Journal of the Brazilian Computer Society, 2026, 32:1, doi: 10.5753/jbcs.2026.8229

SPL4SH: Designing a Systematic Pipeline for 4D Synthetic Humans in XR Content Creation

David Ohara de Souza Cardoso [**Voxar Labs, Centro de Informática, UFPE** | ***dosc@cin.ufpe.br***]
Willams de Lima Costa [**Voxar Labs, Centro de Informática, UFPE** | ***wlc2@cin.ufpe.br***]
Pedro Azevedo Abrantes de Oliveira [**Voxar Labs, Centro de Informática, UFPE** | ***paao@cin.ufpe.br***]
João Marcelo Xavier Natário Teixeira [**University of São Paulo and Voxar Labs, Centro de Informática, UFPE** | ***jmxnt@cin.ufpe.br***]
Veronica Teichrieb [**Voxar Labs, Centro de Informática, UFPE** | ***vt@cin.ufpe.br***]
Qian Lin [**HP Inc.** | ***qian.lin@hp.com***]

Voxar Labs, Centro de Informática, Universidade Federal de Pernambuco, Av. Jornalista Anibal Fernandes, s/n, Cidade Universitária, Recife - PE, CEP: 50740-560

Received: 16 May 2026 • **Accepted:** 05 August 2026 • **Published:** 20 August 2026

Abstract XR content creation increasingly combines reconstruction, generative modeling, animation synthesis, neural rendering, and real-time engine deployment. However, creating deployable synthetic 4D humans remains fragmented because human assets must preserve body structure, appearance, motion, deformation, and scene-level plausibility across time. This paper proposes SPL4SH, a Systematic Pipeline for 4D Synthetic Humans, to investigate how contemporary AI-assisted techniques can be organized into an XR-oriented production workflow. SPL4SH is built from a stage-based taxonomy covering modeling, rigging, animation, rendering, and interaction. This taxonomy supports a comparative suitability assessment of recent techniques and guides a modular proof-of-concept implementation integrating SMPLify for parametric body modeling and rigging, Kimodo for controllable motion generation, SMPLitex for SMPL-compatible texture generation, Rokoko-based retargeting, Blender-based asset integration, and Unreal Engine deployment. The evaluation combines individual module tests, full-pipeline integration, and scene-level deployment in a meeting-room environment involving human-object, human-human, and human-scene arrangements. Results show that modular 4D human generation is technically feasible: SMPL-X geometry supports rigged deformation, generated motions can be retargeted to animated characters, texture maps improve mannequin-like bodies, and final assets can be placed in real-time XR scenes. However, persistent barriers remain, including format mismatch, manual skeleton alignment, retargeting fragility, texture discontinuities, hallucinated visual regions, motion interpenetration, stiff transitions, and scene-level validation. SPL4SH contributes a grounded framework for selecting, combining, and evaluating independent generative components as XR-ready synthetic human assets.

Keywords: Synthetic 4D humans, XR content creation, Generative AI, Modular pipelines, Digital humans, Spatial intelligence

1 Introduction

XR content creation is increasingly moving beyond fully manual asset production toward hybrid workflows that combine capture, reconstruction, neural rendering, and generative authoring. Volumetric video, neural radiance fields, Gaussian splatting, and AI-assisted 3D generation have expanded the ways in which immersive scenes can be created, reconstructed, and rendered in real time. However, creating deployable human characters for XR remains especially challenging. Unlike static objects or background environments, synthetic humans must preserve identity, geometry, appearance, motion, and interaction plausibility over time. They must also be compatible with real-time engines, animation tools, and spatial scenes in which users can inspect them from different viewpoints and interaction distances [Smolic *et al*., 2022; Heagerty *et al*., 2024; Weng *et al*., 2024; Hu *et al*., 2024a; Sakashita *et al*., 2024; Uzun *et al*., 2025; Zhu *et al*., 2026].

In parallel, digital human generation has progressed from static 3D avatar creation toward temporally coherent 4D human synthesis. Recent methods have improved text- or image-conditioned avatar generation, 360-degree human video synthesis, identity preservation, motion consistency, and pose-dependent deformation. These advances indicate that AI-generated humans are becoming increasingly realistic and controllable, but they also expose persistent challenges in fidelity, temporal coherence, editability, efficiency, and deployment [Cao *et al*., 2024; Shao *et al*., 2024, 2025; Cao *et al*., 2025; Miao *et al*., 2025].

This progress is directly relevant to XR. In immersive environments, synthetic humans are not merely visual assets: they can act as embodied agents, simulated participants, social interlocutors, or controlled sources of human-centered variation for perception and spatial-intelligence datasets. For this reason, an XR-ready 4D human must be more than a plausible rendered body, combining a usable body representation, controllable deformation, temporally coherent motion, coherent appearance, and sufficient interaction grounding to function inside spatial scenes. We use the term XR-oriented intelligent spaces to refer to immersive environments in which synthetic humans support spatial reasoning, simulation, interaction, or perception-oriented analysis [Li *et al.*, 2025b; Miao *et al.*, 2025; Shao *et al.*, 2024; Cao *et al.*, 2025; Dai *et al.*, 2025].

Despite rapid progress in individual techniques, current 4D human generation remains fragmented from the perspective of XR content creation. Mesh-based parametric models provide editability and rigging compatibility, neural and Gaussian representations improve visual fidelity, motion-generation methods support controllable behavior, and texture-generation methods can accelerate appearance authoring. Yet these capabilities are rarely available as a single, coherent workflow that begins with body or appearance conditioning and ends with a deployable, animated human in an XR scene. As a result, designers and researchers must still make difficult integration choices involving representation format, rig compatibility, motion retargeting, texture coherence, engine deployment, and scene-level validation [Kerbl *et al.*, 2023; Long *et al.*, 2024; Hong *et al.*, 2024; Li *et al.*, 2025b; Miao *et al.*, 2025; Shao *et al.*, 2025; Cao *et al.*, 2025].

To address this gap, we propose SPL4SH, a Systematic Pipeline for 4D Synthetic Humans. SPL4SH organizes 4D human creation around five production stages: modeling, rigging, animation, rendering, and interaction. Based on this structure, we analyze recent techniques and instantiate a modular proof-of-concept pipeline that combines parametric body modeling, texture generation, motion synthesis, retargeting, and scene deployment for XR-oriented workflows. The pipeline is evaluated through individual module tests, full-pipeline integration, and scene deployment in Unreal Engine, with particular attention to the practical failures that emerge when independent generative techniques are combined into a single XR asset workflow.

The main contributions of this paper are:

- a stage-based taxonomy of XR-oriented 4D synthetic human creation, organized around modeling, rigging, animation, rendering, and interaction;
- a structured comparison of recent techniques according to their suitability for these stages;
- the design and proof-of-concept evaluation of SPL4SH, a modular pipeline integrating parametric body modeling, texture generation, motion synthesis, retargeting, Blender-based integration, and Unreal Engine deployment;
- an analysis of the main integration barriers that remain for XR deployment, including format mismatch, retargeting fragility, texture artifacts, motion interpenetration, and scene-level validation.

In light of this objective, the study is organized around the following guiding questions:

- **Q1:** Which production stages are required to construct a functional 4D synthetic human for XR-oriented applications, and which categories of techniques are suitable for each stage?
- **Q2:** How can techniques from these stages be aligned into a modular AI-assisted workflow for 4D synthetic human creation?
- **Q3:** To what extent is such a workflow modular, scalable, and effective as a proof of concept, and which stages remain dependent on task-specific supervision, manual intervention, or representation-specific optimization?
- **Q4:** How feasible is the deployment of generated 4D humans into XR-oriented real-time scenes, considering engine integration, scene authoring, interaction alignment, and deployment constraints?
- **Q5:** Which application domains are most directly supported by synthetic 4D humans, beyond spatial-intelligence scenarios, including digital content creation (DCC), games, training and simulation, telepresence, scientific research, and synthetic dataset generation?
- **Q6:** What technical and operational limitations still prevent realistic 4D synthetic humans from being broadly adopted in immersive XR systems?

2 Related Work

Recent surveys have mapped key aspects of the synthetic-human production landscape, but they typically address narrower technical or experiential dimensions than an XR-oriented 4D asset pipeline. Surveys on digital human facial capture and tracking describe workflows that span acquisition, facial codification, asset creation, tracking, solving, animation, rendering, and evaluation, showing how facial-performance pipelines have evolved from capture-driven graphics toward machine-learning-assisted digital humans [Vilchis *et al.*, 2023]. Broader surveys on 3D human avatar modeling organize the field around reconstruction and generation, including pixel-aligned implicit functions, neural radiance fields, Gaussian splatting, diffusion-based generation, and text-conditioned avatar creation, while identifying fidelity, controllability, representation choice, and generalization as persistent challenges [Wang *et al.*, 2024]. Text-to-3D surveys similarly situate avatar generation within the broader AIGC landscape, emphasizing the roles of NeRFs, differentiable rendering, diffusion models, and structured or unstructured 3D representations, while noting limitations in consistency, efficiency, controllability, and applicability [Li *et al.*, 2023]. In XR and virtual-human studies, reviews on virtual co-presence and virtual humans in VR examine how avatar fidelity, expressiveness, sociability, and embodiment affect presence and interaction [Korban and Li, 2022; Sangeeth Chandran *et al.*, 2026]. Related work on virtual-human expressiveness and realism similarly shows that users' perception of virtual humans depends not only on geometry and rendering quality, but also on facial behavior, embodiment, crowd context, and conversational agency [Knob *et al.*, 2024; Montanha *et al.*, 2023]. However, these studies focus primarily on perceptual, experiential, and behavioral dimensions rather than on the production-chain requirements for deployable 4D human assets. These works establish the importance of human representation, generative 3D modeling, and virtual human experience, but they do not explicitly analyze modeling, rigging, animation, rendering, interaction grounding, and engine deployment as a single synthetic human creation workflow.

Generative AI studies for human creation have also shifted from single-output avatar generation to controllable, animatable, and representation-aware systems. AvatarGen introduced a generative model for clothed, animatable human avatars by combining a coarse human body proxy, canonical-space deformation, pose and view control, and 2D image supervision [Zhang *et al.*, 2022]. DreamHuman connected text-to-image diffusion, neural radiance fields, and statistical human body models to synthesize animatable 3D humans from text, addressing appearance diversity and body-shape variation while exposing challenges in anthropometric consistency and controllable animation [Kolotouros *et al.*, 2023]. SEEAvatar further introduced geometry and appearance constraints, human priors, evolving templates, and diffusion-guided physically based texture generation to produce mesh-and-texture assets more compatible with conventional graphics pipelines [Xu *et al.*, 2026]. Text2Avatar addressed text-conditioned avatar creation by leveraging codebook-driven, body-controllable attributes and pseudo-data generation [Gong *et al.*, 2024]. These methods show that human-centered generative systems can combine body priors, neural fields, diffusion guidance, texture generation, and animation-oriented representations. However, they are usually evaluated as generation models rather than as complete production workflows in which body representation, motion, texture, retargeting, DCC tools, and real-time engines must be aligned.

XR studies involving synthetic or reconstructed humans show why deployable 4D humans require more than geometric plausibility. Full-body video-based self-avatar systems for mixed reality investigate constraints on capture, segmentation, communication, and embodiment, showing that MR avatars must satisfy both representation quality and runtime system requirements [Gonzalez Morin *et al.*, 2023]. X-Avatar frames digital humans as assets for telepresence, AR, VR, and related applications by modeling body, hands, facial expression, geometry, deformation, and appearance in a holistic SMPL-X-driven representation [Shen *et al.*, 2023]. HDHumans combines classical deforming character templates with neural radiance fields to improve novel-view rendering, temporal coherence, and generalization to unseen motion, with applications in immersive communication, gaming, entertainment, AR, and VR [Habermann *et al.*, 2023]. XR-oriented virtual-human reviews further indicate that fidelity, facial expressions, sociability, and embodiment influence the user experience, while embodied-agent studies emphasize interaction types, embodiment, and domains such as training, therapy, and social interaction [Sangeeth Chandran *et al.*, 2026]. These studies are highly relevant to XR, but they generally focus on avatar realism, embodiment, user-facing behavior, or capture-and-render systems rather than on modular asset production across body modeling, rig compatibility, motion transfer, texture coherence, authored interaction, and Unreal or Unity deployment.

This literature leaves a specific gap that the present study addresses. Existing surveys and generative systems clarify many components of digital-human creation, but less attention has been given to how contemporary 4D human techniques can be selected, combined, tested, and deployed as XR-ready assets. The current work addresses this gap by proposing the SPL4SH pipeline, which organizes synthetic 4D human creation around modeling, rigging, animation, rendering, and interaction, then using this structure to compare recent techniques, instantiate a modular proof-of-concept pipeline, and analyze the failures that appear when body modeling, texture generation, motion synthesis, retargeting, Blender-based integration, and Unreal Engine deployment are connected into a single workflow. The study is therefore positioned neither as a general survey of all 3D human modeling nor as a new monolithic generative model, but rather as a pipeline-oriented investigation into how independent AI-assisted components can be operationalized for XR content creation.

3 SPL4SH Design

3.1 Technique selection and evidence extraction

To support the design of SPL4SH, we conducted a structured, literature-informed, process to select techniques and extract evidence focused on creating synthetic 4D humans. This process was not intended to constitute a systematic review in the strict sense, nor to provide exhaustive coverage of all available 4D generation methods. Instead, its purpose was to identify representative and technically relevant approaches to inform the design of a modular, XR-oriented pipeline covering modeling, rigging, animation, rendering, and interaction. The selection process therefore prioritized techniques that exposed clear inputs, outputs, representations, control mechanisms, and potential integration roles in spatial intelligence and XR oriented applications, such as games, virtual tours, simulators and DCC, for example.

The identification stage combined query-based retrieval across four indexing and publishing platforms (arXiv, ACM Digital Library, IEEE Xplore, and Springer Nature) with targeted searches in specialized sources. The query-based search used the expression (“4D human” OR “4D animation” OR “4D reconstruction”) AND (“modeling” OR “reconstruction” OR “creation”) AND (“automated pipeline” OR “asset generation” OR “content generation”), considering works published between 2021 and 2026. Targeted searches were also conducted in the CVPR 2025 3D Humans Workshop, the Max Planck Institute research pages, and GitHub repositories related to human texturing and Stable Diffusion-based texture generation. These targeted sources were included because recent 4D human generation methods are often disseminated through preprints, workshop papers, project pages, and implementation repositories before appearing in indexed journals or conference proceedings.

The screening process was guided by the following inclusion criteria:

- IC1: the work proposes or evaluates a technique related to human-centered 3D or 4D modeling, reconstruction, rigging, animation, rendering, texture generation, interaction, or deployment;
- IC2: the technique produces, reconstructs, controls, animates, textures, or renders a human-centered representation that could contribute to a synthetic 4D human pipeline;
- IC3: the work provides sufficient technical information to identify its inputs, outputs, representation type, sup-

ported pipeline stage, and potential integration into XR or digital-content-creation workflows.

The exclusion criteria were:

- EC1: works focused only on non-human objects, static environments, or general 3D generation without a clear human-centered component;
- EC2: works limited to recognition, detection, or analysis tasks without generating, reconstructing, animating, rendering, or controlling human representations;
- EC3: works that did not provide enough methodological or implementation detail to assess their role in the proposed pipeline.

The filtering process followed three stages. First, the identification stage produced 74 records: 50 from the query-based search, one from the Max Planck Institute research pages, 20 from the CVPR 2025 3D Humans Workshop, and three from GitHub-related searches for stable diffusion based human texturing techniques. Second, title and abstract screening applied the inclusion and exclusion criteria described above, retaining 39 studies for full-paper review: 22 from the query-based search, 1 from the Max Planck Institute source, 13 from the CVPR workshop, and 3 from the GitHub-associated search. Third, full-paper reading examined whether each study provided sufficient technical evidence about its representation, input and output modalities, controllability, supported pipeline stage, and integration relevance. This stage reduced the set to 29 studies. Finally, three additional NVIDIA human motion generation works were added via targeted search due to their direct relevance to controllable motion generation and their practical importance in the animation stage. The final set comprised 32 studies used to support the SPL4SH taxonomy, stage-wise comparison, and qualitative suitability assessment.

For each included study, we extracted a common set of attributes into an evidence matrix: research objective, input modality, output representation, method overview, supported SPL4SH stages, representation type, controllability, visual fidelity, editability, compatibility with spatial intelligence, XR applications and DCC workflows, reported strengths, reported limitations, and possible role in a modular 4D human pipeline. These attributes were extracted from the paper text, reported experiments, figures, demonstrations, repositories, and implementation notes when available. The extracted evidence was then used to organize the taxonomy, compare techniques within stage-specific groups, and select representative methods for the proposed SPL4SH instantiation.

Large language model assistance was used only as a support tool to organize preliminary extraction notes and draft initial summaries from the reviewed papers. The authors retained responsibility for screening, categorization, and final interpretation. The assistant was not used to decide whether a study should be included or excluded, to assign final categories, or to determine the final suitability assessment. All preliminary notes produced with tool assistance were checked against the original papers, and unsupported, incorrect, or ambiguous entries were removed or corrected before being used in the taxonomy, comparison, or pipeline analysis.

3.2 Pipeline stages and design requirements

Before defining SPL4SH, we establish the production stages required to construct a functional 4D synthetic human for XR-oriented content creation. In this work, a 4D synthetic human is treated only as a temporally coherent human representation that can be modeled, controlled, animated, visually represented, and situated inside a spatial environment. The five-stage taxonomy was derived through an iterative synthesis of two sources of evidence: recurring technical roles observed in the selected literature and the practical requirements of XR and DCC workflows. Across the reviewed techniques, methods repeatedly addressed one or more of five functional questions: what representation defines the human body, how this representation is controlled or deformed, how it changes over time, how it is visually rendered, and how it is situated with respect to objects, other humans, or the surrounding scene. These questions correspond to modeling, rigging, animation, rendering, and interaction, respectively.

We therefore adopt these five stages because each one captures a necessary layer of an XR-ready 4D human asset: geometric structure, deformation control, temporal behavior, visual appearance, and scene-level situation. Other aspects, such as acquisition, prompting, file conversion, engine import, and evaluation, are treated as conditioning, integration, or validation steps rather than as additional top-level stages because they do not, by themselves, define a distinct layer of the synthetic human representation. The taxonomy was then checked against the selected studies by assigning each technique to one or more stages based on its primary technical contribution and potential role in a modular XR production workflow.

Table 1 summarizes the working taxonomy used in this paper to categorize the selected techniques. The taxonomy defines a stage-based analytical structure for comparing techniques, identifying missing capabilities, and composing the proposed SPL4SH pipeline. Its purpose is practical and design-oriented: to make explicit which asset layer or production responsibility each technique can support when constructing deployable synthetic humans for XR.

Modeling. Modeling defines the geometric and structural representation of the human body. This stage determines whether the human is represented as a parametric mesh, a non-parametric mesh, a point-based structure, a Gaussian representation, or a hybrid representation. For XR workflows, modeling is not only about reconstruction quality. The chosen representation must also support identity consistency, body-shape variation, editability, memory efficiency, and compatibility with downstream animation and engine integration.

Rigging. Rigging defines how the human representation can be controlled and deformed over time. In mesh-based workflows, this usually involves articulated skeletons, skinning weights, blendshapes, or parametric body controls. In neural or Gaussian-based workflows, explicit skeletons may be replaced or complemented by deformation fields, learned correspondences, or direct primitive transformations. For XR content creation, rigging is important because it determines whether motion can be retargeted, corrected, reused,

Table 1. Working taxonomy of production stages and technique categories for XR-oriented 4D synthetic human creation.

Stage	Category	Description
Modeling	Parametric polygonal meshes	Meshes controlled by compact body parameters with stable topology.
	Non-parametric polygonal meshes	Mesh geometry not constrained by fixed body templates.
	Gaussian splatting	Gaussian primitives encoding geometry and appearance.
	Hybrid mesh–Gaussian representations	Mesh-based control combined with Gaussian visual detail.
	Point clouds	Sampled 3D points representing surfaces or structures.
Rigging	Parametric rigging	Deformation control inherited from predefined body models.
	Non-parametric rigging	Custom or inferred controls independent of body templates.
	Forward kinematics	Parent-to-child joint transformations defining articulated pose.
	Linear blend skinning	Vertex deformation through weighted bone transformations.
	Blendshape and facial rigging	Facial motion driven by weighted expression deformation targets.
Animation	Keyframe interpolation	Sparse poses generate motion through interpolated intermediate states.
	Posed mesh sequences	Time-ordered posed meshes representing framewise dynamics.
	Parametric animation	Motion encoded as trajectories in a compact control space.
	Dynamic Gaussian splatting	Motion represented by time-varying Gaussian primitives.
	Audio-driven facial animation	Speech features drive synchronized facial motion.
Rendering	Texture-map rendering	Image textures mapped onto UV coordinates.
	Material rendering	Surface appearance represented through material properties and lighting response.
	Gaussian primitive coloring	Color and appearance encoded directly in Gaussian primitives.
Interaction	Human-object interaction (HOI)	Body poses and motions coordinated with manipulable objects.
	Human-human interaction (HHI)	Multiple bodies coordinated through social, spatial, or physical relations.
	Human-scene interaction (HSI)	Body behavior constrained by scene geometry, affordances, or occupancy.

and adapted to interaction constraints.

Animation. Animation specifies how the human changes over time. This stage includes keyframe interpolation, motion capture, motion generation, motion transfer, posed mesh sequences, facial animation, and dynamic neural or Gaussian representations. For XR deployment, the key issue is not only whether the motion looks plausible, but whether it is controllable, reusable, temporally coherent, and compatible with scene-level authoring. Motions that can be edited, blended, retargeted, or conditioned by interaction constraints are more useful for modular pipelines than isolated visual performances that are difficult to manipulate after generation.

Rendering. Rendering defines how the synthetic human becomes visually observable inside an XR scene. This stage includes texture maps, material definitions, lighting behavior, view-dependent appearance, neural rendering, Gaussian primitive coloring, and compatibility with real-time engines. In XR, rendering decisions affect perceptual presence, visual consistency across viewpoints, compatibility with dynamic scene lighting, and the ability to inspect the human at different distances. A rendering strategy must therefore balance visual fidelity with editability, relighting, performance, and deployment constraints.

Interaction. Interaction determines how a synthetic human is situated with respect to objects, other humans, and the surrounding scene. We consider three interaction levels: Human-Object Interaction (HOI), Human-Human Interaction (HHI), and Human-Scene Interaction (HSI). This stage is included because XR-oriented synthetic humans are not only isolated animated bodies, they must eventually preserve plausible contact, proximity, orientation, timing, and contextual relations inside spatial environments.

Importantly, interaction in SPL4SH refers to a production requirement rather than a guarantee of autonomous behavior. The level of interaction support may vary across implementations, ranging from manually authored scene placement and motion sequencing to more advanced forms of autonomous, adaptive, or physically grounded interaction. The proof-of-concept implementation presented in this paper addresses the former: authored scene-situated HOI, HHI, and HSI arrangements created through placement, timing, and motion sequencing. It does not demonstrate autonomous interaction policies, adaptive behavior, scene-goal inference, or physically grounded contact synthesis. These capabilities remain important targets for future SPL4SH instantiations.

Together, these stages provide the design basis for SPL4SH. They are used in the following subsections to categorize the selected techniques, compare their roles in the synthetic-human creation process, and identify combinations suitable for XR-oriented deployment. The resulting taxonomy should be understood as a working production taxonomy: it is grounded in the reviewed literature and XR authoring practice, but it remains open to extension as future techniques integrate, merge, or automate multiple stages.

3.3 Stage-wise comparison of selected techniques

After defining the SPL4SH stages, we compared the selected techniques according to the pipeline roles they can support. Direct comparison across all methods would be misleading because many techniques address different technical problems, even when they belong to the same broad stage. For example, two methods may both operate on parametric human models while pursuing different goals, such as garment generation in ChatGarment [Bian *et al*., 2025] or whole-body expressive recovery in Expressive Body Capture [Pavlakos *et al*., 2019]. For this reason, we organize the comparison by stage-specific semantic groups rather than by a single global ordering.

Table 2 summarizes the selected techniques across modeling, rigging, animation, rendering, interaction, and full-pipeline groups. Each row compares methods that share a similar technical responsibility, such as parametric body modeling, garment generation, rigged Gaussian avatars, dynamic Gaussian animation, texture-map rendering, or interaction grounding. The full-pipeline group is included separately because some systems combine multiple stages, but they do so with different assumptions, representations, and deployment targets.

The comparison shows that current 4D human generation is not a single unified capability, but a collection of complementary techniques distributed across the synthetic-human production process. Modeling methods primarily define the embodied substrate, ranging from standardized SMPL or SMPL-X bodies to garment-specific, Gaussian, and hybrid representations. Rigging methods determine how this substrate can be controlled, either through parametric body priors, unified model abstractions, skeletal deformation of Gaussian assets, or manually structured non-parametric rigs. Animation methods then define temporal behavior through generated motion, motion transfer, scene-conditioned movement, facial dynamics, cloth-aware motion, or dynamic Gaussian transformations. Rendering methods determine how identity and appearance are represented, either through conventional texture and material assets or through appearance encoded directly in Gaussian primitives. Interaction methods remain comparatively narrower, but they are essential for situating humans with respect to objects, other humans, and scene constraints.

Across these groups, five broader capabilities emerge. First, body-standardization methods support stable topology, editability, and downstream rigging. Second, appearance and garment-specialization methods address the visual incompleteness of body-only representations. Third, motion and temporal control methods provide either reusable or generated behavior, but vary considerably in controllability and scene awareness. Fourth, interaction-grounding methods are beginning to encode contact, proximity, spatial occupancy, and conversational responses, although this remains less mature than modeling, rendering, and animation. Fifth, full-pipeline systems demonstrate partial integration, but their notions of completeness differ: some are complete as parametric motion-authoring systems, others as capture-and-render pipelines, and others as interactive virtual-human systems.

This comparison motivates the modular design of SPL4SH. For XR-oriented content creation, the appropriate pipeline cannot be selected only by visual quality or by whether a method is called “4D”. It must be selected according to the intended function of the synthetic human: editable avatar production, photoreal dynamic rendering, retargetable motion, contact-aware reconstruction, scene-conditioned behavior, or real-time interactive presence. The next subsection, therefore, uses this stage-wise comparison to assess which techniques are most suitable for composing an XR-deployable 4D synthetic human pipeline.

3.4 Comparative suitability assessment

Following the stage-wise comparison, we selected eight representative techniques for a more detailed assessment of suitability. The goal of this assessment was not to identify an objectively superior method across all dimensions, but to determine which techniques are most appropriate for composing an XR-oriented 4D synthetic human pipeline. The selected techniques were chosen because they satisfy at least one of the following conditions: they provide foundational representations or tools used by later methods; they cover multiple SPL4SH stages; or they introduce distinctive capabilities relevant to XR deployment, such as controllable motion generation, cross-model compatibility, photoreal rendering, interaction grounding, or real-time virtual-human behavior.

Table 3 summarizes this assessment across stage-specific criteria. The suitability markers were assigned through a qualitative, evidence-based categorizing procedure. For each selected technique, we inspected the corresponding paper and, when available, its project page, repository, reported results, stated limitations, figures, demonstrations, and videos. A criterion was marked as $\times$ when it was outside the technique’s scope or depended on a feature not provided by the method. A criterion was marked as $\circ$ when the technique partially or indirectly supported the criterion, but was not among the strongest candidates in the selected set. A criterion was

Table 2. Condensed comparison of synthetic 4D human generation techniques across pipeline-stage groups and categories.

Group	Category	Comparison
Modeling	Parametric models	Expressive Body Capture [Pavlakos *et al*., 2019] is the explicit SMPL-X whole-body expressive baseline; MEGA [Fiche *et al*., 2025] shifts toward learned SMPL-topology recovery; PICO [Cseke *et al*., 2025] prioritizes contact-aware body-object consistency; Multi-HMR [Baradel *et al*., 2024] scales SMPL-X recovery to multiple humans in camera space; GENMO [Li *et al*., 2025a] uses SMPL mainly as a multimodal motion-oriented representation.
	Unified parametric models	SOMA [Saito *et al*., 2026] is an interoperability layer across SOMA, SMPL, SMPL-X, MHR, Anny, and garment measurements; Kimodo [Rempe *et al*., 2026] narrows support to SOMA, G1, and SMPLX, but turns the unified body representation toward controllable motion generation.
	Garment generation	HumanLiff [Hu *et al*., 2025] models layered, body-aligned clothing; ChatGarment [Bian *et al*., 2025] converts image/text prompts into garment JSON and sewing patterns; T2C [Yu *et al*., 2025] adds text-guided 4D cloth with body-cloth separation; CharacterGen [Peng *et al*., 2024] favors rapid non-parametric stylized character creation; Disco4D [Pang *et al*., 2025] represents clothing as Gaussian components over SMPL-X.
	Gaussian splatting	4D Facial Capture Pipeline [Tian *et al*., 2024] and CAP4D [Taubner *et al*., 2025] are facial/portrait specialists; Real-time Free-view [Sun *et al*., 2025] uses template-plus-Gaussian full-body rendering for sparse views; Expressive Gaussian [Hu *et al*., 2024b] emphasizes face/hand detail; GauSTAR [Zheng *et al*., 2025] handles topology-changing surfaces; Align Your Gaussians [Ling *et al*., 2024] generalizes text-guided dynamic 3D/4D Gaussian generation.
Rigging	Parametric rigging	Expressive Body Capture [Pavlakos *et al*., 2019] inherits unified body-hand-face SMPL-X rigging; Multi-HMR [Baradel *et al*., 2024] extends this rigging logic to multi-person camera-space scenes; GENMO [Li *et al*., 2025a] uses SMPL rigging primarily as a carrier for multimodal motion generation and estimation.
	Unified parametric rigging	SOMA [Saito *et al*., 2026] abstracts rigging across heterogeneous body-model families; Kimodo [Rempe *et al*., 2026] keeps the same interoperable intent but optimizes a smaller body set for controllable generated motion.
	Rigged gaussians	Expressive Gaussian [Hu *et al*., 2024b] and Vid2Avatar-Pro [Guo *et al*., 2025] use SMPL-X/LBS for controllable clothed Gaussian avatars, respectively emphasizing local expressivity and monocular full-body personalization; Disco4D [Pang *et al*., 2025] separates body and clothing control; CAP4D [Taubner *et al*., 2025] applies facial priors to portrait Gaussian animation.
	Non parametric rigging	Articulated Kinematics Distillation [Li *et al*., 2025c] manually constructs articulated skeletons for static 3D assets and transfers forward-kinematics/LBS control into Gaussian kernels, making it flexible for humanoid and non-humanoid assets but more author-intensive.
Animation	Parametric multi-model animation	SOMA [Saito *et al*., 2026] treats animation as motion reuse across body-model families; Kimodo [Rempe *et al*., 2026] adds diffusion-based motion generation, editable waypoints, and sequence composition for a smaller supported body set.
	Parametric animation	GENMO [Li *et al*., 2025a] supports text/video-conditioned SMPL motion; UniMotion [Li *et al*., 2024] unifies motion-language tasks; Revisit Human-Scene Interaction [Liu *et al*., 2023] conditions motion on scene occupancy; In-2-4D [Nag *et al*., 2025] synthesizes temporal continuity from sparse image endpoints through dynamic Gaussians.
	Clothing animation	T2C [Yu *et al*., 2025] separates body and clothes while incorporating cloth-aware motion; Disco4D [Pang *et al*., 2025] expresses clothing dynamics through SMPL-X-guided Gaussian splatting, privileging dynamic rendered plausibility.
	Face animation	4D Facial Capture Pipeline [Tian *et al*., 2024] reconstructs temporally coherent multi-view facial mesh sequences; Towards Highfidelity 3D Talking Avatar [Li *et al*., 2025d] adds audio-driven speech and dynamic textures; CAP4D [Taubner *et al*., 2025] produces facial-prior-driven Gaussian portrait avatars.
	Dynamic gaussian splatting	Real-time Free-view Human Rendering [Sun *et al*., 2025] and Vid2Avatar-Pro [Guo *et al*., 2025] are pose-driven animatable Gaussian humans; GauSTAR [Zheng *et al*., 2025] preserves topology-changing temporal coherence; Efficient4D [Pan *et al*., 2025] targets fast short-span single-view 4D generation; Gaussian See, Gaussian Do [Bekor *et al*., 2025] performs motion transfer; Articulated Kinematics Distillation [Li *et al*., 2025c] animates static Gaussian assets through distilled articulation.
Rendering	Texture map and material rendering	TexDreamer [Liu *et al*., 2024] and SMPLitex [Casas and Comino-Trinidad, 2023] generate complete SMPL UV textures, respectively from multimodal prompts and single images; CharacterGen [Peng *et al*., 2024] refines stylized canonical UVs; T2C [Yu *et al*., 2025] and HumanLiff [Hu *et al*., 2025] render clothing through cloth textures and layered tri-plane/NeRF-style humans; ChatGarment [Bian *et al*., 2025] produces garment assets ready for downstream rendering and simulation.
	Gaussian primitives coloring rendering	Expressive Gaussian [Hu *et al*., 2024b] and Vid2Avatar-Pro [Guo *et al*., 2025] render clothed Gaussian avatars with expressive or pose-dependent detail; Real-time Free-view [Sun *et al*., 2025] improves sparse-view photoreal rendering via double unprojection; Disco4D [Pang *et al*., 2025] optimizes disentangled body/clothing appearance; CAP4D [Taubner *et al*., 2025] targets expression-dependent portraits; Efficient4D [Pan *et al*., 2025] and Splat4D [Yin *et al*., 2025] emphasize fast or temporally consistent general dynamic 4D rendering.
Interaction	Human interaction	Kimodo [Rempe *et al*., 2026] enables pseudo-interaction through waypoint editing and motion composition; PromptHMR [Wang *et al*., 2025] reconstructs prompt-conditioned multi-person HHI with contact/proximity cues; PICO [Cseke *et al*., 2025] reconstructs dense HOI body-object contact but not animation; Revisit Human-Scene Interaction [Liu *et al*., 2023] generates scene-conditioned HSI motion under 3D occupancy constraints.
Full Pipeline	Full-pipeline systems	Kimodo [Rempe *et al*., 2026] and PromptHMR [Wang *et al*., 2025] cover large parametric-motion portions but remain visually incomplete without external texture/material work; Neural Free-Viewpoint [Sun *et al*., 2021] is stronger for capture-and-render HOI performances with texture completion and free-viewpoint rendering; An Interactive Conversational 3D Virtual Human [Shaw *et al*., 2026] combines Gaussian head-body avatar construction, real-time rendering, body/facial animation, and conversational interaction.

marked as ⋆ when the technique provided strong suitability for that criterion, based on explicit method design, reported results, visual evidence, or demonstrated compatibility with the corresponding SPL4SH stage.

The procedure was performed by a researcher and reviewed for consistency by a senior researcher. When the assignment was ambiguous, the final marker was chosen conservatively, giving preference to ∘ rather than ⋆ unless the technique explicitly demonstrated strong support for the criterion. These markers should therefore be interpreted as design-oriented qualitative suitability indicators rather than quantitative scores. Their purpose is to support pipeline design decisions by exposing complementarities and gaps among techniques.

The assessment reinforces the main conclusion of the previous comparison: no single technique satisfies all requirements of an XR-ready 4D human pipeline. Parametric body methods provide a strong foundation for modeling, deformation, editability, and engine compatibility, but they are limited when detailed clothing, hair, material appearance, or photorealistic rendering is required. Gaussian and neural rendering methods offer higher visual fidelity but often reduce editability, retargeting flexibility, and compatibility with conventional DCC or game-engine workflows. Motion-oriented methods improve animation control and retargeting, but they do not necessarily solve appearance generation or physical interaction. Interaction-oriented systems provide important progress toward situated behavior, but they remain narrower and more task-specific than modeling, rigging, or rendering methods.

The selected techniques, therefore, occupy complementary roles. SMPLify [Pavlakos *et al*., 2019] provides a stable parametric body and expressive rigging basis. PromptHMR [Wang *et al*., 2025] extends parametric reconstruction toward multi-person scenes and populated XR environments. SOMA [Saito *et al*., 2026] contributes cross-family body-model compatibility. Kimodo [Rempe *et al*., 2026] is most relevant for controllable motion authoring, motion composition, and retargeting. Gaussian See, Gaussian Do [Bekor *et al*., 2025] supports motion transfer for Gaussian-based assets, although it depends on compatible source and target representations. Neural Free-Viewpoint Performance Rendering [Sun *et al*., 2021] is strongest for captured human-object performances with clothed appearance and situated behavior. Articulated Kinematics Distillation [Li *et al*., 2025c] provides broad support for rigging and animation of static assets, but with weaker human-specific realism. ICo3D [Shaw *et al*., 2026] is most relevant when the target is a high-fidelity conversational virtual human with dynamic rendering and social response.

This suitability assessment motivates a modular rather than monolithic interpretation of the field. For XR-oriented synthetic human creation, the designer must select techniques based on the intended function of the generated human: editable avatar production, photorealistic dynamic rendering, retargetable motion, contact-aware reconstruction, scene-conditioned behavior, or real-time interactive presence. SPL4SH builds on this observation by composing techniques that cover complementary stages of the pipeline while explicitly exposing the integration barriers that remain between them.

3.5 SPL4SH pipeline synthesis

The previous subsections show that current 4D human generation methods are mostly complementary rather than self-sufficient. Some techniques provide stable and editable body representations, others support motion generation or retargeting, and others focus on visual appearance or interaction-specific behavior. Based on this observation, SPL4SH is designed as a modular pipeline rather than a single end-to-end generative model. Its goal is to combine techniques spanning different production stages while making explicit the integration points needed to transform independent outputs into an XR-deployable 4D human.

The proposed SPL4SH instantiation combines SMPLify [Pavlakos *et al*., 2019], Kimodo [Rempe *et al*., 2026], SMPLitex [Casas and Comino-Trinidad, 2023], Rokoko Add-on-based motion retargeting (version 1.4.4 beta), Blender (version 5.0.1), and Unreal Engine (version 5.5.4). SMPLify [Pavlakos *et al*., 2019] provides the pipeline's modeling and rigging basis by producing a SMPL-X parametric human body with a stable topology and an articulated skeletal structure. This representation is suitable for downstream animation because it preserves editability, supports body-shape variation, and is compatible with conventional digital content creation workflows via the SMPL-X Blender Add-on. Kimodo [Rempe *et al*., 2026] contributes generated motion, SMPLitex [Casas and Comino-Trinidad, 2023] provides SMPL-compatible texture maps, Rokoko mediates motion retargeting, Blender serves as the asset integration environment, and Unreal Engine supports scene-level deployment and validation. Figure **??** summarizes the resulting workflow.

At the implementation level, the asset flow followed a sequential body/appearance/motion/deployment process. SMPL-X bodies were created and edited in Blender through the SMPL-X Blender Add-on. SMPLitex texture candidates were applied to the SMPL-X material setup and visually inspected on the 3D body before selection. Kimodo generated motion as BVH files, which were imported into Blender and retargeted to the SMPL-X body using Rokoko. Because the Kimodo/SOMA skeleton and the SMPL-X skeleton use different rest-pose conventions and bone-naming structures, retargeting required manual pose alignment, armature relinking, source/target bone matching, and correction of mismatched limb mappings. Final animated assets were exported to Unreal Engine in FBX format, as previous GLB export attempts resulted in corrupted imports.

Kimodo [Rempe *et al*., 2026] is used to cover the animation stage. Its role in the pipeline is to generate or compose human motion using parametric multi-model body logic, allowing the pipeline to move beyond static avatar creation toward temporally defined 4D characters. However, Kimodo [Rempe *et al*., 2026] and SMPLify [Pavlakos *et al*., 2019] are not directly interchangeable as production components. Their animation outputs and expected input formats differ, which prevents a simple import-and-apply workflow. For this reason, SPL4SH introduces Rokoko's Blender AddOn as an intermediate retargeting layer that transfers Kimodo [Rempe *et al*., 2026] generated motion to the SMPL-X character inside Blender. This retargeting stage is not merely an implementation detail; it is one of the central integration mechanisms that allows

Table 3. Qualitative suitability assessment of selected synthetic 4D human generation techniques across SPL4SH pipeline criteria. Symbols indicate whether a technique does not substantially address a criterion (×), partially or indirectly addresses it (○), or shows strong suitability for that criterion within the selected set (⋆).

Stage	Criterion	SMPLify	PromptHMR	SOMA	Kimodo	GSGD	Neural FVP	AKD	ICo3D
Modeling	Model reconstruction	⋆	○	×	×	×	⋆	×	⋆
	Model quality	⋆	○	⋆	○	×	⋆	×	⋆
	Editability	⋆	⋆	⋆	○	×	○	×	○
	Layered modeling	×	○	×	×	×	⋆	×	⋆
	Clothing physics	×	×	×	×	×	⋆	×	⋆
	Performance in XR	○	⋆	○	○	×	⋆	×	⋆
Rigging	Deformation quality	⋆	⋆	⋆	⋆	×	×	○	×
	Rigging compatibility	○	○	⋆	⋆	×	×	⋆	×
	Expressive rigging	⋆	○	○	×	×	×	×	×
Animation	Animation length	×	⋆	○	⋆	○	⋆	○	○
	Animation constraints	×	×	×	⋆	×	○	×	○
	Motion reconstruction	×	⋆	×	×	○	⋆	×	○
	Motion complexity	×	⋆	○	⋆	○	○	○	○
	Motion plausibility	×	○	○	⋆	⋆	⋆	○	○
	Motion physics	×	○	○	⋆	○	⋆	○	○
	Animation retargeting	×	○	○	⋆	⋆	×	⋆	×
Rendering	Photorealism	×	×	×	×	○	⋆	○	⋆
	Material rendering	×	×	×	×	○	○	○	⋆
	Rendering modularity	×	×	×	×	×	⋆	×	×
	Rendering reconstruction	×	×	×	×	○	⋆	×	⋆
Interaction	Contact plausibility	×	○	×	×	×	⋆	×	×
	Context affordance	×	○	×	×	×	×	×	⋆
	Context modularity	×	⋆	×	○	×	○	×	○
	Context representation	×	⋆	×	×	×	○	×	○
Full pipeline	End-to-end representation	×	×	×	×	×	○	×	⋆
	Scalability	×	⋆	×	○	×	○	×	○
	3D engine compatibility	×	⋆	×	○	×	○	×	○
	XR applicability	×	⋆	×	○	×	○	×	⋆

Technique headers: SMPLify = Expressive Body Capture; GSGD = Gaussian See, Gaussian Do; Neural FVP = Neural Free-Viewpoint Performance Rendering; AKD = Articulated Kinematics Distillation; ICo3D = An Interactive Conversational 3D Virtual Human.

the independently generated body and motion components to function as a single animated asset.

SMPLitex [Casas and Comino-Trinidad, 2023] is incorporated to address the appearance gap left by the body and motion components. Although SMPLitex [Casas and Comino-Trinidad, 2023] was not selected among the highest-ranked techniques in the comparative suitability assessment, it is included because the implemented pipeline requires a lightweight texture-generation component compatible with SMPL-style UV maps. In this role, SMPLitex [Casas and Comino-Trinidad, 2023] contributes to the rendering stage by generating texture maps for the SMPL-X body, enabling the parametric mannequin to acquire skin, clothing, and identity-related visual cues. Its inclusion, therefore, reflects a pipeline-completion requirement rather than a claim that it is the strongest rendering technique overall.

The final pipeline is assembled in Blender and deployed in Unreal Engine. Blender serves as the integration environment in which the SMPL-X body, generated texture maps, and retargeted Kimodo [Rempe *et al.*, 2026] motion via Rokoko are combined and inspected. Unreal Engine is then used to evaluate whether the resulting animated human can be placed inside an XR-oriented scene and coordinated with scene-level authoring tools such as the Level Sequencer. In this deployment stage, interaction is treated as scene-situated behavior rather than as fully autonomous physical interaction. The pipeline supports authored HOI, HHI, and HSI arrangements through placement, motion sequencing, and transform adjustment, but these interactions still require manual validation and refinement.

In summary, SPL4SH operationalizes the stage taxonomy by connecting body modeling, rigging, texture-based appearance, generated motion, retargeting, and engine deployment into a single modular workflow. This design allows the pipeline to produce animated synthetic humans for XR prototyping while also exposing the practical barriers that remain

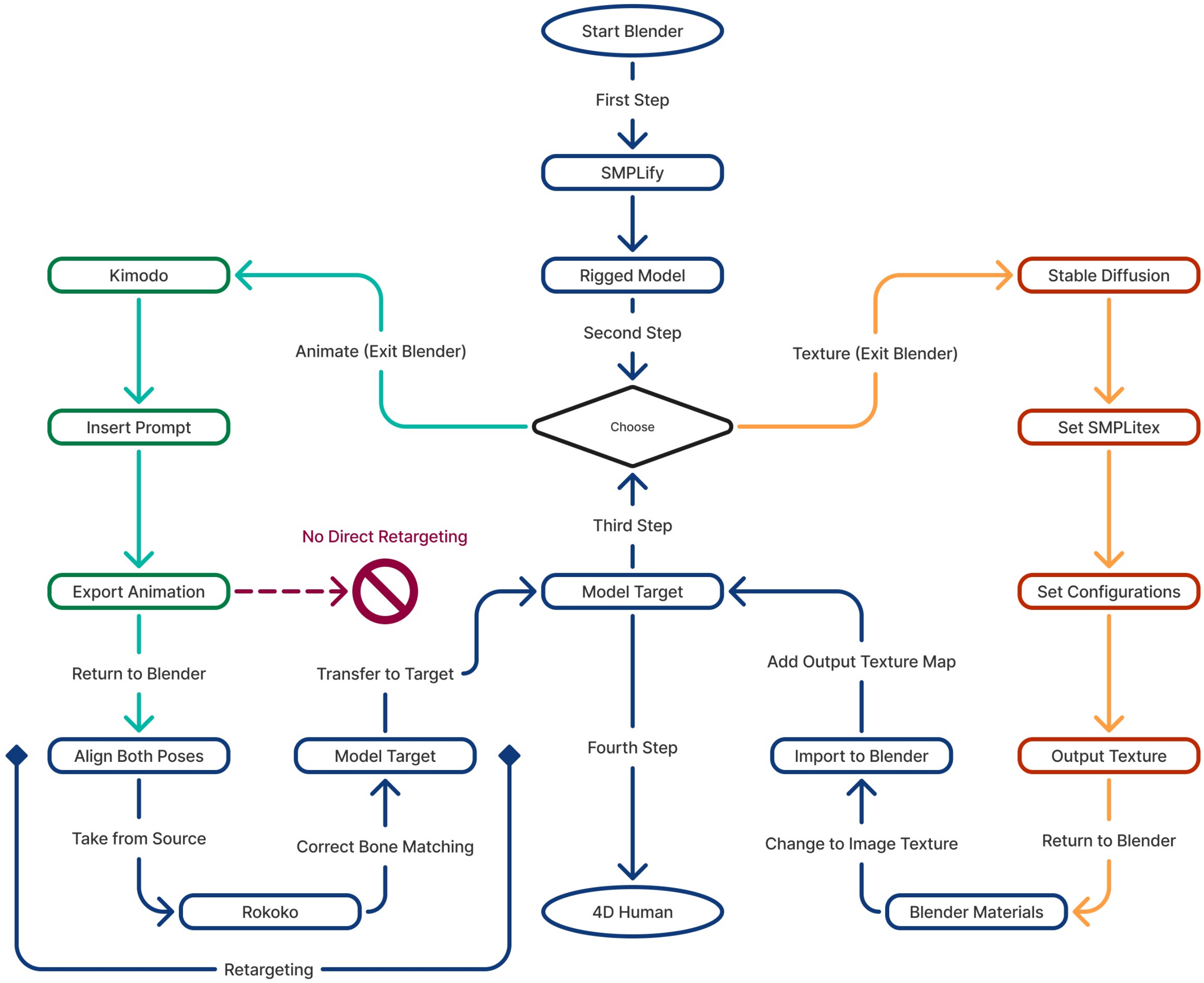


Figure 1. SPL4SH proof-of-concept workflow. The implemented configuration combines SMPLify/SMPL-X for body modeling and rigging, Kimodo [Rempe *et al.*, 2026] for motion generation, SMPLitex [Casas and Comino-Trinidad, 2023] for texture-map generation, Rokoko for motion retargeting, Blender for asset integration, and Unreal Engine for scene-level deployment. The diagram highlights the main integration steps required to transform independent body, motion, and appearance outputs into an authored XR-oriented scene asset.

between current generative techniques, including format mismatch, skeleton alignment, texture coherence, motion artifacts, and scene-level interaction validation.

4 Results and Discussion

4.1 Results of the proposed pipeline

Because SPL4SH is evaluated as a pipeline-oriented proof of concept rather than as a new reconstruction, animation, or rendering algorithm, we report practical production indicators instead of formal perceptual or runtime benchmarks. These indicators include generation time, curation effort, number of selected assets, retargeting effort, integration time, and observed manual interventions. They are intended to characterize the practical cost of this SPL4SH instantiation and should not be interpreted as hardware-independent performance benchmarks.

This section, then, reports the behavior of the SPL4SH pipeline at three levels: individual component behavior, complete pipeline integration, and scene-level deployment. This separation is important because each component contributes a different layer to the construction of a synthetic 4D human. SMPLify [Pavlakos *et al.*, 2019] provides the parametric body structure and deformation basis, Kimodo [Rempe *et al.*, 2026] contributes motion synthesis, and SMPLitex [Casas and Comino-Trinidad, 2023] provides SMPL-compatible texture maps for appearance generation, supporting both text-prompted texture synthesis and image-conditioned texture estimation. These components do not fail in the same way: structural consistency, motion expressiveness, texture coherence, retargeting quality, and scene deployment expose different limitations.

The first level of analysis examines each technique separately, identifying intrinsic properties of each component, such as mesh stability, motion diversity, texture compatibility, artifact frequency, and curation burden. The second level examines the complete pipeline, focusing on how body geometry, rigging, texture maps, generated motion, and retargeting

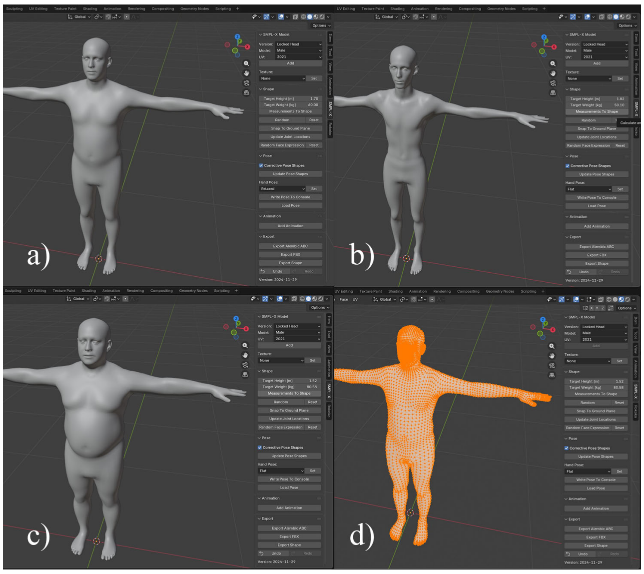


Figure 2. SMPLify's [Pavlakos *et al.*, 2019] Blender AddOn can modify body shape to meet the user's needs and generate diverse humans through variables at any time (a, b, and c). However, despite shape variations, the SMPL-X model still keeps its vertex and polygon count (d).

combine to form a single animated asset. The third level examines scene integration, evaluating whether the resulting animated humans can be placed and authored inside an Unreal Engine scene. This organization allows us to distinguish between limitations that appear in isolated outputs and limitations that emerge only when the components are combined into an XR-oriented workflow.

4.1.1 Per-technique individual results comparison

The individual tests show that each technique contributes a distinct asset layer to the SPL4SH pipeline. SMPLify [Pavlakos *et al.*, 2019] provides the geometric and skeletal basis of the character, enabling body-shape variation while preserving a stable mesh structure across deformations. Kimodo [Rempe *et al.*, 2026] contributes motion synthesis, including simple actions, more complex movements, interpolated action sequences, and constraint-guided motions that can approximate interaction-oriented behaviors, such as pulling a chair, opening a door, or performing gestures. SMPLitex [Casas and Comino-Trinidad, 2023] contributes SMPL-compatible texture maps, either from text prompts or image-conditioned estimation, allowing the mannequin-like SMPL-X body to acquire face, skin, and clothing appearance cues.

The results reveal a clear separation between structural reliability, motion expressiveness, and visual stability. SMPLify [Pavlakos *et al.*, 2019] is the most stable component at the structural level, used to automatically generate a SMPL-X human model with male, female, or neutral variations in Blender's engine. As shown in Figure 2, it supports body-shape variation while preserving a consistent topology and a UV layout compatible with SMPL-style texture maps. However, its visual realism remains limited by the absence of explicit geometry for hair, garments, shoes, and accessories. These elements can be suggested through texture, but they do not acquire independent volume, silhouette, collision behavior, or secondary motion.

Kimodo [Rempe *et al.*, 2026] is the strongest component

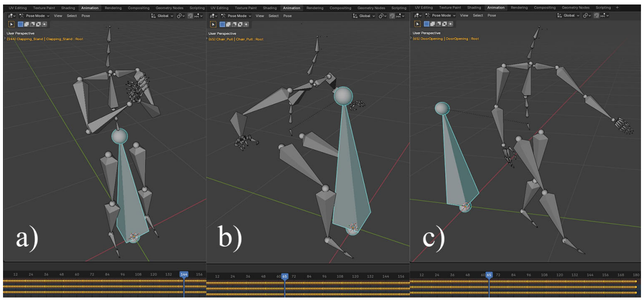


Figure 3. Kimodo [Rempe *et al.*, 2026] outputs .bvh format motion capture files, preserving motion and bones without deformation weights. Kimodo [Rempe *et al.*, 2026] can present some artifacts at hand collisions, where hands interpenetrate, which can occur with some limb-colliding tasks (a); however, it reaches good results at portraying complex animation that does not need any limb intercontact, like sitting on a chair (b) or opening a door while walking (c).

for motion generation. It can produce a broad range of actions and supports prompt- or constraint-guided variation to generate .bvh motion capture files from interpolated motion sequences. Figure 3 presents some frames from the output animations, and using the door opening while walking sequence, for example, those motion sequences are generated with a text prompt like "*A person opens a doorknob while walking*", and configurations such as seed 42 (default seed) and 3 keyframes pose constraints to make the arm open the door to its parallel side. However, its outputs still require inspection before integration. In the tested motions, repeated-contact actions, such as clapping or crossing legs, could produce interpenetration artifacts. In addition, the lack of reusable in-place locomotion clips limits compatibility with common real-time character workflows, where the motion cycle and the character trajectory are often controlled separately. For motion generation, we executed Kimodo [Rempe *et al.*, 2026] on an NVIDIA GeForce RTX 4090 GPU with 24 GB of VRAM. Generated motions used the default seed 42, prompts following the structure "a person [performs an activity]", and animation lengths between 6 and 10 seconds. The average generation time was approximately 15–20 seconds per motion. Most generated motions did not require constraint editing, but two interaction-oriented cases required manual constraint adjustment: opening a door while walking and pulling a chair before sitting. Each of these cases required approximately 10 minutes of iterative constraint editing, regeneration, and visual inspection.

SMPLitex [Casas and Comino-Trinidad, 2023] contributes the appearance layer of the implemented SPL4SH configuration. Figure 4 shows examples of SMPLitex texture maps applied to SMPL-X characters in Blender. Although the method can generate recognizable cues for face, skin, clothing, and overall character appearance, it was also the component with the highest curation burden in our implementation.

In total, 2,089 texture candidates were generated using Stable Diffusion Forge Neo on an NVIDIA RTX 5060 Ti GPU with 16 GB of VRAM. The generation settings were: Euler A sampling, Karras schedule, 32 sampling steps, 512×512 resolution, CFG scale 2.5, and seed -1 to produce prompt-based variations. Texture generation required approximately 3–5 minutes per batch of 50 images. Because SMPLitex relies on a constrained prompt structure, such as "a sks texturemap of ...", small prompt variations were useful for obtaining dif-

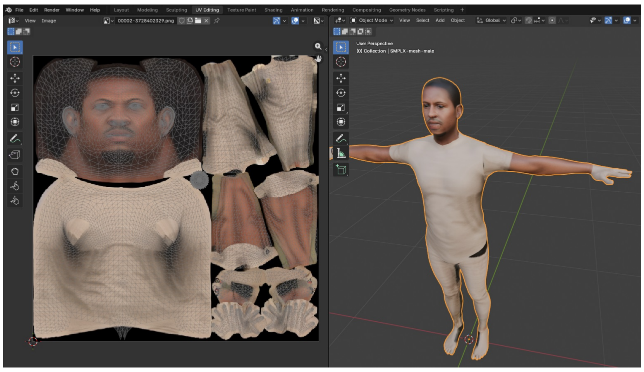


Figure 4. A representation of a SMPLitex [Casas and Comino-Trinidad, 2023] texture map in .png format, generated with its Stable Diffusion checkpoint, which is applied in Blender at the unwrapped model's faces (left) and at the SMPL-X model character (right). Despite almost perfectly matching the SMPL texture map layout, some artifacts persist, including texture discontinuities, poor fitting, and confusing lighting and shading.

ferent clothing or character appearances, but overly detailed prompts tended to increase hallucinated or structurally invalid outputs.

The generated candidates were filtered in two stages. First, each image was visually screened for approximate SMPL texture-map plausibility. At this stage, retention indicated complete body coverage, recognizable body-part placement, and absence of critical artifacts that would immediately prevent testing, such as duplicated faces, misplaced limbs, portrait-like outputs, missing body regions, or severe texture discontinuities. This first screening retained 65 of the 2,089 generated candidates for Blender-based testing and required approximately 40 minutes. Second, the retained candidates were applied to SMPL-X bodies in Blender and inspected from multiple viewpoints to evaluate whether the texture remained plausible once wrapped onto the 3D body. This Blender-based inspection took approximately 30 minutes and yielded 5 textures for use in the final pipeline.

Figure 5 illustrates this two-stage curation process. The raw generated candidates in Figure 5a include examples rejected during the first screening, such as abstract non-texture outputs, misaligned or poorly scaled texture layouts, hallucinated facial regions in torso areas, diffuse segmentation boundaries, monochromatic skin-clothing collapse, and portrait-like images instead of complete unwrapped texture maps. The candidates in Figure 5b passed the first visual screening but were not selected after Blender inspection because wrapping them onto the SMPL-X body revealed scale mismatches, holes, unclear hand and foot regions, skin-clothing blending artifacts, and discontinuities between garment regions. The final textures in Figure 5c were selected because they fitted the SMPL-X body with only minor artifacts, such as small lighting inconsistencies, mild color blending between texture regions, or small placement mismatches on the unwrapped faces.

Overall, the SMPLitex [Casas and Comino-Trinidad, 2023] stage had a first-screening retention rate of approximately 3.1% and a final selection rate of approximately 0.24% relative to all generated candidates. This indicates that, in the implemented SPL4SH configuration, texture generation was not limited by raw generation time alone, but by the manual effort required to identify texture maps that remained usable after 3D application.

Taken together, the isolated results indicate that none of the components is sufficient for high-quality XR deployment on its own. SMPLify [Pavlakos *et al.*, 2019] provides a reliable parametric body and rig, but not a complete visual appearance. Kimodo [Rempe *et al.*, 2026] provides useful motion, but not guaranteed contact correctness or engine-ready locomotion structure. SMPLitex [Casas and Comino-Trinidad, 2023] provides fast appearance generation, but requires visual curation and, in some cases, manual editing or inpainting. These findings justify the need to evaluate SPL4SH not only as a set of independent tools but as an integrated pipeline in which body, texture, motion, retargeting, and scene deployment interact.

4.1.2 Complete pipeline results

The complete SPL4SH pipeline combines the SMPL-X body and rig produced by SMPLify [Pavlakos *et al.*, 2019], the SMPL-compatible texture maps generated by SMPLitex [Casas and Comino-Trinidad, 2023], and the motion sequences generated by Kimodo [Rempe *et al.*, 2026]. In the integrated workflow, Blender serves as the assembly environment where the body, texture, and animation layers are combined, inspected, and corrected before deployment. The resulting character can therefore be treated as a complete animated synthetic human asset: SMPLify [Pavlakos *et al.*, 2019] provides the stable parametric geometry and deformation structure, SMPLitex [Casas and Comino-Trinidad, 2023] adds face, skin, and clothing appearance cues, and Kimodo [Rempe *et al.*, 2026] contributes the temporal motion layer.

The main integration gap appears at the animation-transfer stage. Although Kimodo [Rempe *et al.*, 2026] and SMPL-X are both parametric human-body representations, their outputs are not directly compatible within the implemented workflow. At first, we try to replace the model's skeleton with the .bvh motion capture's bone structure, but since it does not preserve the mesh's pose deformations, as Kimodo [Rempe *et al.*, 2026] outputs lack LBS deformation weights, Rokoko was therefore introduced as an intermediate retargeting tool for transferring Kimodo [Rempe *et al.*, 2026] generated motion to the SMPL-X character in Blender. This retargeting process is presented in Figure 6 and requires manual alignment, bone-structure matching, and correction, particularly because early tests produced errors due to pose and joint mismatches. Examples include differences between the SOMA shoulder joints and the SMPL-X collarbone structure, as well as mismatches in the starting pose and bone placement. More accurate retargeting was achieved when the source and target skeletons were manually aligned before transfer. Retargeting required approximately six manual preparation steps per animation: baking the SMPL-X shape keys, aligning the static SMPL-X body to the first frame of the Kimodo BVH animation, applying the pose as the rest pose, relinking the armature, matching source and target bone structures in Rokoko, and correcting mismatched limb mappings before retargeting. The Rokoko retargeting operation itself required approximately 10 seconds after setup, but the full preparation process required approximately 4-6 minutes per animation depending on pose complexity, as more challenging poses need more joints editing. The final assets were exported using the FBX format because previous GLB export attempts resulted in corrupted

(a) Raw generated candidates (sampled from 2,089)

(b) Passed first visual screening (sampled from 65)

(c) Final selected textures used in pipeline (sampled from 5)

Figure 5. SMPLitex texture-generation curation process in the SPL4SH pipeline. From 2,089 generated texture candidates, 65 passed the first visual screening and were retained for Blender-based testing, and 5 were selected for use in the final pipeline after inspection on the SMPL-X body. (a) Raw-generated candidates rejected during the first screening, including abstract non-texture outputs, misaligned layouts, hallucinated body regions, monochromatic skin/clothing collapse, and portrait-like images. (b) Candidates that passed first visual screening but revealed problems after Blender inspection, including scale mismatches, holes, unclear hand or foot regions, skin/clothing blending, and garment discontinuities. (c) Final selected textures used in the pipeline, which still contain minor artifacts but remained usable when applied to the SMPL-X body.

imports in Unreal Engine.

Complex spatial actions remained more difficult to integrate. Sitting-like motions, for example, required additional sequencing or correction because some joint-position transformations were poorly transferred or unsupported in the retargeting process. This indicates that the pipeline is more reliable for general body motion and gesture transfer than for tightly constrained interactions involving furniture, contact, or precise scene alignment. Texture integration also introduced a second manual correction stage. When SMPLitex [Casas and Comino-Trinidad, 2023] texture maps were applied to the animated SMPL-X body, some discontinuities became more visible under deformation, especially around facial regions, seams, limbs, and clothing boundaries. In these cases, image-editing tools or Stable Diffusion img2img inpainting were required to correct mismatches, colors, and garment details before the asset became visually acceptable.

Despite these limitations, the complete pipeline works as a modular proof of concept for generating 4D humans through interchangeable body, texture, and motion components. The combined techniques enable the transformation of a parametric SMPL-X mannequin into a textured and animated human character. SMPLitex [Casas and Comino-Trinidad, 2023] improves the visual completeness of the body by adding skin tone, facial appearance, and simple clothing cues, while the SMPL-X rig provides believable linear-blend-skinning deformation when driven by retargeted Kimodo [Rempe *et al.*, 2026] animations. This modularity is valuable for XR authoring because different body shapes, texture maps, and motion clips can be recombined to produce multiple character variants.

The complete pipeline results also clarify the current limits of this workflow. Kimodo [Rempe *et al.*, 2026], SMPLify [Pavlakos *et al.*, 2019], and SMPLitex [Casas and Comino-Trinidad, 2023] were not originally designed as a unified system, so the pipeline depends on format mediation, retargeting, manual pose matching, texture inspection, and scene-level

Figure 6. Rokoko animation retargeting from the animation source (left skeleton) to the target that should be animated (right human model). Tests with matching starting poses and correct bone locations resulted in more accurate animation retargeting. To achieve better results, the starting pose must be the same for both source and target skeletons.

correction. Kimodo [Rempe *et al.*, 2026] transitions between activity sequences may remain stiff, subtle jitter becomes more visible once motion is applied to a textured human body, and SMPLitex [Casas and Comino-Trinidad, 2023] facial or seam artifacts can become more uncanny when stretched by animation. These issues do not prevent SPL4SH from functioning as an XR-oriented prototype pipeline, but they show that current modular 4D human generation still requires human supervision before producing polished, deployable assets.

4.1.3 Pipeline scene integration

The scene-integration tests evaluated whether the animated humans produced by SPL4SH could be imported, placed, and authored inside a real-time XR-oriented environment. The final assets were integrated into Unreal Engine 5.5.4 using the Level Sequencer and placed in a meeting-room scene containing HOI, HHI, and HSI arrangements, as shown in

Figure 7. These tests were not intended to validate fully autonomous interaction or physical simulation. Instead, they evaluated whether the generated humans could be situated inside a spatial scene and coordinated through engine-level authoring tools. The Unreal Engine scene was authored in Level Sequencer by adjusting the character's position, rotation, and timing to make the generated motions plausible within the pre-existing environment. Also, HOI cases (door opening and chair pulling) required synchronizing human and objects animations, positions and rotations to perform proper interaction timing. The scene-level animation setup took approximately 1 hour, excluding the placement of static scene objects. This effort indicates that the implemented pipeline supports authored, scene-situated interactions but still relies on manual timing and spatial adjustment rather than on autonomous scene understanding or physically grounded interaction synthesis.

The results show that SPL4SH outputs can be deployed as animated characters in an Unreal Engine scene, making the pipeline suitable for XR prototyping, virtual production, synthetic scene construction, game-level blocking, and early-stage spatial-intelligence dataset creation. End users can adjust transforms, timing, and placement to adapt the generated characters to the target scene. This indicates that the pipeline can cross the boundary between isolated asset generation and scene-level XR composition. The scene integration results are also shown in the supplementary video, which presents the animated SPL4SH characters deployed in the Unreal Engine meeting-room environment.

However, scene integration also exposes limitations that are less visible during isolated component testing. The final characters still require scene-level authoring and visual inspection to ensure plausible placement, timing, contact, and orientation. In the meeting-room tests, some results remained rough or uncanny because motion artifacts, texture discontinuities, and interaction misalignments became more apparent once the characters were placed in a spatial context. As a result, SPL4SH does not eliminate the need for animation polishing, interaction alignment, texture correction, or manual validation before use in high-quality XR simulations or spatial-intelligence datasets.

Overall, the scene-integration results support the use of SPL4SH as a modular prototyping workflow for spatially situated synthetic humans. The pipeline can produce animated human assets that are deployable in a real-time engine, but the current implementation should be understood as a supervised authoring workflow rather than a fully automatic production pipeline.

4.2 Discussion

This subsection interprets the SPL4SH results beyond the descriptive presentation of individual modules, pipeline integration, and scene deployment. The discussion focuses on three aspects: the reliability of the modular pipeline, the relevance of synthetic 4D humans for XR-oriented content creation, and the persistent barriers that still limit broader adoption in immersive systems. This distinction is important because SPL4SH is not proposed as a monolithic generative model but as a workflow that connects parametric body modeling, texture map generation, motion synthesis, retargeting, Blender-based integration, and Unreal Engine deployment. The discussion, therefore, examines not only whether animated synthetic humans can be produced, but also how component-level strengths persist or degrade when independent tools are combined into a single XR-oriented asset pipeline [Pavlakos *et al.*, 2019; Rempe *et al.*, 2026; Casas and Comino-Trinidad, 2023].

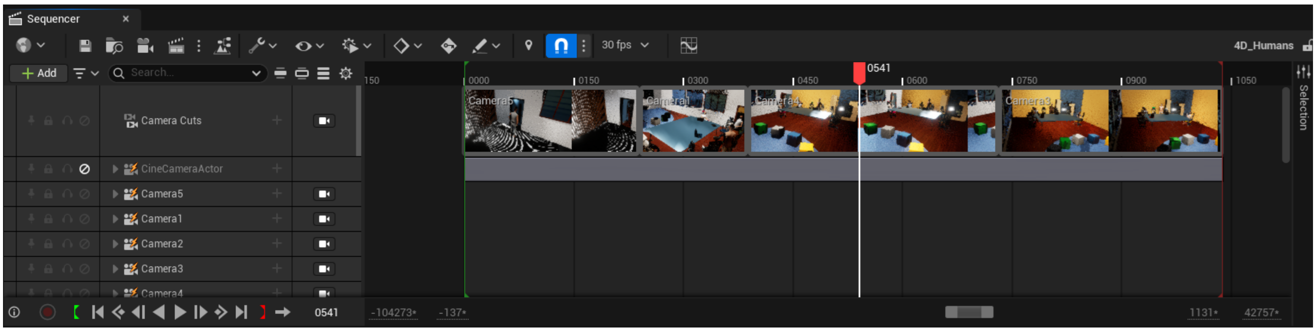
The guiding questions structure the interpretation of the SPL4SH results. Q1, Q2, and Q3 focus on pipeline reliability: which production stages are required, how independent techniques can be aligned into a modular workflow, and which parts still depend on manual intervention, task-specific supervision, or representation-specific optimization. Q4 and Q5 shift the analysis toward XR deployment and application potential, including scene-level integration and use cases such as games, simulators, virtual tours, digital content creation, telepresence, scientific research, and synthetic dataset generation. Q6 addresses the technical and operational barriers that still prevent synthetic 4D humans from becoming broadly reliable in immersive XR systems. Together, these questions connect the SPL4SH implementation to broader issues of representation, controllability, interaction grounding, deployment feasibility, and usage in spatial intelligence applications.

4.2.1 Pipeline strengths

The main strength of SPL4SH is its demonstrated feasibility as a modular path to producing synthetic 4D human characters by combining body modeling, rigging, texture-based appearance generation, motion synthesis, retargeting, and real-time engine deployment. The results show that a functional animated human asset can be assembled without relying on a single monolithic end-to-end model. Instead, the pipeline operates through interchangeable body, texture, and motion components that can be combined and inspected at different stages of the workflow.

In this arrangement, SMPLify [Pavlakos *et al.*, 2019], used through the SMPL-X Blender Add-on, provides a stable parametric body representation with controllable body-shape variation, consistent topology, and a skeletal structure suitable for deformation. SMPLitex [Casas and Comino-Trinidad, 2023] complements this mannequin-like body by generating SMPL-compatible texture maps that introduce visual cues such as facial details, skin tone, and simple clothing appearance. Kimodo [Rempe *et al.*, 2026] introduces a temporal layer that generates motions ranging from simple actions to more complex interpolated sequences, including constraint-guided motions that can approximate human-object, human-scene, and human-human arrangements.

Once retargeted through Rokoko, Kimodo [Rempe *et al.*, 2026] motions can drive the SMPL-X mesh with plausible linear-blend-skinning deformations, allowing the character to function as an animated asset rather than a static, textured body. This is important because it shows that independently generated body, appearance, and motion components can be operationalized within a shared production workflow. The pipeline also preserves a useful degree of authoring control: body shapes, texture maps, and motion clips can be recom-



(a) Manual authoring and camera sequencing in Unreal Engine Level Sequencer.

(b) Scene-situated synthetic humans in a meeting-room arrangement.

(c) Wider rendered view of the authored multi-character scene.

Figure 7. Unreal Engine integration of the implemented SPL4SH configuration. The scene was authored by manually adjusting placement, timing, camera cuts, and motion sequencing. The rendered views show textured, animated synthetic humans situated in an XR-oriented meeting room environment. This implementation demonstrates authored scene-level human–object, human–human, and human–scene arrangements, not autonomous, adaptive, or physically grounded interaction synthesis.

bined to create different character variants, while placement, timing, and transforms can still be adjusted inside Blender or Unreal Engine.

This modularity gives SPL4SH practical value for early-stage XR prototyping. It does not remove the need for expert inspection, correction, or polishing, but it provides a reusable workflow for assembling spatially situated synthetic humans from existing AI-assisted tools. The contribution is therefore not only the final animated character, but the demonstration that body representation, generated appearance, generated motion, retargeting, and scene deployment can be connected into a coherent pipeline while keeping the remaining integration failures visible for analysis.

4.2.2 Reliability of modular technique composition

The reliability of a 4D human pipeline cannot be assessed only by whether it produces a visually plausible body. For XR applications, the generated asset must remain usable across modeling, rigging, animation, rendering, and interaction stages. The comparative assessment shows that no single technique dominates all of these requirements. SMPLify [Pavlakos *et al.*, 2019] is reliable as a parametric modeling and rigging basis, SOMA [Saito *et al.*, 2026] is relevant for cross-model compatibility, Kimodo [Rempe *et al.*, 2026] is stronger for controllable motion generation and retargeting, Neural Free-Viewpoint [Sun *et al.*, 2021] and ICo3D [Shaw *et al.*, 2026] are stronger for visually rich or interactive virtual-human representations, and PromptHMR [Wang *et al.*, 2025] contributes to populated and multi-person XR scenarios. This directly addresses Q1: a functional XR-oriented 4D human is not obtained by selecting a single best method, but by arranging complementary techniques according to their stage-specific roles.

SPL4SH answers Q2 by demonstrating how these stages can be aligned into a modular AI-assisted workflow. In the implemented pipeline, SMPLify [Pavlakos *et al.*, 2019] provides the SMPL-X body and rigging foundation; Kimodo [Rempe *et al.*, 2026] contributes generated or composed motion; SMPLitex [Casas and Comino-Trinidad, 2023] addresses the appearance layer through SMPL-compatible texture maps; Rokoko mediates motion retargeting; Blender supports asset integration; and Unreal Engine enables scene-level deployment. This sequence shows that AI-driven 4D human pipelines can be assembled from specialized components, but only when conversion, retargeting, inspection, and engine-deployment stages are treated as central parts of the workflow rather than as secondary implementation details.

Q3 requires a distinction between modular feasibility and operational robustness. SPL4SH is modular in the sense that body shapes, texture maps, and motion clips can be recombined to create different character variants, which is valuable for XR authoring, synthetic scene construction, and early spatial-intelligence experiments. As a proof of concept, it demonstrates that independently generated body, motion, and appearance components can be assembled into animated humans and placed inside Unreal Engine scenes. However, the pipeline remains only partially scalable for generating many fully reliable XR characters without human validation. Kimodo [Rempe *et al.*, 2026] and SMPLify [Pavlakos *et al.*, 2019] are not natively interchangeable, so motion transfer depends on Rokoko-based retargeting and manual pose or skeleton alignment. Complex actions, such as sitting-like motions or scene-constrained postures, may require additional sequencing or correction. SMPLitex [Casas and Comino-Trinidad, 2023] also introduces a strong curation burden because generated textures may contain discontinuities, hallucinated body regions, mismatched seams, confusing clothing folds, or color inconsistencies that become more visible under deformation.

These findings show that AI-assisted 4D human pipelines are technically plausible, but not automatically robust. The required components already exist across body modeling, rigging, animation, appearance generation, retargeting, and deployment; however, practical success is not achieved by simply connecting state-of-the-art modules in sequence. It depends on modular design choices, shared representations, format compatibility, motion-transfer reliability, texture coherence, and the target XR use case. A pipeline designed for editable-avatar production may prioritize SMPL-X topology, retargetable skeletons, and texture-map compatibility, whereas a pipeline designed for photorealistic free-viewpoint playback may rely more heavily on neural or Gaussian rendering. Similarly, a system intended for spatial-intelligence datasets must prioritize contact plausibility, scene alignment, temporal coherence, and controllable variation rather than visual quality alone. SPL4SH therefore shows that target-specific constraints must guide the selection and integration of every pipeline component.

4.2.3 4D Humans Generation Importance to XR's Virtual Content

The integration of 4D human generation pipelines into XR applications is feasible, but the current level of feasibility is better understood as deployable prototyping than as automatic production. Unreal Engine 5.5.4 scene-integration tests show that SPL4SH outputs can be placed in a meeting room and arranged at HOI, HHI, and HSI interaction levels. This is important for Q4 because it demonstrates that generated humans can bridge the boundary between isolated asset creation and engine-level XR scene composition. At the same time, the tests make it clear that deployment does not eliminate constraints on rendering quality, latency-sensitive interaction, or scene authoring. The resulting characters remain visually rough in some cases: texture artifacts may become uncanny during animation; subtle jitter and stiff transitions become more visible when motion is applied to a textured body; and interaction alignment still requires transform adjustments and visual inspection. Therefore, the pipeline is feasible for rapid XR prototyping, spatial layout studies, synthetic scene blocking, and early dataset generation, but high-quality XR deployment still requires animation polishing, texture correction, contact validation, and scene-level refinement before the assets can support demanding immersive applications.

Beyond spatial intelligence, synthetic 4D humans enable several promising XR and digital content domains, although each domain inherits distinct limitations from the pipeline. For Q5, the strongest application areas identified include games, simulators, virtual tours, digital content creation, telepresence-oriented virtual humans, scientific research, and synthetic dataset generation. In games, SPL4SH-like workflows can support rapid non-player-character prototyping, body variation, motion blocking, and background-character generation, although final hero characters would still require higher-quality materials and textures, properly layered object geometry, animation cleanup, and rig validation. In simulators, generated humans can populate training environments and represent gestures or object-oriented actions, but the usefulness of the output depends on reliable contact, correct posture, and avoidance of interpenetration. Virtual tours can benefit from human figures that convey scale, social use, and spatial atmosphere, while digital content creation workflows can leverage the pipeline as a generative scaffold for previsualization and design iteration. For synthetic datasets, the value lies in controllable variation across body shape, texture, motion, and scene placement; however, this promise is limited by manual retargeting, visual curation, and the need to verify that generated artifacts do not corrupt the spatial or semantic signals expected by perception models.

The results further suggest that AI-created 4D humans should not be reduced to avatar visualization. Their relevance to XR emerges when they function as situated, temporal, and interactive entities inside spatial environments. SPL4SH shows that generated humans can be used to prototype embodied agents, populate scenes, test human-object and human-scene arrangements, support motion-based content creation, and contribute to synthetic data workflows for spatial intelligence.

Their role also extends to scalable digital production, because designers can permute body parameters, texture maps, and generated motions to create multiple character variants more rapidly than in fully manual workflows. However, this potential must be interpreted in light of the current technical limitations. The pipeline does not yet provide autonomous, physically grounded, or fully scene-aware humans. Instead, it offers a modular path for producing animated human assets that can be authored, inspected, and refined within XR scenes. The broader value of 4D humans, therefore, depends on improving interaction grounding, temporal stability, visual coherence, and engine-level controllability.

4.2.4 Persistent challenges

The remaining barriers to the broad adoption of realistic 4D humans in immersive XR systems span representation, motion, appearance, integration, and validation. In response to Q6, the SPL4SH results show that the main bottleneck is not in a single pipeline stage but in preserving coherence across stages. Parametric bodies such as SMPL-X are structurally reliable but visually incomplete: hair, clothing, shoes, and accessories are mostly represented by textures rather than by independent geometry, material thickness, collision, or secondary motion. Kimodo [Rempe *et al.*, 2026] provides useful motion synthesis, but repeated-contact movements can exhibit interpenetration, transitions between activities may remain stiff, and spatially constrained actions may still require manual correction. SMPLitex [Casas and Comino-Trinidad, 2023] accelerates appearance generation, but its texture maps may contain discontinuities, duplicated or hallucinated body regions, misplaced hands or feet, unclear clothing boundaries, and color inconsistencies that become more visible during deformation. These appearance-related failures are especially important because prior work on virtual-human rendering has shown that skin-tone representation and rendering biases can affect the quality and reliability of virtual human outputs [Menezes *et al.*, 2025]. At the integration level, Kimodo [Rempe *et al.*, 2026] and SMPLify [Pavlakos *et al.*, 2019] do not form a native unified system: motion transfer depends on Rokoko, skeleton and pose alignment require human intervention, and Unreal Engine deployment still requires scene-level adjustment through the Level Sequencer. These issues indicate that XR-ready 4D human generation requires not only stronger individual modules but also more reliable interoperability among body structure, rigging, texture, motion, interaction, and real-time deployment.

The results also show that human-specific artifacts are especially disruptive in immersive environments because the body is a perceptually demanding object. Small errors in static props or background geometry may be tolerated, but unnatural posture, unstable motion, incorrect contact, uncanny facial or skin appearance, and mismatches between body deformation and clothing cues can immediately weaken the plausibility of human presence. This concern is consistent with studies on the uncanny valley effect in virtual humans, including work showing that skin appearance and demographic representation can influence how virtual characters are perceived [de Andrade Araujo *et al.*, 2023]. In SPL4SH, artifacts such as hand interpenetration, leg collision, stiff transitions, jitter, texture seams, duplicated facial or limb regions, and poor sitting alignment are therefore not merely aesthetic problems. They affect the credibility of embodied presence and may also reduce the reliability of spatial-intelligence scenarios that depend on correct contact, pose, affordance, and scene occupancy.

The pipeline results further show that this bottleneck is cumulative. An isolated imperfection becomes more disruptive when the body is textured, animated, retargeted, placed in a scene, and observed from an immersive perspective. Broad adoption of synthetic 4D humans in XR will therefore require progress beyond better generative models alone. It will also require stronger format interoperability, automatic retargeting, deformation-aware texture validation, scene-aware motion correction, contact-aware animation refinement, and systematic engine-level evaluation.

5 Conclusion

This study proposed SPL4SH, a modular framework for organizing synthetic 4D human creation for spatial intelligence and XR applications. The central goal was not to introduce a new reconstruction, rendering, animation, or generative model, but to examine how contemporary AI-assisted techniques for human modeling, rigging, animation, rendering, and interaction can be aligned into a coherent production workflow for immersive environments. To this end, the paper combined a structured literature-informed technique search, a stage-based taxonomy, a comparative suitability assessment, and a proof-of-concept implementation integrating SMPLify [Pavlakos *et al.*, 2019], Kimodo [Rempe *et al.*, 2026], SMPLitex [Casas and Comino-Trinidad, 2023], Rokoko-based retargeting, Blender, and Unreal Engine.

The proof-of-concept implementation demonstrates the feasibility of one SPL4SH configuration for producing textured and animated synthetic human assets that can be situated in authored XR-oriented scenes. In this configuration, SMPLify [Pavlakos *et al.*, 2019] provided stable parametric geometry and rigging, Kimodo [Rempe *et al.*, 2026] contributed controllable motion generation, SMPLitex [Casas and Comino-Trinidad, 2023] supplied SMPL-compatible texture candidates, and Unreal Engine deployment demonstrated scene-level integration. These results should not be interpreted as a general validation of all possible SPL4SH instantiations. Instead, they show how a stage-based framework can expose concrete integration bottlenecks across representation, motion, appearance, retargeting, file-format compatibility, and scene authoring.

The implementation also showed that current synthetic 4D human workflows remain fragile. The main barriers involved representation and format mismatch, manual retargeting, skeleton alignment, texture curation, motion interpenetration, static hand poses, stiff transitions, and the need for manually authored scene-level interaction. In particular, the SMPLitex stage revealed a substantial curation bottleneck: 2,089 texture candidates were generated, 65 passed the first visual screening, and 5 were selected for use in the final pipeline after Blender-based inspection. Similarly, motion and interaction integration required manual constraint editing, retargeting preparation, and Level Sequencer adjustment. These results indicate that the main challenge is not only generating assets, but making heterogeneous outputs compatible, controllable, and usable inside XR production environments.

The main contribution of this work is therefore to bridge isolated advances in 4D human generation with the practical requirements of XR-ready asset production. SPL4SH does not claim to be a fully automatic production system or a universal solution for all synthetic-human pipelines. Instead, it provides a technically grounded framework for identifying what must be completed for synthetic 4D humans to become operational in XR: stable body representation, controllable rigging, plausible temporal motion, coherent appearance, and interaction-aware scene placement. Furthermore, this contribution is especially relevant to spatial intelligence because synthetic humans in XR are not only visual avatars, but also situated temporal assets that can populate scenes, support human-centered simulation, enable embodied interaction studies, and provide controlled variation for synthetic datasets.

5.1 Study limitations

This study has limitations that should be considered when interpreting its findings. First, the technique selection process was structured and literature-informed, but it was not designed as a strict systematic review or as a fully exhaustive mapping of all available 4D human generation methods. The final set of analyzed studies and tools was selected to support the SPL4SH design and suitability assessment, so some recent, unpublished, commercial, or highly specialized techniques may not have been included.

Second, the empirical evaluation focused on a proof-of-concept implementation rather than on a large-scale quantitative benchmark. The results were analyzed through observed integration behavior, visual and functional inspection, and scene-deployment feasibility, but the study did not include controlled user studies, perceptual realism metrics, runtime performance profiling across XR hardware, or large-scale dataset validation.

Third, the pipeline evaluation used a specific combination of tools, which limits the generalizability of the implementation results to other representations, engines, motion generators, rendering methods, or interaction systems. Finally, the study emphasized technical feasibility and integration barriers, but did not yet evaluate downstream impacts on real XR users, spatial-intelligence model training, collaborative scenarios, or industrial production pipelines.

5.2 Future work

Future work should distinguish between integration points that appear close to practical resolution and those that remain deeper research challenges. Near-term improvements are likely to come from better interoperability and workflow automation, including more robust FBX/GLB export paths, automatic skeleton mapping between generated motion and SMPL-X bodies, improved rest-pose alignment, and semi-automatic retargeting validation. These problems are largely engineering-facing because they concern format compatibility, naming conventions, pose normalization, and asset-transfer procedures between existing tools.

By contrast, several bottlenecks remain farther from resolution. These include production-quality texture generation with reliable SMPL-compatible body-part placement, deformation-aware texture validation under animation, physically plausible hand and contact motion, scene-aware motion adaptation, and autonomous interaction synthesis. These problems require stronger coupling between generative models, body representations, material consistency, motion constraints, and scene geometry. Based on the results of this SPL4SH instantiation, we hypothesize that retargeting interoperability and export reliability are the nearest integration barriers to practical improvement, whereas automated texture curation, contact-aware animation, and physically grounded interaction remain the farthest from production readiness.

Declarations

Authors' contributions

DOSC: Data curation, Formal analysis, Investigation, Software, Visualization, Writing – original draft. **WLC**: Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing. **PAAO**: Investigation, Methodology, Project administration, Writing – review & editing. **JMXNT**: Funding acquisition, Project administration, Resources, Writing – review & editing. **VT**: Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing. **QL**: Funding acquisition, Project administration, Writing – review & editing. All authors read and approved the final manuscript

Competing interests

QL is an employee of HP Inc. This work was conducted within the scope of an R&D cooperation project between Voxar Labs and HP Inc. R&D Brazil. The study does not evaluate, compare, or endorse any HP products or services. The authors declare no other competing interests.

Acknowledgements

This work was carried out in cooperation with HP Inc. R&D Brazil and Voxar Labs.

Funding

This work was supported by HP Inc. R&D Brazil through an R&D cooperation project with Voxar Labs, under incentives of the Brazilian Informatics Law (Lei de Informática, Law No. 8,248/1991).

Availability of data and materials

Data and materials supporting this study are available upon request at `https://spl4sh-jbcs.github.io/`.

References

Baradel, F., Armando, M., Galaaoui, S., Brégier, R., Weinzaepfel, P., Rogez, G., and Lucas, T. (2024). Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. DOI: 10.48550/arXiv.2402.14654.

Bekor, Y., Harari, G. M., Perel, O., and Litany, O. (2025). Gaussian see, gaussian do: Semantic 3d motion transfer from multiview video. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, SA Conference Papers '25, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3757377.3763992.

Bian, S., Xu, C., Xiu, Y., Grigorev, A., Liu, Z., Lu, C., Black, M. J., and Feng, Y. (2025). Chatgarment: Garment estimation, generation and editing via large language models. DOI: 10.48550/arXiv.2412.17811.

Cao, D., Sun, G., Habermann, M., and Bernard, F. (2025). Hyper diffusion avatars: Dynamic human avatar generation using network weight space diffusion. *arXiv preprint*. DOI: 10.48550/arXiv.2509.04145.

Cao, Y., Cao, Y.-P., Han, K., Shan, Y., and Wong, K.-Y. K. (2024). Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. pages 958–968. DOI: 10.1109/CVPR52733.2024.00097.

Casas, D. and Comino-Trinidad, M. (2023). Smplitex: A generative model and dataset for 3d human texture estimation from single image. In *British Machine Vision Conference*. DOI: 10.48550/arXiv.2309.01855.

Cseke, A., Tripathi, S., Dwivedi, S. K., Lakshmipathy, A., Chatterjee, A., Black, M. J., and Tzionas, D. (2025). Pico: Reconstructing 3d people in contact with objects. DOI: 10.48550/arXiv.2504.17695.

Dai, P., Tan, F., Yu, X., Peng, Y., Wang, R., Wang, J., and Liu, Y. (2025). Go-nerf: Generating objects in neural radiance fields for virtual reality content creation. *IEEE Transactions on Visualization and Computer Graphics*, 31(5):3087–3097. DOI: 10.1109/TVCG.2025.3549558.

de Andrade Araujo, V. F., Costa, A. B., and Musse, S. R. (2023). Evaluating the uncanny valley effect in dark colored skin virtual humans. In *2023 36th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 1–6. IEEE. DOI: 10.1109/SIBGRAPI59091.2023.10347145.

Fiche, G., Leglaive, S., Alameda-Pineda, X., and Moreno-Noguer, F. (2025). Mega: Masked generative autoencoder for human mesh recovery. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5366–5378. DOI: 10.1109/CVPR52734.2025.00505.

Gong, C., Dai, Y., Li, R., Bao, A., Li, J., Yang, J., Zhang, Y., and Li, X. (2024). Text2avatar: Text to 3d human avatar generation with codebook-driven body controllable attribute. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16–20. DOI: 10.1109/ICASSP48485.2024.10446237.

Gonzalez Morin, D., Gonzalez-Sosa, E., Perez, P., and Villegas, A. (2023). Full body video-based self-avatars for mixed reality: from e2e system to user study. *Virtual Reality*, 27(3):2129–2147. DOI: 10.1007/s10055-023-00785-0.

Guo, C., Li, J., Kant, Y., Sheikh, Y., Saito, S., and Cao, C. (2025). Vid2avatar-pro: Authentic avatar from videos in the wild via universal prior. DOI: 10.48550/arXiv.2503.01610.

Habermann, M., Liu, L., Xu, W., Pons-Moll, G., Zollhoefer, M., and Theobalt, C. (2023). Hdhumans: A hybrid approach for high-fidelity digital humans. *Proc. ACM Comput. Graph. Interact. Tech.*, 6(3). DOI: 10.1145/3606927.

Heagerty, J., Li, S., Lee, E., Bhattacharyya, S., Bista, S., Brawn, B., Feng, B. Y., Jabbireddy, S., JaJa, J., Kacorri, H., Li, D., Yarnell, D., Zwicker, M., and Varshney, A. (2024). Holocamera: Advanced volumetric capture for cinematic-quality vr applications. *IEEE Transactions on Visualization and Computer Graphics*, 30(5):2767–2775. DOI: 10.1109/TVCG.2024.3372123.

Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., and Tan, H. (2024). Lrm: Large reconstruction model for single image to 3d. *arXiv preprint*. DOI: 10.48550/arXiv.2311.04400.

Hu, E., Li, M., Qian, X., Olwal, A., Kim, D., Heo, S., and Du, R. (2024a). Experiencing thing2reality: Transforming 2d content into conditioned multiviews and 3d gaussian objects for xr communication. In *Adjunct Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–3. ACM. DOI: 10.1145/3672539.3686740.

Hu, H., Fan, Z., Wu, T., Xi, Y., Lee, S., Pavlakos, G., and Wang, Z. (2024b). Expressive gaussian human avatars from monocular rgb video. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C., editors, *Advances in Neural Information Processing Systems*, volume 37, pages 5646–5660. Curran Associates, Inc.. DOI: 10.52202/079017-0183.

Hu, S., Hong, F., Hu, T., Pan, L., Mei, H., Xiao, W., Yang, L., and Liu, Z. (2025). Humanliff: Layer-wise 3d human diffusion model. *International Journal of Computer Vision*, 133(9):5938–5957. DOI: 10.1007/s11263-025-02477-5.

Kerbl, B., Kopanas, G., Leimkühler, T., and Drettakis, G. (2023). 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4). DOI: 10.1145/3592433.

Knob, P., Pinho, G., Silva, G. F., Montanha, R., Peres, V., Araujo, V., and Musse, S. R. (2024). Surveying the evolution of virtual humans expressiveness toward real humans. *Computers & Graphics*, 123:104034. DOI: 10.1016/j.cag.2024.104034.

Kolotouros, N., Alldieck, T., Zanfir, A., Bazavan, E. G., Fieraru, M., and Sminchisescu, C. (2023). Dreamhuman: animatable 3d avatars from text. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.. DOI: 10.48550/arXiv.2306.09329.

Korban, M. and Li, X. (2022). A survey on applications of digital human avatars toward virtual co-presence. *CoRR*. DOI: 10.48550/arXiv.2201.04168.

Li, C., Chibane, J., He, Y., Pearl, N., Geiger, A., and Pons-moll, G. (2024). Unimotion: Unifying 3d human motion synthesis and understanding. DOI: 10.48550/arXiv.2409.15904.

Li, C., Zhang, C., Waghwase, A., Lee, L.-H., Rameau, F., Yang, Y., Bae, S.-H., and Hong, C.-S. (2023). Generative ai meets 3d: A survey on text-to-3d in aigc era. *ArXiv*, abs/2305.06131. DOI: 10.48550/arXiv.2305.06131.

Li, J., Cao, J., Zhang, H., Rempe, D., Kautz, J., Iqbal, U., and Yuan, Y. (2025a). Genmo: A generalist model for human motion. DOI: 10.48550/arXiv.2505.01425.

Li, K., Masuda, M., Schmidt, S., and Mori, S. (2025b). Radiance fields in xr: A survey on how radiance fields are envisioned and addressed for xr research. *IEEE Transactions on Visualization and Computer Graphics*. DOI: 10.1109/TVCG.2025.3616794.

Li, X., Ma, Q., Lin, T.-Y., Chen, Y., Jiang, C., Liu, M.-Y., and Xiang, D. (2025c). Articulated kinematics distillation from video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17571–17581. DOI: 10.1109/CVPR52734.2025.01637.

Li, X., Wang, J., Cheng, Y., Zeng, Y., Ren, X., Zhu, W., Zhao, W., and Yan, Y. (2025d). Towards high-fidelity 3d talking avatar with personalized dynamic texture. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 204–214. DOI: 10.1109/CVPR52734.2025.00028.

Ling, H., Kim, S. W., Torralba, A., Fidler, S., and Kreis, K. (2024). Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8576–8588. DOI: 10.1109/CVPR52733.2024.00819.

Liu, X., Hou, H., Yang, Y., Li, Y.-L., and Lu, C. (2023). Revisit human-scene interaction via space occupancy. *arXiv preprint arXiv:2312.02700*. DOI: 10.48550/arXiv.2312.02700.

Liu, Y., Zhu, J., Tang, J., Zhang, S., Zhang, J., Cao, W., Wang, C., Wu, Y., and Huang, D. (2024). Texdreamer: Towards zero-shot high-fidelity 3d human texture generation. DOI: 10.48550/arXiv.2403.12906.

Long, X. *et al*. (2024). Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint*. DOI: 10.48550/arXiv.2310.15008.

Menezes, E., Leal, H., Moura, J. V., Araujo, V., and Raupp Musse, S. (2025). Evaluating skin tone biases in virtual human rendering. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Posters*, pages 1–2. DOI: 10.1145/3721250.3743031.

Miao, Q., Li, K., Quan, J., Min, Z., Ma, S., Xu, Y., Yang, Y., and Luo, Y. (2025). Advances in 4d generation: A survey. *arXiv preprint*. DOI: 10.48550/arXiv.2503.14501.

Montanha, R., Araujo, V., Knob, P., Pinho, G., Fonseca, G., Peres, V., and Musse, S. R. (2023). Crafting realistic virtual humans: Unveiling perspectives on human perception, crowds, and embodied conversational agents. In *2023 36th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 252–257. IEEE. DOI: 10.1109/SIBGRAPI59091.2023.10347175.

Nag, S., Cohen-Or, D., Zhang, H., and Amiri, A. M. (2025). In-2-4d: Inbetweening from two single-view images to 4d generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, SA Conference Papers '25, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3757377.3763904.

Pan, Z., Yang, Z., Zhu, X., and Zhang, L. (2025). Efficient4d: Fast dynamic 3d object generation from a single-view video. *International Journal of Computer Vision*, 134(1):14. DOI: 10.1007/s11263-025-02615-z.

Pang, H. E., Liu, S., Cai, Z., Yang, L., Zhang, T., and Liu, Z. (2025). Disco4d: Disentangled 4d human generation and animation from a single image. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26331–26344. DOI: 10.1109/CVPR52734.2025.02452.

Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A., Tzionas, D., and Black, M. J. (2019). Expressive body capture: 3d hands, face, and body from a single image. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10967–10977. DOI: 10.1109/CVPR.2019.01123.

Peng, H.-Y., Zhang, J.-P., Guo, M.-H., Cao, Y.-P., and Hu, S.-M. (2024). Charactergen: Efficient 3d character generation from single images with multi-view pose canonicalization. *ACM Trans. Graph.*, 43(4). DOI: 10.1145/3658217.

Rempe, D., Petrovich, M., Yuan, Y., Zhang, H., Peng, X. B., Jiang, Y., Wang, T., Iqbal, U., Minor, D., de Ruyter, M., Li, J., Tessler, C., Lim, E., Jeong, E., Wu, S., Hassani, E., Huang, M., Yu, J.-B., Chung, C., Song, L., Dionne, O., Kautz, J., Yuen, S., and Fidler, S. (2026). Kimodo: Scaling controllable human motion generation. DOI: 10.48550/arXiv.2603.15546.

Saito, J., Li, J., de Ruyter, M., Guerrero, M., Lim, E., Hassani, E., Ribera, R. B., Moon, H., Dadela, M., Lucca, M. D., Wang, Q., Li, X., Kautz, J., Yuen, S., and Iqbal, U. (2026). Soma: Unifying parametric human body models. DOI: 10.48550/arXiv.2603.16858 .

Sakashita, M., Kumaravel, B. T., Marquardt, N., and Wilson, A. D. (2024). Sharednerf: Leveraging photorealistic and view-dependent rendering for real-time and remote collaboration. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM. DOI: 10.1145/3613904.3642945.

Sangeeth Chandran, J. K., Llorens Salvador, M., and Ennis, C. (2026). Virtual humans in virtual reality: a scoping review on sociability, fidelity, and expression. *Frontiers in Virtual Reality*, Volume 7 - 2026. DOI: 10.3389/frvir.2026.1738000.

Shao, R., Xu, Y., Shen, Y., Yang, C., Zheng, Y., Chen, C., Liu, Y., and Wetzstein, G. (2025). Interspatial attention for efficient 4d human video generation. *ACM Transactions on Graphics*, 44(4):1–16. DOI: 10.1145/3731165.

Shao, R. *et al*. (2024). 360-degree human video generation with 4d diffusion transformer. *ACM Transactions on Graphics*, 43(6). DOI: 10.1145/3687980.

Shaw, R., Jang, Y., Papaioannou, A., Moreau, A., Dhamo, H., Zhang, Z., and Pérez-Pellitero, E. (2026). An interactive conversational 3d virtual human. *International Journal of Computer Vision*, 134(4):161. DOI: 10.1007/s11263-025-02725-8.

Shen, K., Guo, C., Kaufmann, M., Zarate, J. J., Valentin, J., Song, J., and Hilliges, O. (2023). X-avatar: Expressive human avatars. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16911–16921. DOI: 10.48550/arXiv.2303.04805.

Smolic, A., Amplianitis, K., Moynihan, M., O'Dwyer, N. C., Ondrej, J., Pagés, R., Young, G. W., and Zerman, E. (2022). Volumetric video content creation for immersive xr experiences. In *London Imaging Meeting 2022: Display Science, LIM 2022*, pages 54–59. Society for Imaging Science and Technology. DOI: 10.2352/LIM.2022.1.1.13.

Sun, G., Chen, X., Chen, Y., Pang, A., Lin, P., Jiang, Y., Xu, L., Yu, J., and Wang, J. (2021). Neural free-viewpoint performance rendering under complex human-object interactions. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 4651–4660, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3474085.3475442.

Sun, G., Dabral, R., Zhu, H., Fua, P., Theobalt, C., and Habermann, M. (2025). Real-time free-view human rendering from sparse-view rgb videos using double unprojected textures. DOI: 10.48550/arXiv.2412.13183.

Taubner, F., Zhang, R., Tuli, M., and Lindell, D. B. (2025). Cap4d: Creating animatable 4d portrait avatars with morphable multi-view diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5318–5330. DOI: 10.1109/CVPR52734.2025.00501.

Tian, Z., Weng, D., Fang, H., Guo, H., and Bao, Y. (2024). 4d facial capture pipeline incorporating progressive retopology approach. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 741–742. DOI: 10.1109/VRW62533.2024.00167.

Uzun, Y., Riedinger, U., Biswas, S., Hetzer, S., and Oppermann, L. (2025). Enabling digital spaces: Bringing gaussian splats alive in xr. In *Extended Reality*, Lecture Notes in Computer Science, pages 82–92. Springer. DOI: 10.1007/978-3-031-97769-5_6.

Vilchis, C., Perez-Guerrero, C., Mendez-Ruiz, M., and Gonzalez-Mendoza, M. (2023). A survey on the pipeline evolution of facial capture and tracking for digital humans. *Multimedia Syst.*, 29(4):1917–1940. DOI: 10.1007/s00530-023-01081-2.

Wang, R., Cao, Y., Han, K., and Wong, K.-Y. K. (2024). A survey on 3d human avatar modeling - from reconstruction to generation. *ArXiv*, abs/2406.04253. DOI: 10.48550/arXiv.2406.04253.

Wang, Y., Sun, Y., Patel, P., Daniilidis, K., Black, M. J., and Kocabas, M. (2025). Prompthmr: Promptable human mesh recovery. DOI: 10.48550/arXiv.2504.06397.

Weng, S. C., Chiou, Y., and Do, E. Y. (2024). Dream mesh: A speech-to-3d model generative pipeline in mixed reality. In *2024 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)*, pages 345–349. IEEE. DOI: 10.1109/AIxVR59861.2024.00059.

Xu, Y., Yang, Z., and Yang, Y. (2026). Photorealistic text-to-3d avatar generation with constraints for decoupled geometry and appearance. *ACM Trans. Multimedia Comput. Commun. Appl.*, 22(2). DOI: 10.1145/3774422.

Yin, M., Cao, Y., Peng, S., and Han, K. (2025). Splat4d: Diffusion-enhanced 4d gaussian splatting for temporally and spatially consistent content creation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, SIGGRAPH Conference Papers '25, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3721238.3730752.

Yu, Z., Zhao, Z., Du, Y., Zheng, Y., Zuo, B., and Wang, Y. (2025). T2c: Text-guided 4d cloth generation. *ACM Trans. Multimedia Comput. Commun. Appl.*, 21(7). DOI: 10.1145/3735642.

Zhang, J., Jiang, Z., Yang, D., Xu, H., Shi, Y., Song, G., Xu, Z., Wang, X., and Feng, J. (2022). Avatargen: A 3d generative model for animatable human avatars. In *Computer Vision – ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, page 668–685, Berlin, Heidelberg. Springer-Verlag. DOI: 10.1007/978-3-031-25066-8_39.

Zheng, C., Xue, L., Zarate, J., and Song, J. (2025). Gaustar: Gaussian surface tracking and reconstruction. DOI: 10.48550/arXiv.2501.10283.

Zhu, S., Guo, C., Lu, Y., Yu, J., Shen, Y., Wu, Z., Hong, Y., Cai, Y., Liu, Y., Zhang, M., Xu, Y., and Xu, L. (2026). A general framework for gaussian splatting-based human-centric volumetric videos. *Visual Intelligence*, 4:8. DOI: 10.1007/s44267-026-00111-7.