Stable Signer V3: 4D Deformable Primitive Splatting Is All You Need for Sign Language Production

Sen Fang Liwei Che Dimitris N. Metaxas
Rutgers University
https://signerx.github.io/Stable-Signer-V3

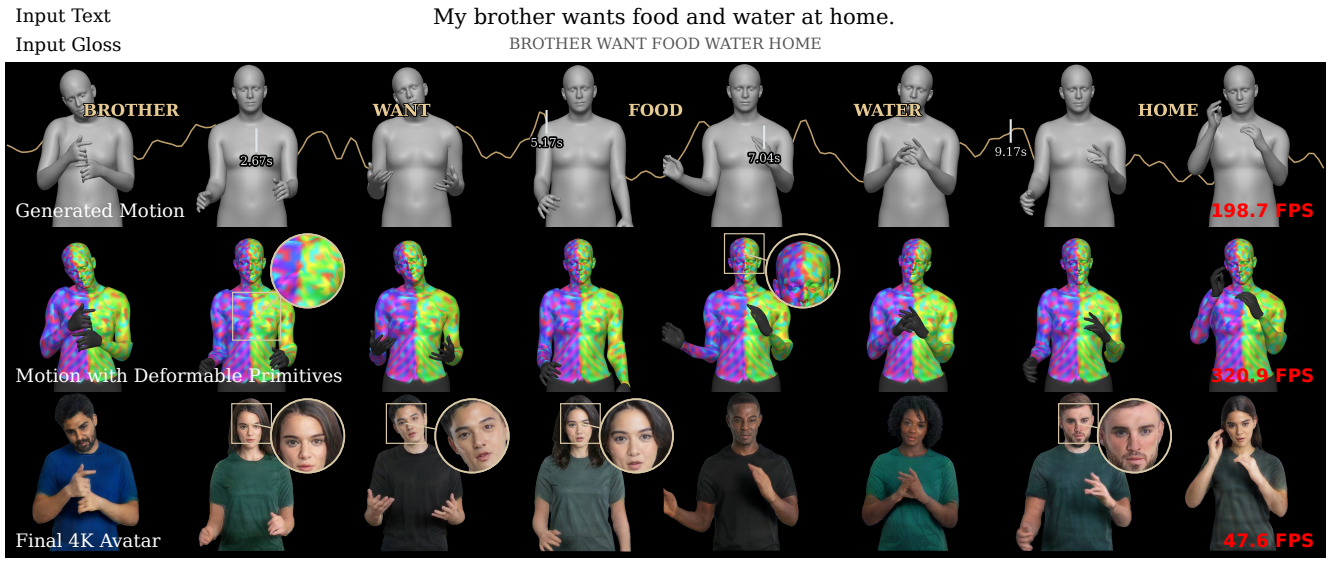


Figure 1. **Stable Signer V3.** Given an input sentence or gloss sequence, our system generates signing motion (top), augments the body with deformable face and body primitives (middle), and renders personalized 4K avatars (bottom). Columns show successive moments of the same motion, with four male and four female avatars alternating in the bottom row.

Abstract

*We present **Stable Signer V3**, an interactive high-resolution sign language avatar system built on **4D Deformable Primitive Splatting**. It addresses three limitations of existing avatars: restricted facial expression, inaccurate identity-specific body deformation, and loss of fine appearance details. **Shared Face Primitive Splatting** adds transferable superquadric primitives for richer expression control; **Test-Time Body Primitive Splatting** learns an identity-specific virtual muscle layer for more faithful motion; and **Test-Time Tunneling Super-Resolution** uses TunnelingNet to recover identity-specific high-resolution details. Both test-time modules are optimized once per signer identity; inference then runs entirely from the stored identity parameters. Together, they reduce facial geometry error by 71.4%, body-boundary error by 35.1% across ten identities, and high-resolution RGB reconstruction error by 24.2% on held-out evaluations, while retaining interactive streaming.*

1. Introduction

Sign language production aims to translate text or lexical sequences into intelligible, natural, and expressive signing videos [7, 49, 57]. Sign language conveys information through coordinated manual gestures, facial expressions, head poses, and body movements [2, 43, 52, 78]. A practical production system must therefore go beyond predicting skeletons or parametric poses: it must consistently express these motions through avatars with diverse identities and appearances, preserve subtle expressions and visual details, and meet the latency requirements of interactive ap-

plications [16, 26, 62].

Existing approaches broadly follow two routes. *Text2Pose2Video* synthesizes signing videos from intermediate poses with pose-conditioned image or video generation models [8, 48, 51, 54, 57, 58, 63]. *Text2Avatar* predicts language-conditioned 3D signing motion and uses it to animate explicit characters [1, 13, 20, 34, 77, 86]. Pose-guided human-video models provide additional tools for this pipeline [5, 9, 35, 84]. Dictionary-based production also exploits reusable signing segments and their transitions [61, 63]. Pose generation models provide structured representations of signing motion [7, 49, 52, 70, 73]. Within these pipelines, 3D Gaussian splatting [25] enables efficient rendering, with recent extensions supporting signing avatars [13, 20] and real-time upper-body animation [77]. Fine-grained signing remains a central rendering challenge even with accurate 3D pose prediction and fast rendering. The resulting motion is jointly constrained by the local expressiveness of the parametric human model, the compatibility between template deformation and individual geometry, and the image decoder's ability to reconstruct fine details. We investigate how compact adaptation mechanisms can improve signing expression by building on existing motion and avatar priors.

Three challenges motivate this work. First, the compact expression spaces of parametric face models such as FLAME [30] suppress subtle local motion and impose a geometric ceiling on facial detail. Second, shared skinning and deformation rules cannot consistently accommodate identity-specific body shape and local surface motion [15, 28, 38, 85]. An avatar that appears plausible in its reference pose can therefore develop local shape errors as the pose changes. Third, general-purpose decoders can discard identity-specific high-frequency details when reconstructing images from compact Gaussian features [24, 71, 77]. These challenges concern facial control, identity-specific body deformation, and high-resolution appearance, respectively. Together, they limit the quality of existing sign language production systems.

To address these challenges, we introduce **Stable Signer V3**, a sign language production system based on **4D Deformable Primitive Splatting**. Our key idea is to introduce a learnable local deformation layer between the parametric human model and the avatar, separating shared facial motion capabilities from identity-specific body and appearance adaptation. We instantiate the deformable primitives as superquadrics, allowing local residual deformations to vary with pose and expression while preserving the existing human control interface and avatar bindings. Given text or gloss input, the system uses SignSparK [34] for motion completion and GUAVA [77] for avatar initialization and feature rendering. A shared facial layer learns local residual deformations driven by FLAME expression, jaw, and eye controls and transfers them to the avatar through surface bindings. For the body, a pose-conditioned virtual muscle layer is fitted to each identity from reference videos; it corrects geometry before skinning so that bound Gaussians follow the adjusted surface, with displacement and smoothness constraints discouraging implausible deformations. For appearance, an independently normalized Softmax tunneling path uses terminal decoder semantics to learn identity-specific high-resolution detail while preserving the early Gaussian feature pathway. Test-time adaptation forms a one-time per-identity character-preparation stage. The resulting body and detail parameters serve every subsequent sequence, cleanly separating identity adaptation from online generation.

Our contributions are summarized as follows:

- **Shared Face Primitive Splatting.** A shared deformable primitive layer enables more complex and controllable facial expressions.
- **Test-Time Body Primitive Splatting.** An identity-specific virtual muscle layer adapts body deformation from reference videos while keeping the avatar generation backbone frozen.
- **Test-Time Tunneling Super-Resolution.** A late tunneling path learns identity-specific high-resolution detail from reference videos and stores it in the deployed decoder.

2. Related Work

Sign Language Production. Resources such as WLASL [29], How2Sign [6], JWSign [10], and YouTube-SL-25 [59] support research across lexical signs, continuous signing, and multiple languages. Pose-conditioned video synthesis has been explored through Text2Sign [57], SignSynth [58], and Everybody Sign Now [48], with subsequent work on appearance control, large-scale co-articulation, and diffusion-based synthesis [8, 51, 54]. Broader human-video synthesis develops pose-guided generation through DreaMoving [9], Follow Your Pose [35], and MimicMotion [84], with ControlVideo [83] and Wan-Animate [5] further exploring controllable video generation and character animation. The 3D route combines signing-motion generation [1, 34, 86] with controllable avatar rendering [13, 20, 77]. For motion generation, some methods retrieve and concatenate signing clips from existing dictionaries [54, 61, 63, 76], while others learn to predict continuous motion from language inputs using progressive prediction or motion primitives [49, 50, 52]. Motion modeling has also explored adversarial training, skeletal graph structure, and non-autoregressive generation [17, 47, 53]. Related text-to-human-motion research studies diffusion-based synthesis and language-like motion modeling [21, 60, 80], providing a broader context for language-conditioned animation. Recent motion generators explore discrete diffusion for gloss-

to-pose synthesis [70], autoregressive motion tokens [73], and multilingual generation from text and prompts [7]. SignSparK [34] combines sparse keyframe control with conditional flow matching to generate dense 3D signing motion. Appearance transfer in skeletal pose sequences further studies how signer-specific characteristics can be changed while retaining signing motion [37]. Together, these directions address motion accuracy, co-articulation, appearance control, and efficient generation [16, 62]. We build on this motion completion capability and focus on facial, body, and appearance adaptation for personalized avatar rendering.

3D Humans and Gaussian Avatars for Sign Language. SMPL-X [40] provides joint body, hand, and facial controls for animatable humans, while FLAME [30] models facial shape, expression, and articulation. These parametric representations support pose-driven signing avatars [1, 86]. Whole-body reconstruction methods, including SMPLer-X and component-aware mesh recovery [3, 32], and dedicated hand reconstruction [41] support detailed avatar controls; DWPose provides complementary whole-body keypoint estimates [72]. Complementary work improves 3D pose estimation for sign language [18] and resolves hand-to-hand intersections [19], supporting more reliable motion inputs. Accurate facial and body motion is particularly relevant to signing, where non-manual cues contribute to meaning [26, 78]. Building on 3D Gaussian splatting [25], SplattingAvatar [55] embeds Gaussians on a deformable mesh surface, while GART [28] learns forward skinning and latent bones to capture articulated and non-rigid motion. Articulated Gaussian models further explore cage-based deformation in D3GA [85], monocular human reconstruction in GauHuman [15], and efficient human rendering in ASH and HiFi4G [24, 38]. Related efforts address monocular Gaussian avatar reconstruction in GVA [33], image-based body estimation in GST [44], and generalizable sparse-view synthesis in GHG [27]. Expressive avatar reconstruction is also studied by X-Avatar [56] and EVA [13], with EVA learning Gaussian human avatars from monocular RGB video. SignSplat [20] adapts Gaussian avatars to signing through sequence-level mesh fitting and Gaussian regularization, emphasizing accurate hands and facial expressions. We retain structured human controls and surface bindings while adding a shared facial residual layer and identity-specific body deformation.

Personalized Avatars and Test-Time Training. Video-based character reconstruction has developed through HumanNeRF [69] and NeuMan [23], with Vid2Avatar exploring self-supervised scene decomposition for videos in the wild [11]. Efficient character preparation is explored by InstantAvatar [22], which learns avatars from monocular video. Subject-specific optimization can recover detailed avatar appearance: Gaussian Head Avatar [71] jointly learns neutral Gaussians and an expression-driven deformation field for high-fidelity 2K head rendering. Reference-conditioned image and video synthesis transfers appearance from exemplar images [65, 75]. High-resolution conditional synthesis and latent diffusion provide complementary image-generation frameworks [46, 64], with ControlNet [79] and ControlNeXt [42] exploring additional conditioning for image and video generation. For 3D avatars, GUAVA [77] predicts an animatable upper-body Gaussian avatar from a single image and refines rendered features with a neural decoder. Lightweight adaptation offers a way to reuse pretrained representations: LoRA [12] learns low-rank weight updates, and LLaMA-Adapter [82] introduces new conditioning through zero-initialized attention. Adapters also connect pretrained visual representations to dense prediction tasks [4]. These adaptation mechanisms motivate compact trainable paths around a frozen backbone. We combine generalizable avatar initialization with identity-specific adaptation of body geometry and high-resolution appearance. Character reference videos supervise a body deformation layer and a detail module at the decoder's terminal stage, with the pretrained backbone held fixed. Both modules are optimized once per signer, and online signing directly reuses the stored parameters.

3. Method

3.1. Overview

Given text or gloss input, SignSparK [34] completes dictionary-retrieved sparse poses into body, hand, and facial controls $\{\mathbf{q}_t\}_{t=1}^{T}$. GUAVA [77] initializes a surface-bound Gaussian avatar from a reference image. We augment this pipeline with shared facial deformation and identity-specific body and appearance adaptation.

For signer s, aligned reference images, controls, and cameras $\mathcal{V}_s = \{I_t, \mathbf{q}_t, \pi_t\}$ supervise image formation:

$$\widehat{I}_t = \mathcal{D}_{\psi_s}(\mathcal{R}_{\pi_t}[\mathcal{G}_s(\mathbf{q}_t; \phi, \theta_s)]) \,, \tag{1}$$

where ϕ is the shared facial controller, θ_s is the body adaptation, and $\mathcal{G}_s$ transports Gaussians with the deformed surface. The rasterizer $\mathcal{R}$ projects Gaussian features, and $\mathcal{D}_{\psi_s}$ decodes them into high-resolution RGB. Adaptation occurs once per identity; inference runs from stored parameters.

3.2. Deformable Primitive Representation

Here *4D* denotes pose- and time-conditioned primitive transformations across motion sequences. A sparse field blends control-dependent affine residuals over superquadric supports. Primitive k has center $\mathbf{c}_k$, orthonormal frame $\mathbf{R}_k$, positive axis lengths $\mathbf{a}_k$, and normalized distance

$$d_k(\mathbf{x}) = \left\| \mathrm{diag}(\mathbf{a}_k)^{-1} \mathbf{R}_k^{\mathsf{T}} (\mathbf{x} - \mathbf{c}_k) \right\|_4 . \tag{2}$$

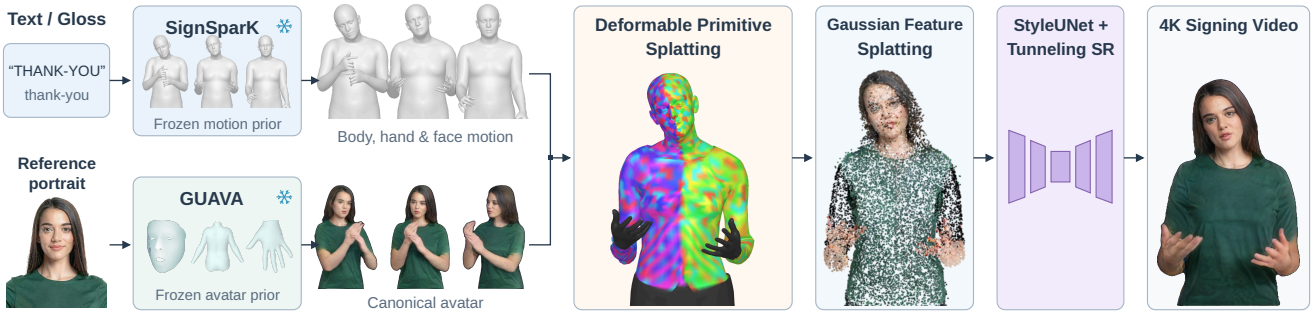


Figure 2. **Overview of Stable Signer V3.** Motion controls drive an avatar through deformable primitives. Gaussian feature splatting and a tunneling-enhanced decoder produce high-resolution signing videos.

The level set $d_k(\mathbf{x}) = 1$ defines a superellipsoid with fixed shape exponents. Centers are initialized in bilateral pairs, and local surface covariance determines orientations and scales. The supports remain fixed during controller learning. Each surface sample receives a sparse neighborhood $\mathcal{N}(\mathbf{x})$ with weights

$$w_k(\mathbf{x}) = \frac{\exp[-2d_k(\mathbf{x})^2]}{\sum_{j\in\mathcal{N}(\mathbf{x})} \exp[-2d_j(\mathbf{x})^2]}. \tag{3}$$

Fixed reference-surface bindings blend controller-predicted affine residuals $(\mathbf{A}_{k,t}, \mathbf{b}_{k,t})$:

$$\Delta\mathbf{x}_t = \sum_{k\in\mathcal{N}(\mathbf{x})} w_k(\mathbf{x}) \left[\mathbf{A}_{k,t}(\mathbf{x}_t - \mathbf{c}_k) + \mathbf{b}_{k,t}\right], \tag{4}$$

Here $\mathbf{x}_t$ uses local coordinates. Zero-initialized controller outputs preserve geometry. The field deforms the surface, while bound Gaussians supply appearance.

3.3. Shared Face Primitive Splatting

The shared facial field is conditioned on

$$\mathbf{u}_t^{\text{face}} = [\mathbf{e}_t, \mathbf{j}_t, \mathbf{o}_t, \boldsymbol{\ell}_t], \tag{5}$$

comprising 50 expression coefficients, three jaw parameters, six eye-pose parameters, and two eyelid controls. Global head orientation is handled by the body rig. A multilayer perceptron maps normalized facial controls to the affine residuals in Eq. 4. Common facial surface correspondences provide the interface for transferring the shared controller across identities.

Paired geometry from a tracked FLAME sequence uses ten expression coefficients for the base and fifty for the target, with other controls identical. For corresponding surface samples $\mathbf{x}_{ti}$ and $\mathbf{y}_{ti}$, we minimize

$$\begin{aligned}\mathcal{L}_{\text{face}} = \frac{1}{|\Omega|} \sum_{(t,i)\in\Omega} \omega_i \rho(\mathbf{x}_{ti} + \Delta\mathbf{x}_{ti} - \mathbf{y}_{ti}) \\ + \lambda_f \mathbb{E}_t \|\mathbf{T}_t\|_F^2,\end{aligned} \tag{6}$$

where ρ is a smooth L_1 penalty, Ω indexes training samples, and $\mathbf{T}_t$ collects affine coefficients. Weights ω_i emphasize lips and eyes. Appearance, opacity, and the decoder remain frozen; compatible facial bindings transfer the controller across identities.

3.4. Test-Time Body Primitive Splatting

An identity-specific virtual muscle layer covers the torso, shoulders, arms, and wrists. Static and pose-dependent affine corrections capture individual surface motion:

$$\mathbf{T}_{s,t} = \tanh\left(\mathbf{T}_s^{\text{static}} + f_{\theta_s}(\overline{\mathbf{p}}_t)\right), \tag{7}$$

where $\overline{\mathbf{p}}_t$ is the normalized 63-dimensional body pose. Static and dynamic components capture shape and articulation effects, with bounded coefficients limiting distortion.

The blended field modifies the canonical carrier before skinning:

$$\widetilde{\mathbf{V}}_{s,t} = \text{LBS}(\mathbf{V}_s^{\text{can}} + \mathcal{S}(\Delta\mathbf{V}_{s,t}), \mathbf{q}_t). \tag{8}$$

Here $\mathcal{S}$ smooths offsets along body mesh edges, preserving the head, hands, and motion controls. Template and UV Gaussians follow corrected vertices and parent triangles, respectively.

Differentiable rendering supervises a dilated visibility region of body-bound Gaussians. Raw and decoded RGB, gradients, SSIM [68], silhouette, and boundary losses enforce video alignment:

$$\begin{aligned}\mathcal{L}_{\text{body}} = \lambda_r\mathcal{L}_{\text{raw}} + \lambda_c\mathcal{L}_{\text{rgb}} + \lambda_g\mathcal{L}_{\text{grad}} + \lambda_s\mathcal{L}_{\text{ssim}} \\ + \lambda_m\mathcal{L}_{\text{mask}} + \lambda_b\mathcal{L}_{\text{boundary}} + \mathcal{L}_{\text{reg}}.\end{aligned} \tag{9}$$

Regularization penalizes affine magnitude, neighboring translations, and mean vertex displacement. The avatar generator, appearance, and renderer remain frozen. Held-out reconstruction guides selection.

3.5. Test-Time Tunneling Super-Resolution

A late Softmax tunnel adapts terminal decoder features into identity-specific detail before RGB prediction.

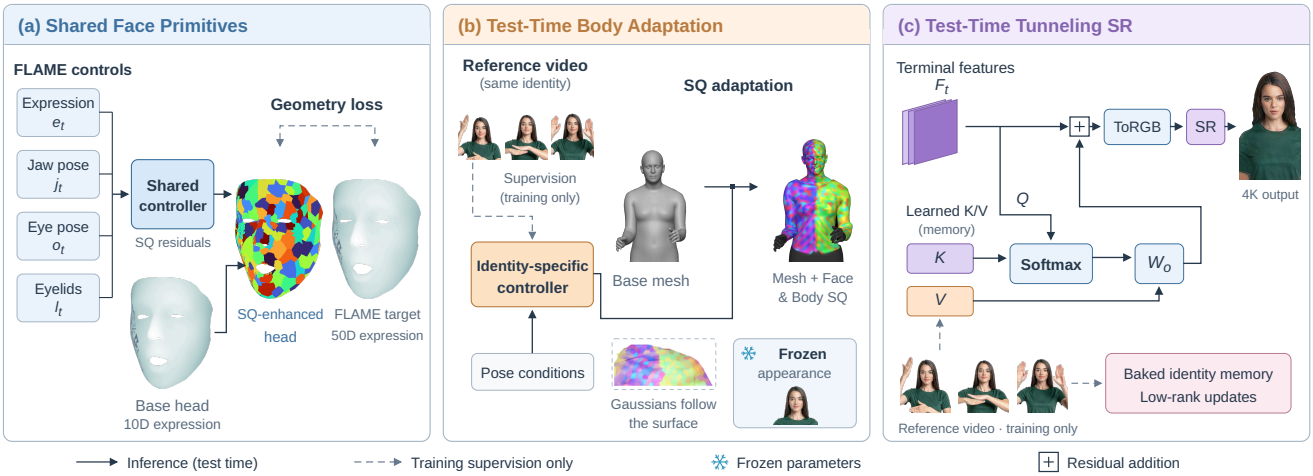


Figure 3. **Deformation and detail adaptation.** (a) Shared facial primitives learn geometric residuals with FLAME supervision. (b) Identity-specific body primitives adapt the avatar carrier. (c) A late Softmax tunneling pathway learns high-resolution appearance. Dashed arrows denote training supervision; solid arrows denote inference.

Let $\mathbf{h}_t(\mathbf{p})$ be the terminal feature at location $\mathbf{p}$. Learnable spatial keys and values $\{\mathbf{k}_{s,n}, \mathbf{v}_{s,n}\}_{n=1}^{N}$ are initialized from reference frames and optimized by video reconstruction. Interpolation to the feature resolution gives

$$\alpha_{t,n}(\mathbf{p}) = \frac{\exp[\mathbf{z}_t(\mathbf{p})^\top \mathbf{k}_{s,n}(\mathbf{p})/\sqrt{r}]}{\sum_{m=1}^{N} \exp[\mathbf{z}_t(\mathbf{p})^\top \mathbf{k}_{s,m}(\mathbf{p})/\sqrt{r}]}, \tag{10}$$

$$\widetilde{\mathbf{h}}_t(\mathbf{p}) = \mathbf{h}_t(\mathbf{p}) + W_o \sum_{n=1}^{N} \alpha_{t,n}(\mathbf{p}) \mathbf{v}_{s,n}(\mathbf{p}), \tag{11}$$

Here N is the number of memory slots, $\mathbf{z}_t(\mathbf{p}) = W_q \mathbf{h}_t(\mathbf{p})$, and r is the query/key dimension. Projections W_q and W_o map into and out of the tunnel, while $\alpha_{t,n}$ weights slot n. Softmax normalizes slots per location; zero-initialized W_o preserves the original feature path.

Low-rank updates $\Delta W = BA$, with trainable factors A and B, complement the tunnel with pretrained weights frozen. Training and inference use one StyleUNet [66] pass; stored identity memory supplies the deployed inference path.

High-resolution reconstruction. Let J_t be the decoded RGB image and $\mathcal{U}_q$ denote spatial upsampling by factor q. The intermediate reconstruction is

$$\widehat{I}_t^{\text{mid}} = \mathcal{U}_q(J_t) + \mathcal{H}_s(\widetilde{\mathbf{h}}_t, J_t), \tag{12}$$

where $\mathcal{H}_s$ is an identity-adapted PixelShuffle residual head. A shared fully convolutional head refines aligned native-resolution crops for 4K output. High-resolution target patches preserve supervision detail.

Appearance objective. We learn appearance parameters from aligned video frames with fixed geometry and Gaussian features. Intermediate-resolution supervision combines RGB, image-gradient, multiscale high-pass, detail-band, local-contrast, perceptual, and low-frequency penalties. The native-resolution head receives RGB, gradient, high-pass, and local-detail supervision:

$$\mathcal{L}_{\text{detail}} = \lambda_{\text{mid}} \mathcal{L}_{\text{mid}} + \mathbb{E}_{t,\mathcal{P}} \sum_{a} \lambda_a \mathcal{L}_a(\widehat{I}_t^{\mathcal{P}}, I_t^{\mathcal{P}}), \tag{13}$$

Here $\mathcal{L}_{\text{mid}}$ is the intermediate loss above; λ_{mid} and λ_a are nonnegative weights. The crop $\mathcal{P}$ is shared by prediction and target, and a indexes the native-resolution loss terms. Patch sampling preserves target-frequency information while bounding training memory. Body and detail adaptation are optimized with their respective frozen components, yielding a geometric controller and an appearance decoder for each identity. Online signing combines these modules with the shared facial field and newly generated motion.

4. Experiments

4.1. Experimental Setup

The evaluation covers facial expression reconstruction, identity-specific body adaptation, and high-resolution detail recovery.

Data and protocol. Controlled face and body studies use chronological 80%/20% training–evaluation splits and fixed optimization budgets; checkpoint selection uses the training segment exclusively. Detail adaptation is evaluated on held-out frames, and cross-method rendering uses identical poses, cameras, backgrounds, and target-defined crops for

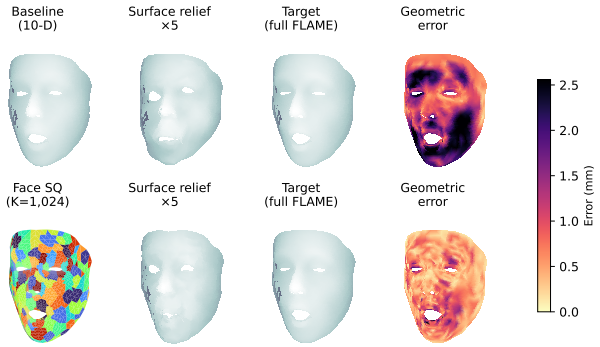


Figure 4. **Face primitive reconstruction on a held-out frame.** Top: restricted FLAME. Bottom: SQ-corrected geometry, colored by primitive assignment. The second column amplifies surface deviations from the target by $5\times$ for visibility. The third column shows the full-expression target; the last shows unamplified geometric error on a shared scale (darker indicates larger error).

every method. Scores are first aggregated per identity and then macro-averaged.

Metrics. Facial geometry is measured by Euclidean point error in the native shared tracked coordinates. Body adaptation reports refined RGB L1, boundary L1, and SSIM [68]. Rendering comparisons use foreground PSNR [67] together with SSIM and LPIPS [81] inside a shared ground-truth person crop. Runtime measurements distinguish independently warmed stage throughput from measured resident-pipeline wall time. Semantic preservation is evaluated through fixed back-translation using BLEU-4 [39] and ROUGE-L F1 [31] on poses re-extracted from the rendered RGB videos.

4.2. Ablation I: Shared Face Primitives

We train the shared face layer on the tracked FLAME sequence provided with GUAVA [77], using a chronological 80%/20% training–evaluation split and a fixed optimization budget. The restricted FLAME carrier and the full-expression target share identical articulation, isolating the expression geometry that our layer must recover.

Table 1 reports Euclidean error over all retained head samples, including template- and UV-bound samples: mean head error falls from 1.267 to 0.362 mm (71.4%), and facial error falls from 1.492 to 0.518 mm. Figure 4 shows the held-out midpoint with a shared error scale and $5\times$ deviation magnification. The shared primitive controller reduces head error by 71.4% and recovers fine expression geometry across the face; region-wise results are provided in the supplement.

4.3. Ablation II: Test-Time Detail Tunneling

Figure 6 compares frozen GUAVA, an Adapter, and our tunneling decoder through synchronized facial crops. The 512-resolution baseline is enlarged for display. Both adapted outputs are generated at 1024 resolution, and the visualization presents the combined effect of adaptation and native-resolution decoding. All trainable variants use the same frames, Gaussian features, targets, optimizer, and update budget. On held-out frames, Adapter reduces mean RGB L1 by 13.5%, while tunneling reduces it by 24.2% relative to the frozen baseline.

Method	Head	Face	Lips
Restricted FLAME	1.267	1.492	1.739
Face primitives	0.362	0.518	0.990
Reduction (%)	71.4	65.3	43.1

Table 1. **Face reconstruction error (mm, lower is better).** Average over the held-out 20% of the sequence. Head aggregates all retained samples; face and lip metrics use their semantic subsets. Held-out full-expression FLAME tracking provides the target geometry.

Body model	RGB L1 ↓	Boundary L1 ↓	SSIM ↑
Baseline	0.0649	0.0350	0.9038
$K = 512$	0.0599	0.0282	0.9076
$K = 1{,}024$	0.0575	0.0270	0.9096

Table 2. **Body primitive count.** Baseline and Body SQ use the same chronological 80%/20% train–evaluation split for one identity. Both SQ variants share the adaptation budget and loss weights. Metrics are computed in the body region; RGB L1 uses refined renderings.

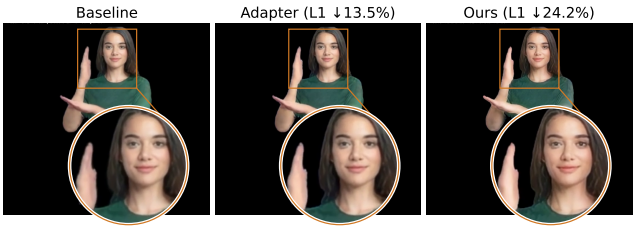


Figure 6. **Detail adaptation.** Frozen baseline, Adapter, and our tunneling decoder at the same timestamp. Circular insets magnify identical facial regions. The baseline is enlarged from 512 resolution; both adapted outputs are generated at 1024. Parentheses report mean held-out RGB L1 reductions relative to the frozen baseline.

4.4. Ablation III: Test-Time Body Primitives

Under a fixed 249-step adaptation budget and chronological 80%/20% training–evaluation split, Table 2 shows that increasing K from 512 to 1,024 reduces refined RGB L1 from 0.0599 to 0.0575 and boundary L1 from 0.0282 to 0.0270.

Figure 5 compares frozen GUAVA with identity-specific Body SQ under the same frame, pose, and camera; the learned layer locally adapts the carrier around the torso and upper arms.

Reference overlays locate the learned patches, carrier

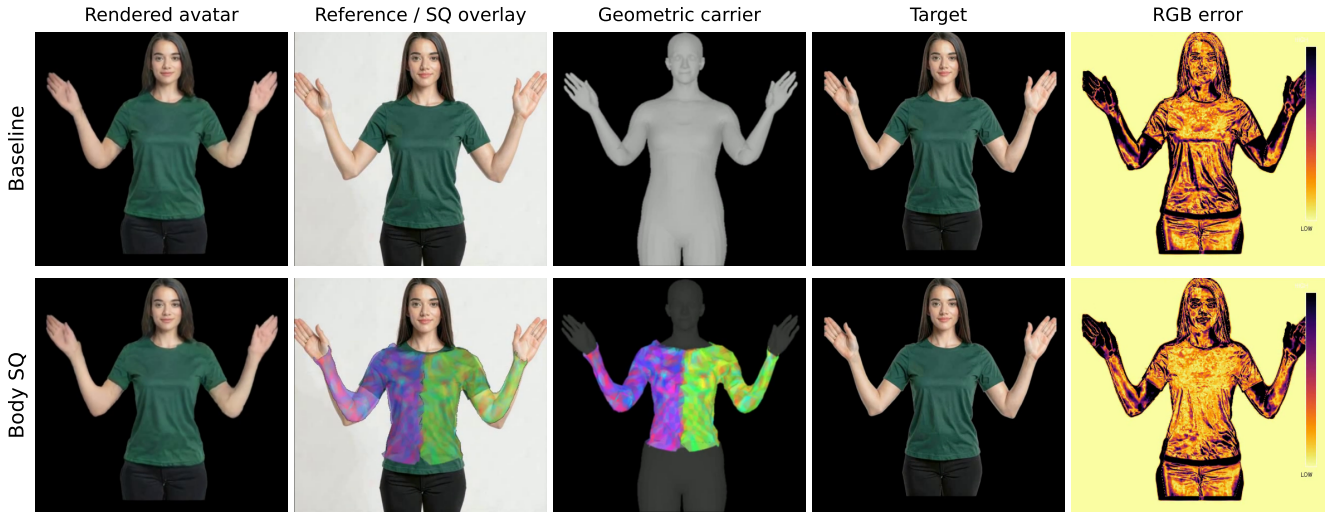


Figure 5. **Test-time body adaptation.** Frozen GUAVA (top) and identity-specific Body SQ (bottom). Columns show the rendered avatar, reference frame with adapted SQ patches overlaid at 60% opacity, geometric carrier, target, and RGB error on a shared scale. The top-row reference has no SQ overlay. We show the held-out moment with the largest body-region RGB error reduction in this clip.

views expose their geometric support, and RGB errors use the same $[0, 0.08]$ range with darker colors indicating larger error, separating surface alignment from its effect on rendering. Figure 5 highlights the resulting geometric fit, and Table 2 confirms simultaneous gains in RGB fidelity, boundary alignment, and structural similarity.

4.5. Comparison with Existing Methods

Online efficiency. Table 3 reports warmed batch-one component throughput on a single Quadro RTX 8000. Motion synthesis and primitive deformation run at 198.7 and 320.9 FPS, while HD avatar rendering and 4K detail refinement reach 52.0 and 47.6 FPS. In the resident pipeline, stages overlap across streaming chunks. A warmed 784-frame, three-chunk text-to-HD-avatar run completes in 31.0 seconds wall time (25.3 generated FPS), including composition, motion synthesis, streamed rendering, encoding, and chunk concatenation. Body and detail adaptation are performed once for each new avatar and baked into the deployed model; subsequent sequences run directly from the baked parameters.

Back-translation evaluation. Table 4 tests whether the proposed visual enhancement components alter the linguistic content of the output. Both configurations render the same motion sequence; the full configuration additionally activates shared face primitives, test-time body primitives, and detail tunneling. We re-extract signing poses from each rendered video and translate them with the same frozen back-translation model. Activating all three components raises BLEU-4 from 25.6450 to 26.8897 (4.9%) and ROUGE-L from 59.4771 to 61.7298 (3.8%). The complete enhancement stack therefore improves geometric and rendering fidelity while fully preserving the underlying signs.

Stage	Output	Throughput ↑
Motion synthesis	SMPLest-X [74] motion	198.7 FPS
Primitive deformation	Face and body primitives	320.9 FPS
Avatar rendering	HD RGB	52.0 FPS
Detail refinement	4K RGB	47.6 FPS
Resident pipeline	Text → HD RGB	25.3 FPS

Table 3. **Online efficiency.** Component rows report warmed batch-one throughput on one Quadro RTX 8000. The resident row reports measured wall-clock throughput over 784 frames, including composition, motion synthesis, HD rendering, encoding, and concatenation.

Configuration	**B-T BLEU-4↑**	**ROUGE-L↑**
Enhancement disabled	25.6450	59.4771
Enhancement enabled	**26.8897**	**61.7298**

Table 4. **Semantic preservation under visual enhancement.** Both configurations render identical motions. The enabled configuration activates shared face primitives, test-time body primitives, and detail tunneling. Signing poses are re-extracted from the final RGB videos and translated by the same frozen model.

Rendering fidelity. Table 5 first compares our system with four reproducible open-source avatar methods at one shared held-out pose for each of four input reference videos. We compute PSNR inside the shared ground-truth foreground mask and SSIM/LPIPS within the same ground-truth person bounding box, then macro-average identities. The second block evaluates our identity-adapted renderer against frozen GUAVA using matched cached features and crops sampled directly from held-out 4K targets.

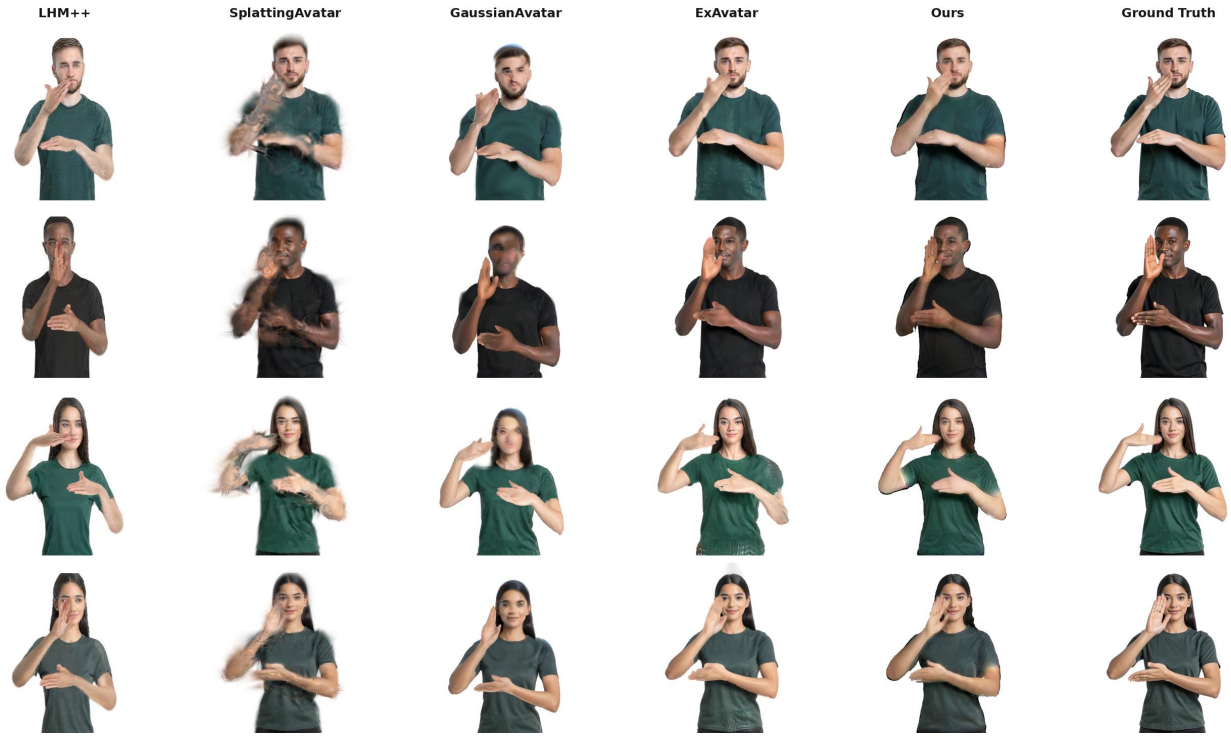


Figure 7. **Qualitative rendering comparison.** Each column corresponds to one method: LHM++, SplattingAvatar, GaussianAvatar, ExAvatar, Ours, and Ground Truth. Each row shows the same held-out identity and signing pose across methods; the comparison focuses on facial detail, finger structure, and garment boundaries.

The second block measures identity adaptation by comparing frozen GUAVA and our complete renderer on the same held-out subset from each of ten reference videos. Both methods receive identical cached Gaussian features, crop coordinates, and 768×768 crops sampled directly from the held-out 4K frames. Macro-averaged over identities, Stable Signer V3 raises PSNR from 19.418 to 21.790 dB and SSIM from 0.833 to 0.860, while reducing LPIPS from 0.245 to 0.226. This corresponds to a 2.372 dB PSNR gain and an 8.0% relative LPIPS reduction over the frozen renderer.

4.6. Qualitative Rendering Comparison

Figure 7 compares methods column by column on the same sign-language sequence. The four published baselines are followed by our method and Ground Truth. The shared framing makes facial detail, finger articulation, and garment boundaries directly comparable across columns. Across identities and poses, our method preserves facial character, articulated fingers, silhouettes, and garment boundaries more consistently than the reproduced baselines.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Four-identity shared held-out-pose evaluation			
LHM++ [45]	8.987	0.646	0.576
SplattingAvatar [55]	18.396	0.764	0.348
GaussianAvatar [14]	10.246	0.684	0.460
ExAvatar [36]	18.141	0.775	0.291
Stable Signer V3	**19.415**	**0.787**	**0.276**
Matched held-out reference-video evaluation (10 identities)			
Frozen GUAVA [77]	19.418	0.833	0.245
Stable Signer V3	**21.790**	**0.860**	**0.226**

Table 5. **Rendering fidelity.** Macro-averaged results under two matched protocols: shared held-out poses for four identities (top) and crops sampled directly from held-out 4K targets for ten identities (bottom). PSNR uses GT foreground; SSIM and LPIPS use the shared GT person crop.

5. Conclusion

We presented Stable Signer V3, a sign language production framework combining shared facial primitives, test-time identity-specific body primitives, and a tunneling decoder for high-resolution appearance. Together, they unify expressive control and personalized rendering through a reusable deformable representation. Reference-video adaptation is performed once, after which the stored identity parameters drive subsequent signing sequences. On held-out evaluations, the three modules reduce head reconstruction error by 71.4%, body-boundary error by 35.1% across ten identities, and high-resolution RGB reconstruction error by 24.2%, while retaining interactive inference. Future work will broaden signer and motion coverage and include evaluation by proficient sign-language users.

References

[1] Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. Neural sign actors: A diffusion model for 3d sign language production from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3

[2] Diane Brentari, Joshua Falk, Anastasia Giannakidou, Annika Herrmann, Elisabeth Volk, and Markus Steinbach. Production and Comprehension of Prosodic Markers in Sign Language Imperatives. *Frontiers in Psychology*, 2018. 1

[3] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Yanjun Wang, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. Smpler-x: Scaling up expressive human pose and shape estimation, 2023. 3

[4] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022. 3

[5] Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025. 2

[6] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2Sign: A Large-Scale Multimodal Dataset for Continuous American Sign Language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[7] Sen Fang, Chen Chen, Lei Wang, Ce Zheng, Chunyu Sui, and Yapeng Tian. Signllm: Sign language production large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 6681–6693, 2025. 1, 2, 3

[8] Sen Fang, Chunyu Sui, Yanghao Zhou, Xuedong Zhang, Hongbin Zhong, Yapeng Tian, and Chen Chen. Signdiff: Diffusion model for american sign language production. In *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–11, 2025. 2

[9] Mengyang Feng, Jinlin Liu, Kai Yu, Yuan Yao, Zheng Hui, Xiefan Guo, Xianhui Lin, Haolan Xue, Chen Shi, Xiaowen Li, Aojie Li, Xiaoyang Kang, Biwen Lei, Miaomiao Cui, Peiran Ren, and Xuansong Xie. Dreamoving: A human video generation framework based on diffusion models, 2023. 2

[10] Shester Gueuwou, Sophie Siake, Colin Leong, and Mathias Müller. Jwsign: A highly multilingual corpus of bible translations for more diversity in sign language processing. *arXiv preprint arXiv:2311.10174*, 2023. 2

[11] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[12] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 3

[13] Hezhen Hu, Zhiwen Fan, Tianhao Wu, Yihan Xi, Seoyoung Lee, Georgios Pavlakos, and Zhangyang Wang. Expressive gaussian human avatars from monocular rgb video. In *NeurIPS*, 2024. 2, 3

[14] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 8

[15] Shoukang Hu, Tao Hu, and Ziwei Liu. GauHuman: Articulated Gaussian Splatting from Monocular Human Videos . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20418–20431, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 2, 3

[16] Wencan Huang, Wenwen Pan, Zhou Zhao, and Qi Tian. Towards Fast and High-Quality Sign Language Production. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 2, 3

[17] Eui Jun Hwang, Jung-Ho Kim, and Jong C. Park. Non-autoregressive sign language production with gaussian space. In *The 32nd British Machine Vision Conference (BMVC 21)*. British Machine Vision Conference (BMVC), 2021. 2

[18] Maksym Ivashechkin, Oscar Mendez, and Richard Bowden. Improving 3d pose estimation for sign language. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5, 2023. 3

[19] Maksym Ivashechkin, Oscar Mendez, and Richard Bowden. Two hands are better than one: Resolving hand to hand intersections via occupancy networks. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–9, 2024. 3

[20] Maksym Ivashechkin, Oscar Mendez, and Richard Bowden. Signsplat: Rendering sign language via gaussian splatting, 2025. 2, 3

[21] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *arXiv preprint arXiv:2306.14795*, 2023. 2

[22] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[23] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. NeuMan: Neural human radiance field from a single video. In *ECCV*, 2022. 3

[24] Yuheng Jiang, Zhehao Shen, Penghao Wang, Zhuo Su, Yu Hong, Yingliang Zhang, Jingyi Yu, and Lan Xu. Hifi4g: High-fidelity human performance rendering via compact gaussian splatting. In *Proceedings of the IEEE/CVF Confer-*

ence on Computer Vision and Pattern Recognition (CVPR), pages 19734–19745, 2024. 2, 3

[25] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 2, 3

[26] Michael Kipp, Quan Nguyen, Alexis Heloir, and Silke Matthes. Assessing the Deaf User Perspective on Sign Language Avatars. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*, 2011. 2, 3

[27] Youngjoong Kwon, Baole Fang, Yixing Lu, Haoye Dong, Cheng Zhang, Francisco Vicente Carrasco, Albert Mosella-Montoro, Jianjin Xu, Shingo Takagi, Daeil Kim, Aayush Prakash, and Fernando De la Torre. Generalizable human gaussians for sparse view synthesis. In *Computer Vision – ECCV 2024*, pages 451–468, Cham, 2025. Springer Nature Switzerland. 3

[28] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models, 2023. 2, 3

[29] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1459–1469, 2020. 2

[30] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 2, 3

[31] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, 2004. 6

[32] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21159–21168, 2023. 3

[33] Xinqi Liu, Chenming Wu, Jialun Liu, Xing Liu, Jinbo Wu, Chen Zhao, Haocheng Feng, Errui Ding, and Jingdong Wang. Gva: Reconstructing vivid 3d gaussian avatars from monocular videos, 2024. 3

[34] Jianhe Low, Alexandre Symeonidis-Herzig, Maksym Ivashechkin, Ozge Mercanoglu Sincan, and Richard Bowden. Signspark: Efficient multilingual sign language production via sparse keyframe learning, 2026. 2, 3

[35] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos, 2024. 2

[36] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3D gaussian avatar. In *ECCV*, 2024. 8

[37] Amit Moryossef, Gerard Sant, and Zifan Jiang. Pose-based sign language appearance transfer, 2025. 3

[38] Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. Ash: Animatable gaussian splats for efficient and photoreal human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1165–1175, 2024. 2, 3

[39] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *ACL*, 2002. 6

[40] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[41] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024. 3

[42] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation, 2025. 3

[43] Roland Pfau, Josep Quer, et al. *Nonmanuals: Their Grammatical and Prosodic Roles*. 2010. 1

[44] Lorenza Prospero, Abdullah Hamdi, Joao F. Henriques, and Christian Rupprecht. Gst: Precise 3d human body from a single image with gaussian splatting transformers, 2024. 3

[45] Lingteng Qiu, Peihao Li, Heyuan Li, Qi Zuo, Xiaodong Gu, Yuan Dong, Weihao Yuan, Rui Peng, Siyu Zhu, Xiaoguang Han, Guanying Chen, and Zilong Dong. Lhm++: An efficient large human reconstruction model for pose-free images to 3d, 2026. 8

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 3

[47] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Adversarial Training for Multi-Channel Sign Language Production. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2020. 2

[48] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Everybody sign now: Translating spoken language to photo realistic sign language video, 2020. 2

[49] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Progressive Transformers for End-to-End Sign Language Production. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[50] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Mixed signals: Sign language production via a mixture of motion primitives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1919–1929, 2021. 2

[51] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. AnonySign: Novel Human Appearance Synthesis for Sign Language Video Anonymisation. *arXiv preprint arXiv:2107.10685*, 2021. 2

[52] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Continuous 3D Multi-Channel Sign Language Production via Progressive Transformers and Mixture Density Net-

works. *International Journal of Computer Vision (IJCV)*, 2021. 1, 2

[53] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Skeletal Graph Self-Attention: Embedding a Skeleton Inductive Bias into Sign Language Production. *arXiv preprint arXiv:2112.05277*, 2021. 2

[54] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5141–5151, 2022. 2

[55] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 3, 8

[56] Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. X-avatar: Expressive human avatars. *Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[57] Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. Text2Sign: Towards Sign Language Production using Neural Machine Translation and Generative Adversarial Networks. *International Journal of Computer Vision (IJCV)*, 2020. 1, 2

[58] Stephanie Stoll, Simon Hadfield, and Richard Bowden. Signsynth: Data-driven sign language video generation. In *European Conference on Computer Vision*, pages 353–370. Springer, 2020. 2

[59] Garrett Tanzer and Biao Zhang. Youtube-sl-25: A large-scale, open-domain multilingual sign language parallel corpus. *ArXiv*, abs/2407.11144, 2024. 2

[60] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model, 2022. 2

[61] Harry Walsh, Ben Saunders, and Richard Bowden. Select and reorder: A novel approach for neural sign language production. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14531–14542, Torino, Italia, 2024. ELRA and ICCL. 2

[62] Harry Walsh, Ed Fish, Ozge Mercanoglu Sincan, Mohamed Ilyes Lakhal, Richard Bowden, Neil Fox, Bencie Woll, Kepeng Wu, Zecheng Li, Weichao Zhao, Haodong Wang, Wengang Zhou, Houqiang Li, Shengeng Tang, Jiayi He, Xu Wang, Ruobei Zhang, Yaxiong Wang, Lechao Cheng, Meryem Tasyurek, Tugce Kiziltepe, and Hacer Yalim Keles. Slrtp2025 sign language production challenge: Methodology, results, and future work, 2025. 2, 3

[63] Harry Thomas Walsh, Ben Saunders, and Richard Bowden. Sign stitching: A novel approach to sign language production, 2024. 2

[64] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

[65] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot Video-to-Video Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3

[66] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9168–9178, 2021. 5, 1

[67] Zhou Wang and Alan C. Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, 2009. 6

[68] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 2004. 4, 6

[69] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, 2022. 3

[70] Pan Xie, Qipeng Zhang, Taiyi Peng, Hao Tang, Yao Du, and Zexian Li. G2p-ddm: Generating sign pose sequence from gloss sequence with discrete diffusion model, 2023. 2, 3

[71] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1941, 2024. 2, 3

[72] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 3

[73] Aoxiong Yin, Haoyuan Li, Kai Shen, Siliang Tang, and Yueting Zhuang. T2s-gpt: Dynamic vector quantization for autoregressive sign language production from text, 2024. 2, 3

[74] Wanqi Yin, Zhongang Cai, Ruisi Wang, Ailing Zeng, Chen Wei, Qingping Sun, Haiyi Mei, Yanjun Wang, Hui En Pang, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Atsushi Yamashita, Lei Yang, and Ziwei Liu. Smplest-x: Ultimate scaling for expressive human pose and shape estimation, 2025. 7

[75] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3

[76] Jan Zelinka and Jakub Kanis. Neural Sign Language Synthesis: Words Are Our Glosses. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020. 2

[77] Dongbin Zhang, Yunfei Liu, Lijian Lin, Ye Zhu, Yang Li, Minghan Qin, Yu Li, and Haoqian Wang. Guava: Generalizable upper body 3d gaussian avatar. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14205–14217, 2025. 2, 3, 6, 8, 1

[78] Han Zhang, Rotem Shalev-Arkushin, Vasileios Baltatzis, Connor Gillis, Gierad Laput, Raja Kushalnagar, Lorna C Quandt, Leah Findlater, Abdelkareem Bedri, and Colin Lea. Towards ai-driven sign language generation with non-manual markers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2025. Association for Computing Machinery. 1, 3

[79] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 3

[80] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model, 2022. 2

[81] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[82] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *CoRR*, abs/2303.16199, 2023. 3

[83] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation, 2023. 2

[84] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance, 2025. 2

[85] Wojciech Zielonka, Timur M. Bagautdinov, Shunsuke Saito, Michael Zollhofer, Justus Thies, and Javier Romero. Drivable 3d gaussian avatars. *ArXiv*, abs/2311.08581, 2023. 2, 3

[86] Ronglai Zuo, Rolandos Alexandros Potamias, Evangelos Ververas, Jiankang Deng, and Stefanos Zafeiriou. Signs as tokens: A retrieval-enhanced multilingual sign language generator. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23806–23816, 2025. 2, 3

Stable Signer V3: 4D Deformable Primitive Splatting Is All You Need for Sign Language Production

Supplementary Material

Region	Restricted FLAME (mm)↓	Face primitives (mm)↓	Reduction (%)↑
Full head	1.267	**0.362**	71.4
Face	1.492	**0.518**	65.3
Lips	1.739	**0.990**	43.1
Left eye	1.029	**0.587**	43.0
Right eye	1.483	**0.626**	57.8
Forehead	1.047	**0.339**	67.6
Nose	1.009	**0.477**	52.7

Table 6. **Region-wise face reconstruction error.** Mean Euclidean point error on the held-out 20% of the tracked sequence. Both methods use identical articulation and coordinates.

Model	RGB L1↓	Boundary L1↓	SSIM↑	Disp. (m)↓
Frozen carrier	0.0649	0.0350	0.9038	–
K=512, matched budget	0.0599	0.0282	0.9076	0.0111
K=1,024, matched budget	0.0575	0.0270	0.9096	0.0111
K=1,024, final checkpoint	**0.0562**	**0.0230**	**0.9115**	0.0125

Table 7. **Body adaptation over the held-out segment.** The first three rows reproduce the controlled capacity study; the final row reports the converged model on the complete held-out segment.

Method	RGB L1↓	Boundary L1↓	SSIM↑
Frozen carrier	0.0661	0.0308	0.9086
Body primitives	**0.0551**	**0.0200**	**0.9155**

Table 8. **Ten-identity body adaptation.** Macro-average over temporally held-out segments. The same model configuration and evaluation code are used for all identities.

6. Evaluation Protocols

We expand the evaluation protocols, component analyses, and deployment measurements presented in the main paper. Supervision and evaluation occupy temporally disjoint segments, and checkpoint selection uses the supervision segment exclusively. Paired methods share motion, camera, background, crop, and target data. The controlled face and body studies use fixed optimization budgets.

6.1. Face primitive evaluation

We evaluate facial deformation on the tracked FLAME sequence distributed with GUAVA [77]. The first 80% trains the shared primitive controller, and the final 20% measures generalization to later motion. Restricted-FLAME and full-expression geometry share the tracked global pose, jaw, eye, and eyelid articulation, assigning their displacement directly to expression detail. We compute Euclidean error in the common tracked coordinates.

The controller contains K=1,024 local primitives, four-neighbor blending, and a compact residual MLP with 2.42M trainable parameters. We optimize it for 1,200 iterations using batch size 8 and a learning rate of 2×10^{-3}. Table 6 resolves the aggregate head error into semantic regions. Face primitives reduce error across the lips, eyes, forehead, and nose, confirming that the improvement spans the expressive facial surface.

6.2. Body adaptation evaluation

The body capacity study uses a chronological 80%/20% split of one identity-specific reference video. We uniformly sample a fixed evaluation subset and train the K=512 and K=1,024 models for the same 249-step budget. Both configurations share four-neighbor blending, optimizer, supervision, and loss weights. Their objective combines raw and refined RGB reconstruction, silhouette and boundary alignment, image gradients, SSIM, anchor regularization, and spatial smoothness.

Table 7 further evaluates the final K=1,024 checkpoint over the complete held-out segment. With the carrier and renderer fixed, continued adaptation improves appearance and boundary alignment. Its 1.25-cm mean displacement shows that compact local corrections are sufficient to capture identity-specific body geometry.

We then fit the same K=1,024 configuration independently to ten characters. Table 8 reports macro-averages over their complete held-out segments. Body primitives improve refined RGB L1, boundary L1, and SSIM for every character, demonstrating consistent per-character adaptation across the full identity set.

7. Detail Tunneling Analysis

Native-resolution reference video supervises the detail module with the Gaussian predictor and StyleUNet [66] backbone fixed. Character preparation bakes the learned identity parameters into the decoder, and online sequences run directly through this personalized model. The design concentrates reference evidence into a one-time adaptation stage and preserves a single-stream inference path.

Table 9 compares the trainable alternatives under a shared 900-step budget. LoRA and Adapter both improve the frozen decoder; the independent late Softmax pathway achieves the largest reconstruction gain. Increasing mem-

Method	RGB L1↓	PSNR↑
Frozen decoder	0.0260	22.16
LoRA (r=32)	0.0234	22.62
Adapter (r=32)	0.0225	22.90
Tunneling (r=8, 64 slots)	**0.0197**	23.77
Tunneling (r=8, 128 slots)	0.0196	**23.85**

Table 9. **Equal-budget detail adaptation.** Every trainable method is evaluated after 900 update steps on the same identity, held-out frames, Gaussian features, and native-resolution targets.

ory from 64 to 128 slots adds only 0.08 dB. We therefore adopt rank-8 tunneling with 64 slots as the default operating point.

At convergence, the selected 64-slot model reaches 0.0181 RGB L1 and 24.43 dB PSNR; the 128-slot model reaches 0.0178 and 24.48 dB. This 0.05-dB gap shows that 64 slots capture nearly all of the detail gain, supporting the selected default.

8. Per-Identity 4K Results

Table 10 reports native-4K results for all ten identities. Each renderer receives identical Gaussian features and deterministic aligned 768×768 crops from held-out 4K targets. We compute each identity score first and then macro-average, giving every character equal weight.

	PSNR↑		SSIM↑		LPIPS↓	
Identity	GUAVA	Ours	GUAVA	Ours	GUAVA	Ours
S01	20.146	**22.956**	0.8487	**0.8715**	0.2243	**0.1988**
S02	18.903	**21.538**	0.8452	**0.8719**	0.2123	**0.1900**
S03	20.324	**23.713**	0.8316	**0.8715**	0.2189	**0.2004**
S04	19.554	**22.304**	0.8105	**0.8556**	0.2373	**0.2220**
S05	19.213	**21.766**	0.8460	**0.8689**	0.2294	**0.2049**
S06	20.653	**22.549**	0.8483	**0.8670**	0.2453	**0.2220**
S07	18.199	**19.631**	0.8240	**0.8387**	0.2862	**0.2748**
S08	18.434	**20.675**	0.7829	**0.8195**	0.2873	**0.2744**
S09	19.557	**21.838**	0.8421	**0.8664**	0.2485	**0.2273**
S10	19.198	**20.928**	0.8527	**0.8687**	0.2645	**0.2440**
Mean	19.418	**21.790**	0.8332	**0.8600**	0.2454	**0.2259**

Table 10. **Per-identity native-4K rendering fidelity.** The adapted renderer improves all three metrics for every evaluated identity. Identity labels are anonymized and fixed across columns.

9. Runtime and Deployment

Stage throughput and latency. On one Quadro RTX 8000, warmed batch-one throughput reaches 198.7 FPS for motion synthesis, 320.9 FPS for primitive deformation, 52.0 FPS for avatar rendering, and 47.6 FPS for 4K detail refinement. The resident pipeline overlaps stages across chunks. A warmed 784-frame, three-chunk text-to-HD-avatar request completes in 31.0 seconds, yielding 25.3 FPS across composition, diffusion, streaming render, encoding, and concatenation. This wall-clock measurement captures the concurrency of the deployed pipeline.

One-time character preparation. The shared face controller transfers directly across characters. Character preparation fits the body and detail modules once and stores them with the avatar. Each final K=1,024 body model contains 2.43M trainable parameters and occupies 9.6 MB. Across ten identities, body adaptation averages 40.2 minutes on one Quadro RTX 8000, with a 22.6–53.5 minute range driven mainly by clip length. Online inference loads the resulting character package directly.

Meaning of 4D deformable primitives. In our formulation, *4D* indexes primitive transformations by pose and time throughout a motion sequence. The primitives supply local deformable support for the surface-bound Gaussian avatar, and Gaussian splatting rasterizes the resulting features. This factorization adds expressive local control to the pretrained parametric carrier while preserving its avatar representation and rendering path.

10. Interactive User Guide

Figures 8 and 9 demonstrate the complete generation workflow in the Stable Signer V3 interface. A user can begin with English text, an instruction-conditioned prompt, or an editable gloss sequence, then select an avatar identity from its reference thumbnail. After a single prewarm operation, *Generate Sign Language Avatar* executes gloss composition, body, hand, and face motion synthesis, followed by Gaussian-avatar rendering. The interface streams intermediate results and exposes the generated SMPL-X motion for direct inspection. The same job can then be viewed as the Gaussian avatar on black or white backgrounds, while completed requests remain available under *Recent Generations*. This interface turns the full text-to-4K-avatar system into a short, inspectable workflow and requires no manual exchange of intermediate files.

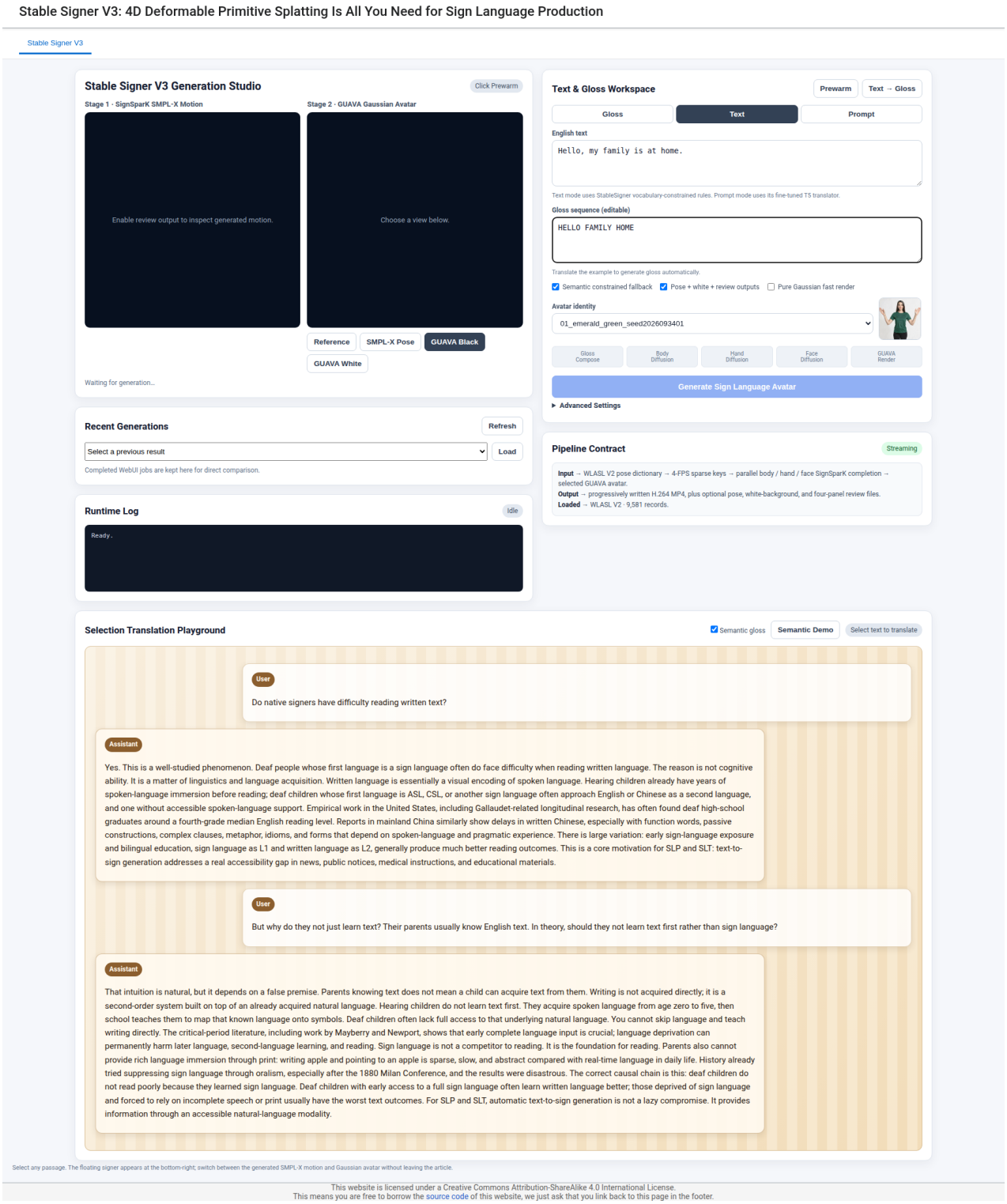
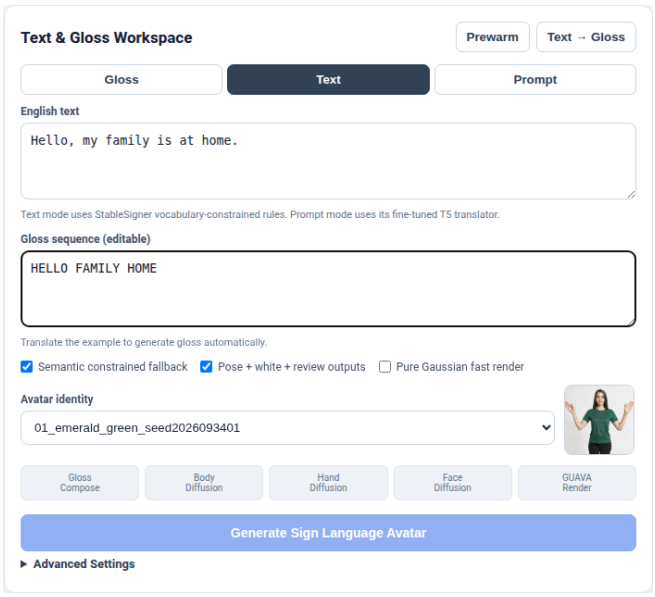


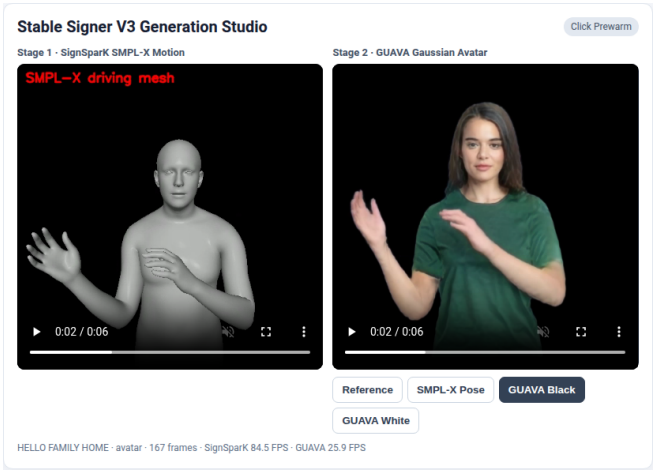
Figure 8. **Stable Signer V3 interface overview.** The complete interface presents text or gloss input, signer identity, staged motion and avatar outputs, execution state, and reusable results within one workspace.



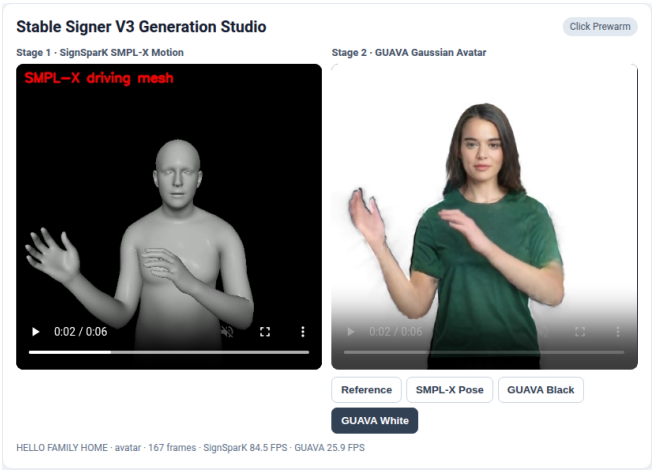
(a) Enter text or gloss and select the signer identity.



(b) Inspect the synthesized SMPL-X signing motion.



(c) Review the Gaussian avatar rendering.



(d) Switch to the white-background final output.

Figure 9. **Stable Signer V3 interactive generation workflow.** The interface exposes input translation, identity selection, stage-level progress, motion inspection, and final avatar outputs in one view. The example loads the same completed generation across all output modes.

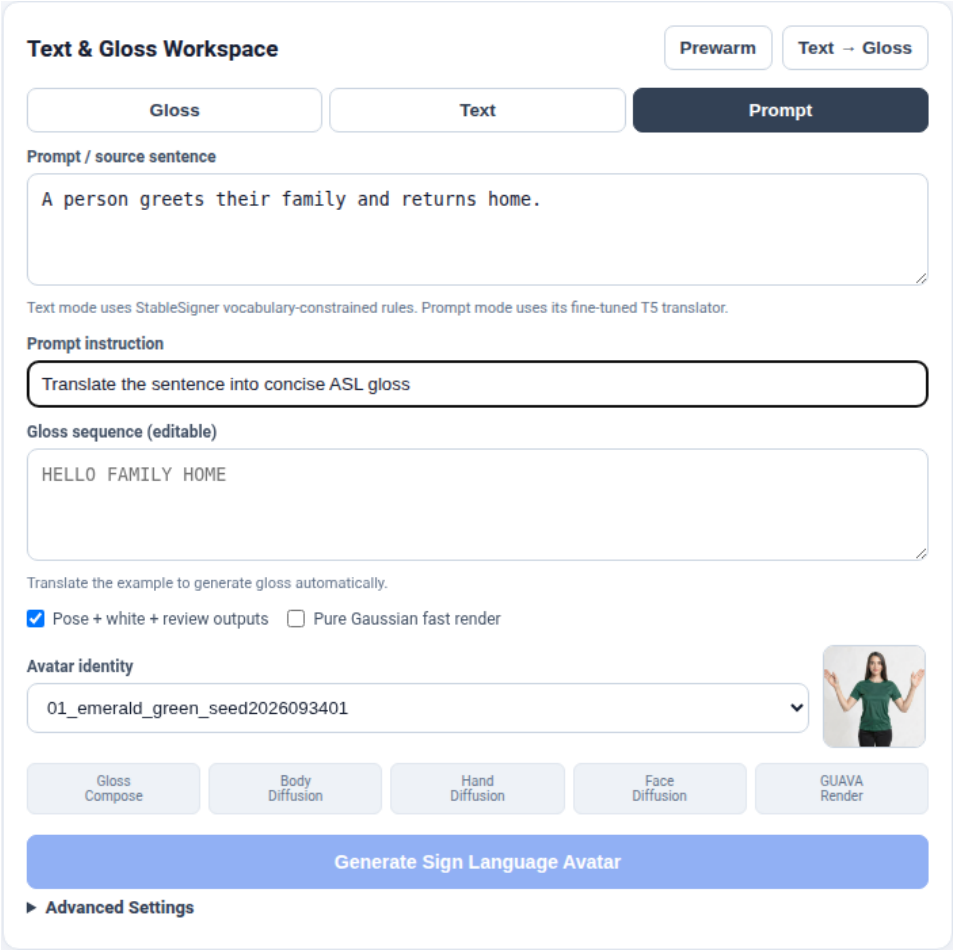
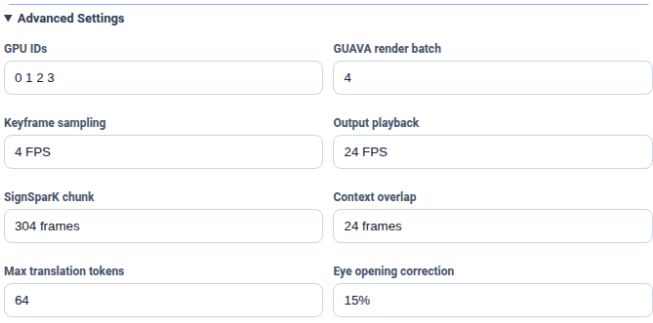
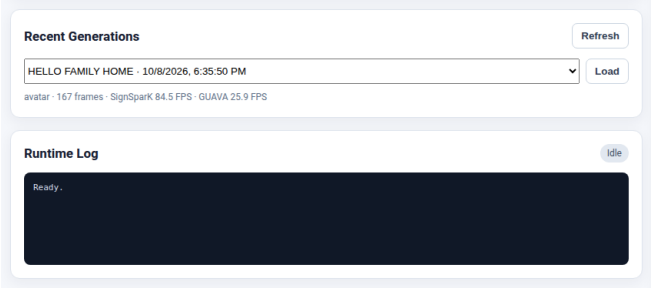


Figure 10. **Instruction-conditioned generation.** Prompt input, editable gloss, semantic fallback, signer identity, and the complete generation command remain visible within one uninterrupted workspace.

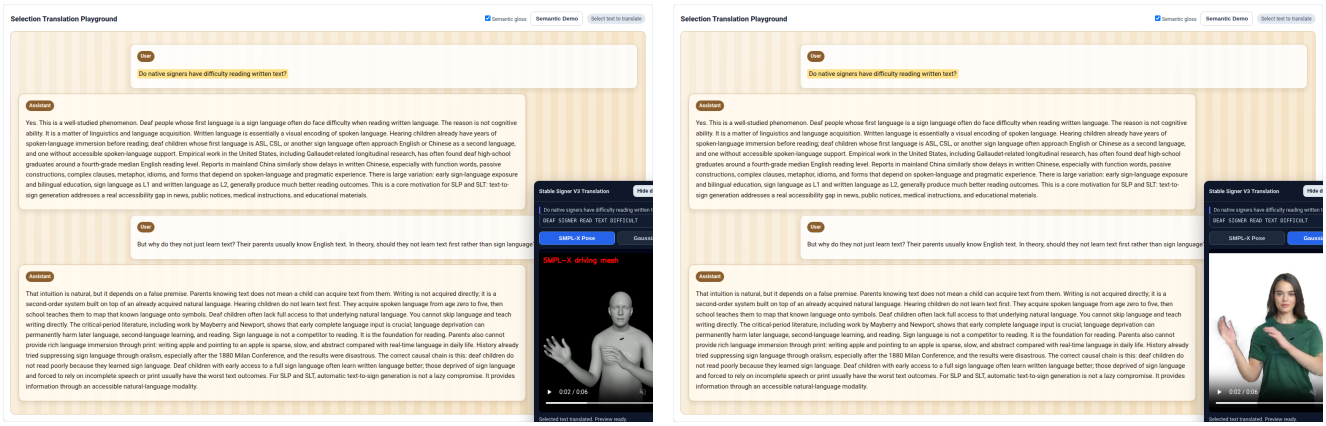


(a) Runtime controls for devices, batching, temporal sampling, and sequence overlap.



(b) Persistent generation history, measured throughput, and asynchronous runtime logs.

Figure 11. **Execution and inspection controls.** Stable Signer V3 exposes reproducible runtime settings and keeps completed generations available for immediate comparison.



(a) Selected text opens a bottom-right signer with SMPL-X motion playback.

(b) The same selection switches directly to the rendered Gaussian avatar.

Figure 12. **Contextual text-to-sign access.** Highlighting a passage invokes the floating signer in place and provides both its 3D motion and photorealistic avatar views without leaving the document.