the 16 views for training. For in-the-wild sign videos, we uniformly sample the frames with the interval as 1 to split the training and testing frames. The number of training and testing frames are both 150.

Baselines. We compare our method with the-state-the-art human avatar modeling methods, including SplattingAvatar [47], GaussianAvatar [20], Animatable-Gaussians [30], EVA [19] and Mmlphuman [57], implemented by the official code. Among these methods, EVA focuses on expressive avatar modeling and explicitly accounts for sign language. Note that recent text-to-video generators for sign language (e.g., SignGen [43], SignDiffs [11]) follow a 2D image synthesis paradigm with different inputs and do not model a 3D drivable avatar, thus they are not directly comparable. Since some of the baselines [20, 47] are animated with SMPL [33] which has no control over the hands and facial expressions. We replace the driven model with SMPL-X [39] and increase the pose-conditioning dimension accordingly, while keeping network, losses, training schedule, and hyperparameters unchanged. For fairness, all methods use the same train/test split, camera views, images, fitted SMPL-X supervision, and masks.

Implementation Details. Our framework is implemented with PyTorch and all experiments are performed on a single NVIDIA 3090 GPU. We set the optimization hyperparameters $\lambda_{\text{SSIM}} = 0.1$, $\lambda_{\text{LPIPS}} = 0.1$, $\lambda_{\text{hand}} = 3$, $\lambda_{\text{head}} = 3$ and $\lambda_{\text{offset}} = 0.005$. The resolution of the posed position maps for body, hand and head part are 1024×512, 256×256 and 256×128, respectively. Please refer to the supplementary material for more details.

5.2 Dataset Analysis

Unlike previous multi-view human-centric [6,22,41,53] or sign language datasets [4, 9, 55], MVSign captures both multi-view imagery and diverse sign gestures, which are crucial for learning expressive and generalizable sign avatars. As shown in Table 4 and Figure 4, increasing the number of views enhances visual quality, especially for fine-grained hand details. This demonstrates that multi-view supervision is essential for resolving self-occlusions and depth ambiguities that are inherent in complex hand gestures. We further analyze the effect of different

Table 2: Ablation study of hybrid SMPL-X fitting strategy. "↓" indicates that lower values are better, while "↑" means the opposite.

Method	Full			Hand			Face		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
w/o Hybrid Fitting	23.58	0.9637	0.0618	16.82	0.7371	0.2987	16.62	0.7923	0.2435
w Hybrid Fitting	**26.90**	**0.9722**	**0.0370**	**18.85**	**0.7704**	**0.2301**	**20.51**	**0.8499**	**0.1665**

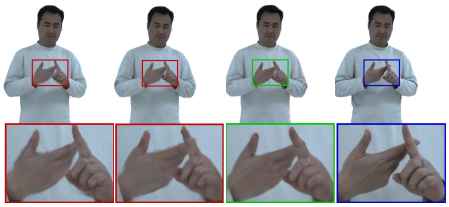

Fig. 4: Visualization of animated avatar trained with different numbers of views.

Table 3: Comparison with other SMPL-X estimation methods on the SGNify [15] mocap dataset. "↓" indicates that lower values are better.

Method	Body↓	Left Hand↓	Right Hand↓
FrankMoCap [46]	78.07	20.47	19.62
PIXIE [12]	60.11	25.02	22.42
PyMAF-X [58]	68.61	21.46	19.19
SMPLify-X [39]	56.07	22.23	18.83
SGNify [15]	55.63	19.22	17.50
NSA [2]	46.42	16.17	15.23
Ours	**40.38**	**14.11**	**13.79**

signing patterns in MVSign. As shown in Table 5, training on only the basic hand shapes subset limits generalization due to the lack of diverse movement patterns. Conversely, training solely on sign sentences makes the learning process more challenging, leading to degraded quality. Combining both subsets resulting in a more robust and realistic sign avatar. These experiments highlight the necessity of MVSign multi-view and diverse sign articulations, distinguishing it from existing sign language datasets.

5.3 Ablation Study

Effectiveness of hybrid SMPL-X fitting. We evaluate our hybrid SMPL-X fitting strategy from two perspectives: accuracy and its effect on modeling. To assess accuracy, we compare our pipeline against prior methods, including SGNify [15] and Neural Sign Actors (NSA) [2], both designed for SMPL-X estimation in sign language contexts. The evaluation is conducted on the SGNify mocap dataset following its official protocol, measuring the mean per-vertex error (mm) on the upper body, left hand, and right hand. As shown in Table 3, our method achieves the lowest error, particularly for the hands, demonstrating its superior ability to capture precise hand motions. To evaluate its effect on avatar modeling, we compare models trained using SMPL-X parameters fitted from DWPose 2D keypoints versus our hybrid pipeline. As shown in Table 2, directly fitting from DWPose 2D keypoints fails to capture complex hand motions, whereas our hybrid fitting provides accurate hand poses, resulting in fine-grained hand details.

Table 4: Impact of the number of views on MVSign dataset. "↓" indicates that lower values are better, while "↑" means the opposite.

Method	Full			Hand			Face		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
1	26.17	0.9711	0.0425	18.49	0.7599	0.2581	18.84	0.8223	0.1927
8	26.42	0.9711	0.0417	18.60	0.7649	0.2524	19.43	0.8286	0.1759
Full (16)	**26.90**	**0.9722**	**0.0370**	**18.85**	**0.7704**	**0.2301**	**20.51**	**0.8499**	**0.1665**

Table 5: Impact of incorporated signing patterns on MVSign dataset. "↓" indicates that lower values are better, while "↑" means the opposite.

Method	Full			Hand			Face		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Only Basic Hand Shapes	24.08	0.9660	0.0524	17.47	0.7425	0.2892	17.55	0.8034	0.2287
Only Sign Sentences	24.36	0.9653	0.0501	17.62	0.7518	0.2768	17.76	0.8122	0.2139
Ours	**26.90**	**0.9722**	**0.0370**	**18.85**	**0.7704**	**0.2301**	**20.51**	**0.8499**	**0.1665**

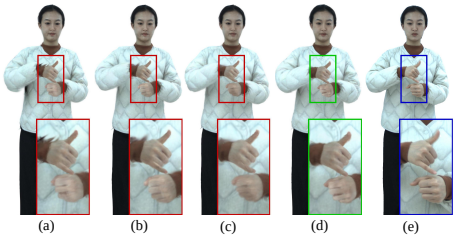

Fig. 5: Ablation study of sampling strategy. (a) Sequential sampling; (b) Isometric sampling; (c) Random sampling; (d) Ours motion-aware sampling; (e) Ground truth image.

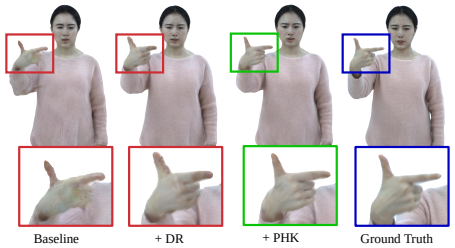

Fig. 6: Visualization of animated avatar trained with different representation. "DR" indicates the decoupled representation, while "PHK" means Partial Hand Kinematic.

Effect of the Data Sampling Strategy. We evaluate four sampling strategies: sequential, isometric, random, and our proposed Motion-aware Data Sampling Strategy to assess how frame selection affects modeling. As shown in Table 6 and Figure 5, filtering out invalid frames using our sampling strategy leading to improved visual quality. In contrast, other strategies often include motion-blurred frames, and overfitting to these low-quality frames degrades the accuracy of fine-grained hand details.

Effects of Decoupled Design. Starting from a single Gaussian map baseline [30], we evaluate decoupling the representation into body, head, and hands, and further applying partial kinematic decoupling for hands. As shown in Table 7 and Figure 6, decoupling alone gives minor gains, whereas incorporating partial kinematic decoupling significantly enhances hand quality, underscoring the importance of fully decoupling hand representations.

Table 6: Ablation study of data sampling strategy. "↓" indicates that lower values are better, while "↑" means the opposite.

Strategy	Full			Hand			Face		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Sequential	26.55	0.9659	0.0372	16.40	0.6734	0.2656	18.66	0.7733	0.1875
Isometric	26.47	0.9659	0.0371	16.29	0.6697	0.2686	18.57	0.7726	0.1906
Random	26.90	0.9663	0.0353	17.40	0.6939	0.2490	18.94	0.7838	0.1784
Motion-aware	**27.67**	**0.9684**	**0.0340**	**18.03**	**0.7161**	**0.2387**	**19.79**	**0.8029**	**0.1672**

Table 7: Ablation study of decoupled sign avatar representation. "DR" indicates the decoupled representation, while "PHK" means Partial Hand Kinematic. "↓" indicates that lower values are better, while "↑" means the opposite.

Method	Full			Hand			Face		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Baseline	25.24	0.9570	0.0490	16.65	0.6912	0.3759	18.47	0.7948	0.2167
+ DR	26.28	0.9607	0.0430	17.53	0.7174	0.3188	20.53	0.8099	0.1731
+ PHK	**27.93**	**0.9665**	**0.0388**	**18.74**	**0.7322**	**0.2666**	**21.50**	**0.8253**	**0.1595**

5.4 Comparison with State-of-the-art Methods

Quantitative and Qualitative Comparison. As shown in Table 10, our method outperforms the baselines across all evaluation metrics. Specifically, on the MVSign dataset, it achieves an improvement of 8.1%, 4.0% and 5.9% relative LPIPS gain on the full, hand and face regions, respectively. On the in-the-wild sign videos, the performance gains are more pronounced, with 15.8%, 11.6% and 12.6% relative LPIPS improvements on the full, hand and face regions. These results demonstrate that our method not only performs strongly on the MVSign dataset but also generalizes effectively to complex real-world sign language videos. The qualitative comparisons in Figure 7 show that baselines struggle with complex hand gestures, producing blurred or coarse hands. In contrast, our decoupled avatar representation captures intricate hand motions, yielding more realistic results.

User Study. In addition to signal-based metrics, we conducted a user study with native Deaf people to evaluate perceptual quality. We distributed 20 survey forms via Deaf community channels and collected 20 valid responses. Participants evaluated 25 sets of videos generated by three representative baseline methods [19, 30, 57] and ours. For each set, participants were asked to select the single video that best satisfied the corresponding criterion:

- **Q1:** Which video best allows you to understand the intended sign-language content? (*Comprehensibility*)
- **Q2:** Which video best presents clear hand gestures and facial expressions? (*Hand/Facial Clarity*)
- **Q3:** Which video shows the most temporally stable visual appearance? (*Temporal Consistency*)

Table 8: User study results comparing SMPL-X mesh rendering and our photorealistic sign avatar.

Representation	SMPL-X Mesh	Photorealistic Sign Avatar
Comprehensibility	17.5%	**82.5%**
Hand/Facial Clarity	33.5%	**66.5%**
Visual Realism	5.0%	**95.0%**
Aesthetic Preference	16.0%	**84.0%**

Table 9: User study results comparing our method with other human avatar modeling methods.

Method	AnimatableGS [30]	EVA [19]	Mmlphuman [57]	Ours
Comprehensibility	8.8%	13.6%	20.8%	**56.8%**
Hand/Facial Clarity	7.8%	11.2%	18.6%	**62.4%**
Temporal Consistency	8.0%	13.8%	25.2%	**53.0%**
Aesthetic Preference	11.2%	15.6%	22.0%	**51.2%**

Table 10: Comparison to the-state-the-art human avatar modeling methods. "↓" indicates that lower values are better, while "↑" means the opposite.

Method	Full			Hand			Face		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Multi-view setting: MVSign dataset									
SplattingAvatar [47]	23.71	0.9625	0.0494	15.64	0.6762	0.3884	16.90	0.7633	0.2423
GaussianAvatar [20]	24.23	0.9633	0.0428	16.30	0.6833	0.3082	17.78	0.7737	0.2086
AnimatableGS [30]	25.09	0.9647	0.0465	16.95	0.7002	0.3135	18.63	0.7842	0.2230
EVA [19]	25.56	0.9667	0.0436	17.43	0.7154	0.2869	19.21	0.8000	0.1964
Mmlphuman [57]	25.91	0.9652	0.0460	17.41	0.7084	0.2675	19.38	0.8025	0.1869
Ours	**27.03**	**0.9689**	**0.0393**	**18.55**	**0.7325**	**0.2568**	**20.60**	**0.8189**	**0.1757**
Monocular setting: Real-world Web video									
SplattingAvatar [47]	18.71	0.9198	0.1223	16.37	0.6572	0.3114	16.97	0.7243	0.2751
GaussianAvatar [20]	18.90	0.9261	0.0918	16.30	0.6668	0.2797	17.29	0.7178	0.2157
AnimatableGS [30]	19.34	0.9252	0.0954	17.42	0.6951	0.2799	17.65	0.7277	0.2561
EVA [19]	19.22	0.9250	0.0966	17.49	0.6958	0.2588	17.75	0.7381	0.2132
Mmlphuman [57]	19.61	0.9239	0.1023	16.90	0.6645	0.2607	17.88	0.7284	0.2374
Ours	**20.39**	**0.9334**	**0.0773**	**18.70**	**0.7189**	**0.2289**	**18.84**	**0.7502**	**0.1864**

– **Q4:** Which video is most visually appealing and acceptable as a sign-language avatar? (*Aesthetic Preference*)

As shown in Table 9, our method achieves the highest ratings across all four aspects, indicating superior perceptual quality from the participants' perspective.

We further conduct a user study within the Deaf community to evaluate the perceptual preference between our photorealistic sign avatar and the SMPL-X mesh. We distributed 20 survey forms via Deaf community channels and collected 20 valid responses. Participants evaluated 10 video pairs, which compare SMPL-X mesh rendering against our photorealistic avatar under the same motion, and rated them from four aspects: comprehensibility, hand/facial clarity, visual realism, and aesthetic preference. As reported in Table 8, participants consistently preferred our photorealistic avatar across these aspects (with 84% favoring ours in aesthetic preference), showing a clear advantage in perceptual quality and user acceptance.

6 Limitations

Since we employ three specialized StyleUNet modules for Gaussian attribute prediction, inference speed is largely dominated by their forward passes. This computational overhead limits real-time applicability. A promising direction is replacing StyleUNet with lightweight architectures to improve efficiency while maintaining acceptable visual fidelity.

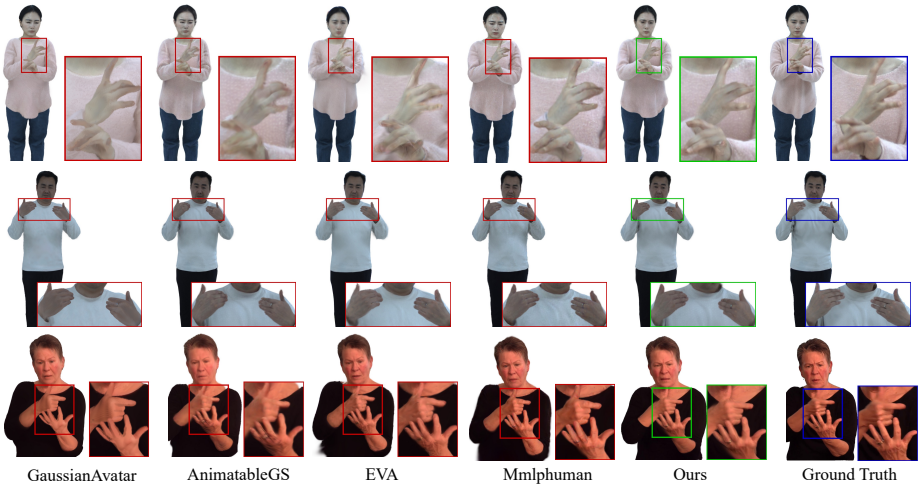

Fig. 7: Qualitative comparison of novel pose synthesis with GaussianAvatar [20], AnimatableGS [30], EVA [19] and Mmlphuman [57]. The first two rows show results on the MVSign dataset, while the last row present results on real-world web videos.

7 Conclusion

In this work, we focus on photorealistic and drivable sign avatar modeling. To this end, we introduce MVSign, the first multi-view Chinese sign language dataset co-designed with Deaf experts, featuring diverse gestures and rich annotations. Our dataset includes comprehensive annotations along with the raw data, including image matting, body part segmentation, 3D keypoints, and SMPL-X parameters. For precise SMPL-X annotation, we propose a hybrid SMPL-X fitting pipeline that integrates outputs from multiple models, producing accurate body, hand, and facial parameters and can also be applied to the monocular setting. Building on MVSign, we introduce a decoupled sign avatar representation that addresses the topological complexity in hand articulation, while a motion-aware data sampling strategy is applied to filter out motion-blurred frames while balancing the distribution of sign articulations. Extensive experiments on MVSign dataset and in-the-wild sign videos demonstrate that our method achieves strong performance both quantitatively and qualitatively.

8 Acknowledgments

This work was supported by the Youth Innovation Promotion Association CAS. It was also supported by the GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC, and the Supercomputing Center of USTC.

References

1. Easymocap - make human motion capture easier. https://github.com/zju3dv/EasyMocap (2021) 6
2. Baltatzis, V., Potamias, R.A., Ververas, E., Sun, G., Deng, J., Zafeiriou, S.: Neural sign actors: A diffusion model for 3d sign language production from text. In: CVPR (2024) 2, 11
3. Cai, Z., Yin, W., Zeng, A., Wei, C., Sun, Q., Yanjun, W., Pang, H.E., Mei, H., Zhang, M., Zhang, L., Loy, C.C., Yang, L., Liu, Z.: SMPLer-X: Scaling up expressive human pose and shape estimation. In: NeurIPS (2023) 5
4. Camgöz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Rwth-phoenix-weather 2014 t: Parallel corpus of sign language video, gloss and translation. In: CVPR (2018) 2, 3, 4, 5, 10
5. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR (2017) 5
6. Cheng, W., Chen, R., Fan, S., Yin, W., Chen, K., Cai, Z., Wang, J., Gao, Y., Yu, Z., Lin, Z., et al.: Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In: ICCV (2023) 2, 3, 5, 10
7. Contributors, M.: Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose (2020) 5
8. Danecek, R., Black, M.J., Bolkart, T.: EMOCA: Emotion driven monocular face capture and animation. In: CVPR (2022) 7
9. Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., Torres, J., Giro-i Nieto, X.: How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In: CVPR (2021) 2, 3, 4, 5, 10
10. Ebling, S., Camgöz, N.C., Braem, P.B., Tissi, K., Sidler-Miserez, S., Stoll, S., Hadfield, S., Haug, T., Bowden, R., Tornay, S., et al.: Smile swiss german sign language dataset. In: LREC (2018) 3, 4
11. Fang, S., Sui, C., Zhou, Y., Zhang, X., Zhong, H., Tian, Y., Chen, C.: Signdiff: Diffusion model for american sign language production. In: 2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–11. IEEE (2025) 10
12. Feng, Y., Choutas, V., Bolkart, T., Tzionas, D., Black, M.J.: Collaborative regression of expressive bodies using moderation. In: 3DV (2021) 11
13. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3D face model from in-the-wild images. ACM TOG (2021) 7
14. Filntisis, P.P., Retsinas, G., Paraperas-Papantoniou, F., Katsamanis, A., Roussos, A., Maragos, P.: Visual speech-aware perceptual 3d facial expression reconstruction from videos. arXiv preprint arXiv:2207.11094 (2022) 7
15. Forte, M.P., Kulits, P., Huang, C.H.P., Choutas, V., Tzionas, D., Kuchenbecker, K.J., Black, M.J.: Reconstructing signing avatars from video using linguistic priors. In: CVPR (2023) 4, 11
16. Guo, Z., Zhou, W., Wang, M., Li, L., Li, H.: Handnerf: Neural radiance fields for animatable interacting hands. In: CVPR (2023) 3
17. Hanke, T.: Hamnosys–representing sign language data in language resources and language processing contexts. In: sign-lang@ LREC 2004 (2004) 4
18. Hong, Y., Peng, B., Xiao, H., Liu, L., Zhang, J.: Headnerf: A real-time nerf-based parametric head model. In: CVPR (2022) 3
19. Hu, H., Fan, Z., Wu, T., Xi, Y., Lee, S., Pavlakos, G., Wang, Z.: Expressive gaussian human avatars from monocular rgb video. In: NeurIPS (2024) 3, 10, 13, 14, 15

20. Hu, L., Zhang, H., Zhang, Y., Zhou, B., Liu, B., Zhang, S., Nie, L.: Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In: CVPR (2024) 2, 3, 8, 10, 14, 15
21. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE TPAMI (2014) 4
22. Işık, M., Rünz, M., Georgopoulos, M., Khakhulin, T., Starck, J., Agapito, L., Nießner, M.: Humanrf: High-fidelity neural radiance fields for humans in motion. ACM TOG (2023) 2, 3, 5, 10
23. Jiang, T., Chen, X., Song, J., Hilliges, O.: Instantavatar: Learning avatars from monocular video in 60 seconds. In: CVPR (2023) 3
24. Jiang, W., Yi, K.M., Samei, G., Tuzel, O., Ranjan, A.: Neuman: Neural human radiance field from a single video. In: ECCV (2022) 3
25. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019) 3
26. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. SIGGRAPH (2023) 3, 9
27. Khirodkar, R., Bagautdinov, T., Martinez, J., Zhaoen, S., James, A., Selednik, P., Anderson, S., Saito, S.: Sapiens: Foundation for human vision models. In: ECCV (2025) 5
28. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: ICCV (2023) 5
29. Lei, J., Wang, Y., Pavlakos, G., Liu, L., Daniilidis, K.: Gart: Gaussian articulated template models. In: CVPR (2024) 2, 3
30. Li, Z., Zheng, Z., Wang, L., Liu, Y.: Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In: CVPR (2024) 2, 3, 8, 10, 12, 13, 14, 15
31. Lin, J., Zeng, A., Wang, H., Zhang, L., Li, Y.: One-stage 3d whole-body mesh recovery with component aware transformer. In: CVPR (2023) 5
32. Liu, L., Habermann, M., Rudnev, V., Sarkar, K., Gu, J., Theobalt, C.: Neural actor: Neural free-view synthesis of human actors with pose control. ACM TOG (2021) 3
33. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. SIGGRAPH Aisa (2015) 10
34. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM (2021) 3
35. Moon, G., Choi, H., Lee, K.M.: Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In: CVPRW (2022) 5
36. Moon, G., Shiratori, T., Saito, S.: Expressive whole-body 3d gaussian avatar. In: ECCV (2024) 3
37. Ni, L.: Chinese Sign Language Tutorial. Fudan University Press (2020) 4
38. Pang, H., Zhu, H., Kortylewski, A., Theobalt, C., Habermann, M.: Ash: Animatable gaussian splats for efficient and photoreal human rendering. In: CVPR (2024) 3
39. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: CVPR (2019) 2, 10, 11
40. Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., Malik, J.: Reconstructing hands in 3D with transformers. In: CVPR (2024) 6

41. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: CVPR (2021) 2, 3, 4, 5, 10
42. Peter, E., Hilsmann, A.: Visual sign language datset. https://cvg.hhi.fraunhofer.de/VisualSignLanguageData.htm (2022) 4
43. Qi, F., Duan, Y., Zhang, H., Xu, C.: Signgen: End-to-end sign language video generation with latent diffusion. In: European Conference on Computer Vision. pp. 252–270. Springer (2024) 10
44. Quandt, L.C., Willis, A., Schwenk, M., Weeks, K., Ferster, R.: Attitudes toward signing avatars vary depending on hearing status, age of signed language acquisition, and avatar type. Frontiers in psychology (2022) 2
45. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. SIGGRAPH Aisa (2017) 6
46. Rong, Y., Shiratori, T., Joo, H.: Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In: ICCV (2021) 11
47. Shao, Z., Wang, Z., Li, Z., Wang, D., Lin, X., Zhang, Y., Fan, M., Wang, Z.: Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. In: CVPR. pp. 1606–1616 (2024) 2, 10, 14
48. Wang, J., Karaev, N., Rupprecht, C., Novotny, D.: Vggsfm: Visual geometry grounded deep structure from motion. In: CVPR (2024) 5
49. Wang, L., Zhao, X., Sun, J., Zhang, Y., Zhang, H., Yu, T., Liu, Y.: Styleavatar: Real-time photo-realistic portrait avatar from a single video. In: SIGGRAPH (2023) 3, 8
50. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE TIP (2004) 9, 10
51. Weng, C.Y., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In: CVPR (2022) 3
52. Wu, M., Wang, Y., Hu, Q., Yu, J.: Multi-view neural human rendering. In: CVPR (2020) 4
53. Xiong, Z., Li, C., Liu, K., Liao, H., Hu, J., Zhu, J., Ning, S., Qiu, L., Wang, C., Wang, S., et al.: Mvhumannet: A large-scale dataset of multi-view daily dressing human captures. In: CVPR (2024) 2, 3, 5, 10
54. Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. In: ICCV (2023) 5, 6
55. Yu, Z., Huang, S., Cheng, Y., Birdal, T.: Signavatars: A large-scale 3d sign language holistic motion dataset and benchmark. In: ECCV (2024) 2, 4, 5, 10
56. Zeng, A., Yang, L., Ju, X., Li, J., Wang, J., Xu, Q.: Smoothnet: A plug-and-play network for refining human poses in videos. In: ECCV (2022) 7
57. Zhan, Y., Shao, T., Yang, Y., Zhou, K.: Real-time high-fidelity gaussian human avatars with position-based interpolation of spatially distributed mlps. In: CVPR (2025) 8, 10, 13, 14, 15
58. Zhang, H., Tian, Y., Zhang, Y., Li, M., An, L., Sun, Z., Liu, Y.: Pymaf-x: Towards well-aligned full-body model regression from monocular images. PAMI (2023) 11
59. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) 9, 10
60. Zhao, F., Yang, W., Zhang, J., Lin, P., Zhang, Y., Yu, J., Xu, L.: Humannerf: Efficiently generated human radiance field from sparse inputs. In: CVPR (2022) 3
61. Zheng, Z., Huang, H., Yu, T., Zhang, H., Guo, Y., Liu, Y.: Structured local radiance fields for human avatar modeling. In: CVPR (2022) 4

62. Zhou, H., Zhou, W., Qi, W., Pu, J., Li, H.: Improving sign language translation with monolingual data by sign back-translation. In: CVPR (2021) 3, 4, 5
63. Zielonka, W., Bagautdinov, T., Saito, S., Zollhöfer, M., Thies, J., Romero, J.: Drivable 3d gaussian avatars. In: 3DV (2025) 3
64. Zielonka, W., Bolkart, T., Thies, J.: Towards metrical reconstruction of human faces. In: ECCV (2022) 7
65. Zuo, R., Wei, F., Chen, Z., Mak, B., Yang, J., Tong, X.: A simple baseline for spoken language to sign language translation with 3d avatars. In: ECCV (2024) 2