SIGGRAPH 2025

Vancouver+ 10-14 August

Human-Centric Capture and Digitalization for Immersive XR Experiences

Zhuo Su

Bytedance

Immersive XR is about bringing real humans into the virtual world

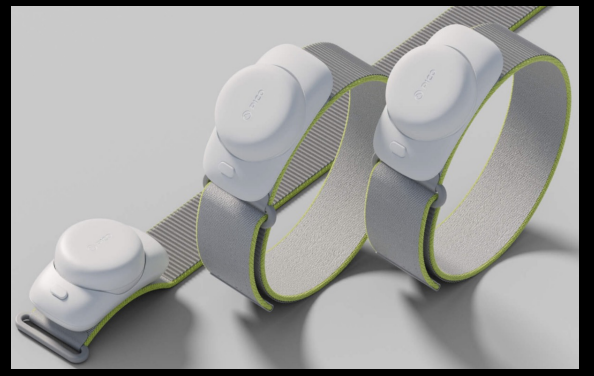
- It’s not just about virtual environments — it’s also about human presence.
- Humans need to be *captured* and *recreated* digitally.
- Core Elements:
  - Motion: how people move
  - Appearance: how they look (3D shape + texture)
  - Animation: how they act

Talk Overview

Stage in Immersive XR	Motion	Appearance	Animation
1. Motion Capture	✅	–	–
2. Reconstruction	–	✅	–
3. Performance Capture	✅	✅	–
4. Avatar Creation	✅	✅	✅

1. MoCap in XR: Why It's Hard

XR Applications



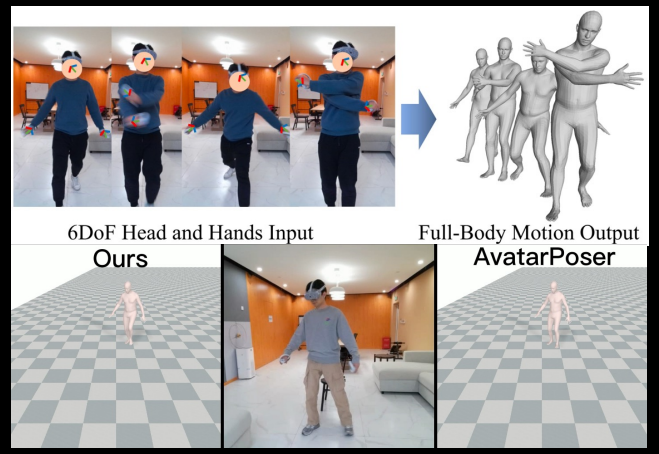
Sparse IMU sensors & ego-centric cameras

- Sparse observations, underdetermined motion: most body parts unobserved
- Human motion is highly varied, XR games involve complex and challenging motion
- Real-time constraints limit use of post processing like IK or physical simulation
- Prone to physical artifacts: sliding, floating, ground penetration

1. MoCap in XR: A Series of Solutions

AvatarJLM

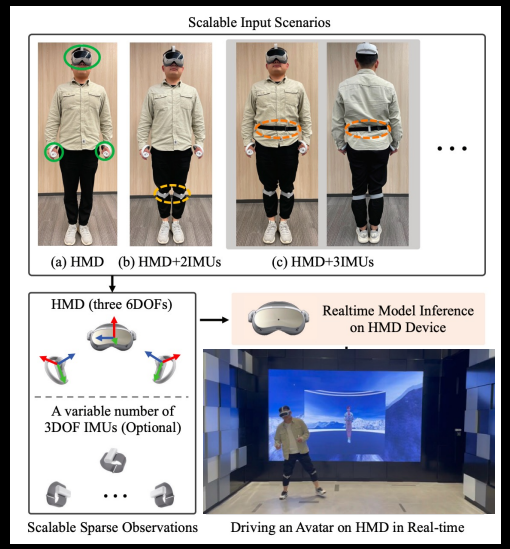
ICCV'23



6DoF of HMD and hand controllers

HMD-Poser

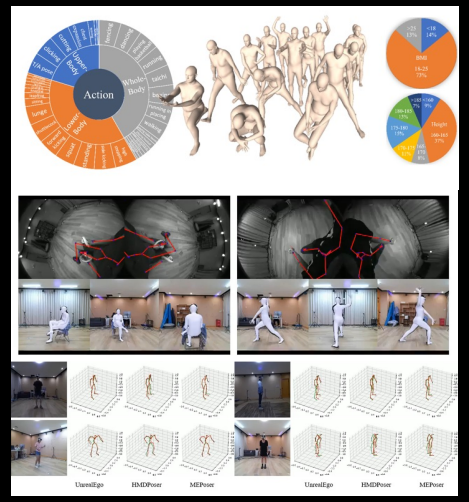
CVPR'24



+ scalable IMUs on legs or pelvis

EMHI (MEPoser)

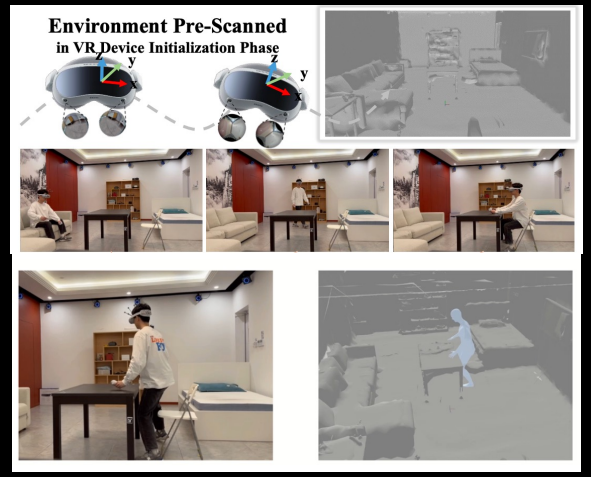
AAAI'25



+ egocentric cameras of HMD

EnvPoser

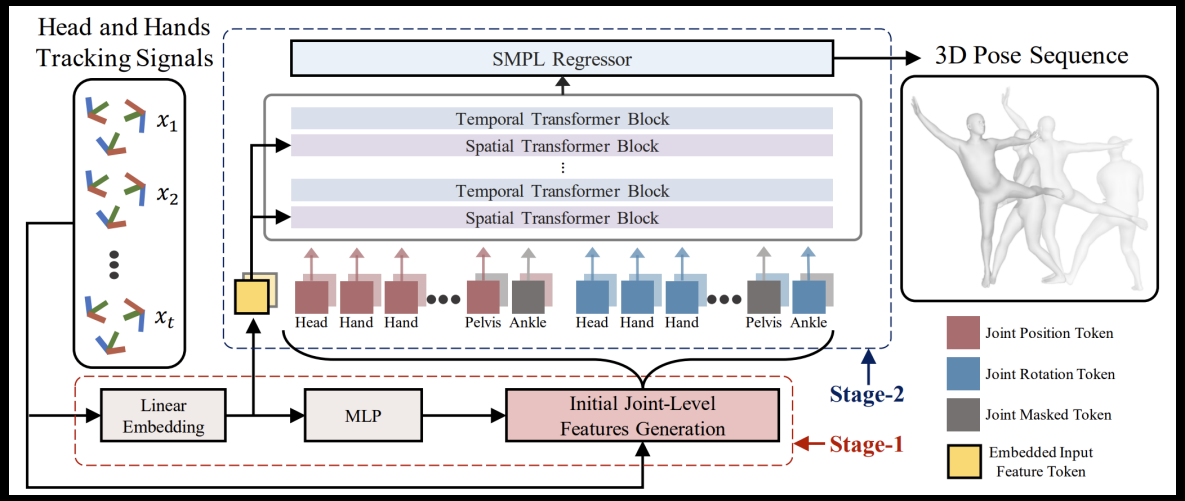
CVPR'25



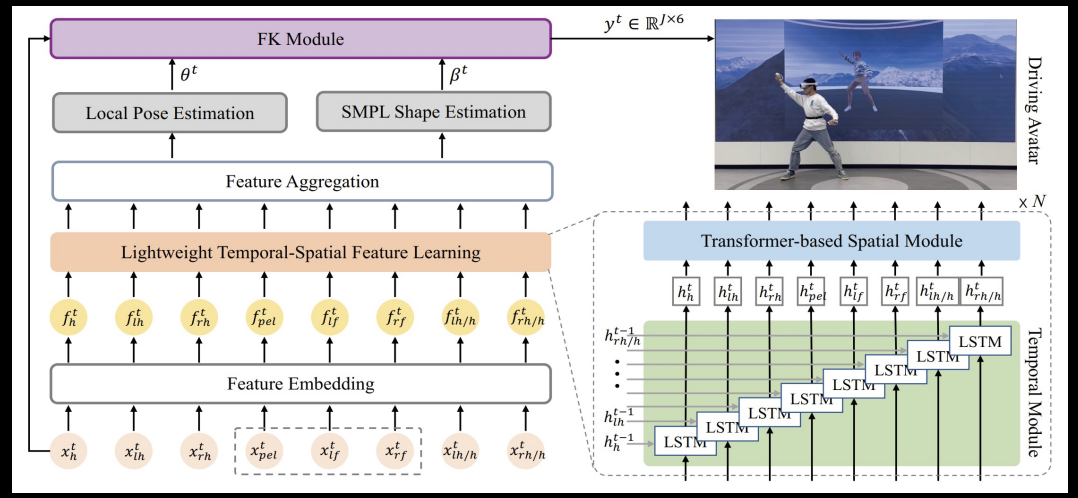
+ 3D environment from VR device

Xiaozheng, Zheng, et al. "Realistic Full-Body Tracking from Sparse Observations via Joint-Level Modeling", ICCV 2023.
Peng, Dai, et al. "HMD-Poser: On-Device Real-time Human Motion Tracking from Scalable Sparse Observations", CVPR 2024.
Fan, Zhen, et al. "EMHI: A Multimodal Egocentric Human Motion Dataset with HMD and Body-Worn IMUs", AAAI 2025.
Songpengcheng, Xia, et al. "EnvPoser: Environment-aware Realistic Human Motion Estimation from Sparse Observations with Uncertainty Modeling", CVPR 2025.

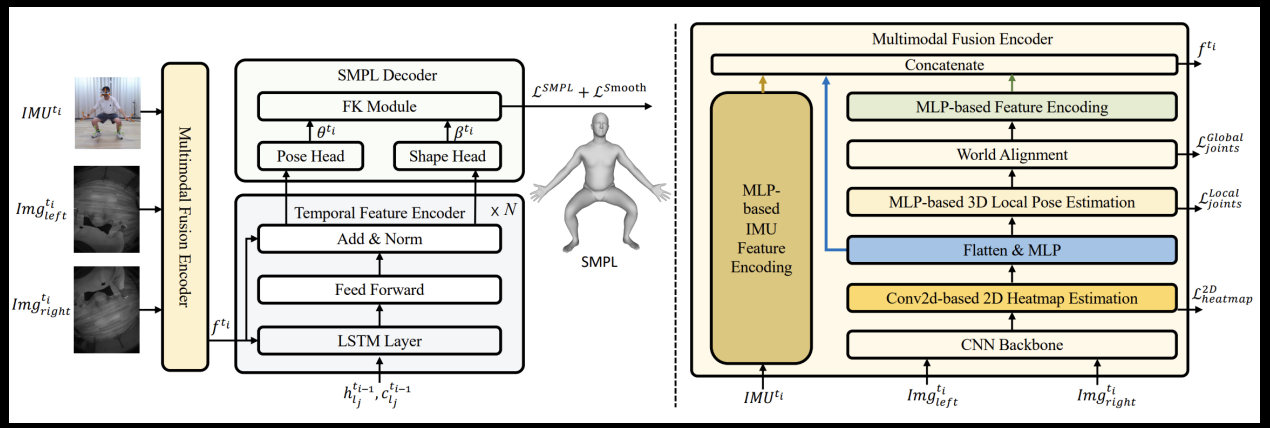
1. Frameworks of XR Mocap



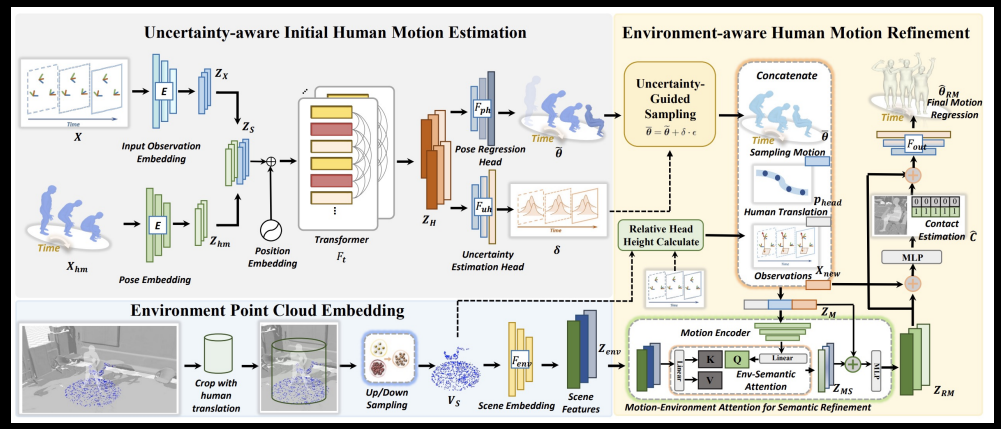
AvatarJLM



HMDPoser

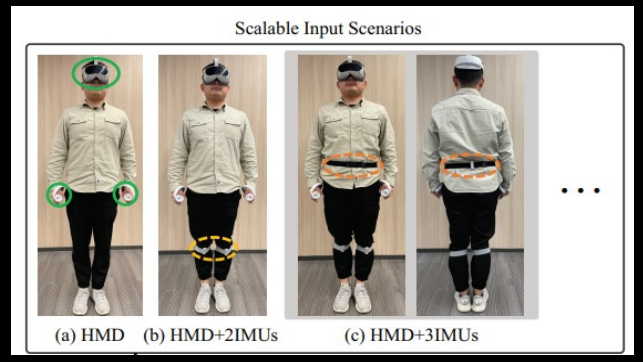


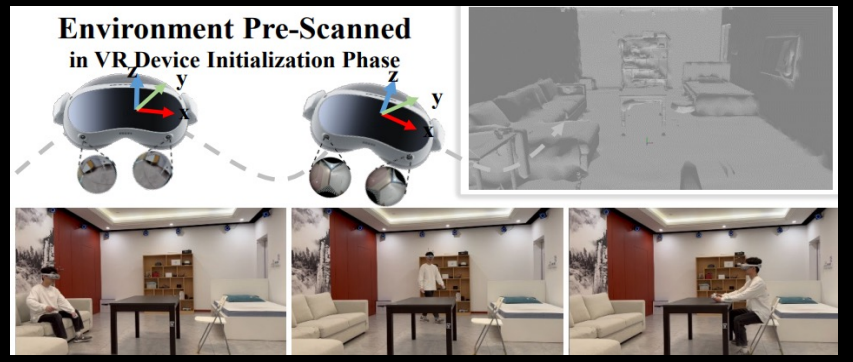
MEPoser



EnvPoser

1. Scalability and Flexibility Are Essential for XR MoCap





HMD + Hands | Legs | Egocentric view | Pre-scanned Environment

Input Embeddings from different modals

Temporal-Spatial Module (Transformer/LSTM)

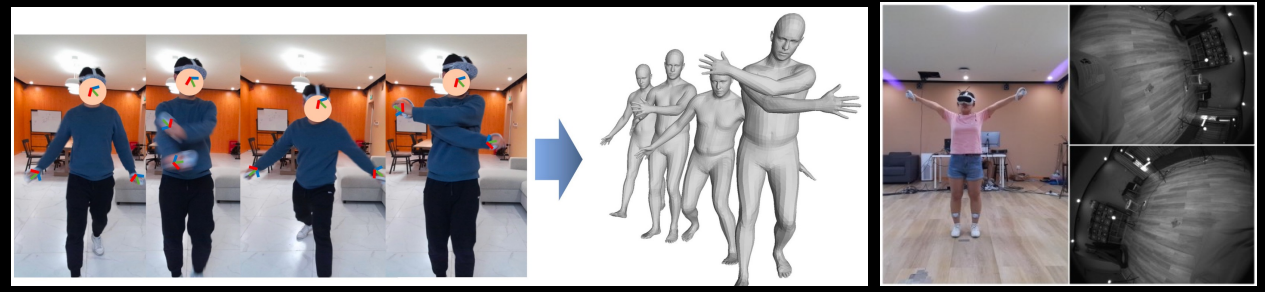
Regression Head(s)

Motion

1. Data is Equally Important for XR MoCap

Underdetermined Input → Ambiguous Motion

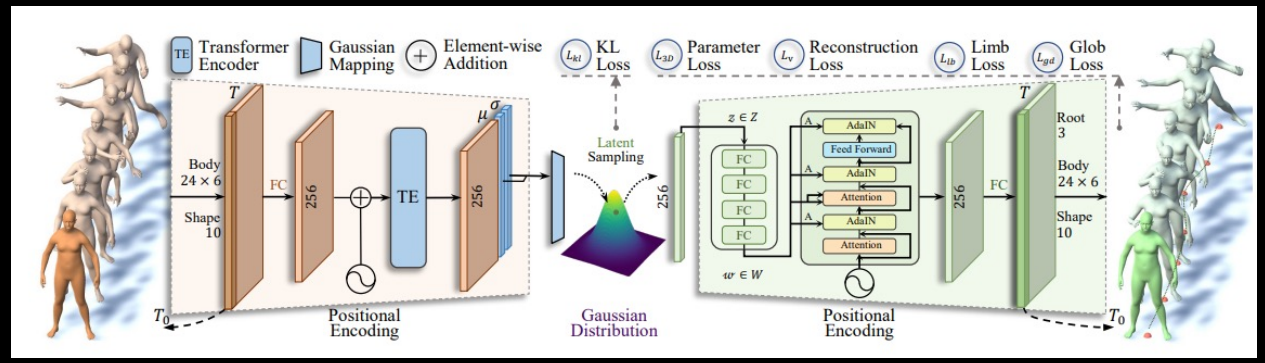
- Sparse sensors, partial observations
- Many possible poses fit the same input



Lower body is unobserved in typical XR setups.

Human Motion Has Structure

- Motion is not random — it follows patterns
- Learning these patterns from **real data** helps resolve ambiguity



Patterns like “Learning Variational Motion Prior.

Xin, Chen, et al. “Learning Variational Motion Prior for Video-based Motion Capture”, arXiv 2022.

1. Data is Equally Important for XR MoCap

SOTA Models Are Not Enough Without the Right Data

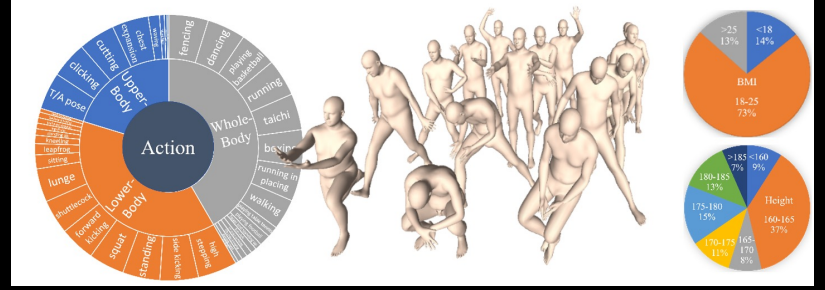
- We achieve SOTA on public benchmarks; code & models open-sourced
- But real XR needs diverse, task-specific data

We Built the Right Data With real XR sensors

- Optical MoCap → highest precision with suits
- Multi-View Marker-less Mocap → diverse clothing

Dataset	Device	Real/Synth.	Sensor Modality		SMPL(x)	Statistic		
			Egocentric Vision	Inertial		Actions	Subjects	Frames
Mo^2Cap^2		Synth.	Monocular Downward-Facing	-	-	3K	700	530K
EgoPW		Real.	Monocular Downward-Facing	-	-	20	10	318K
EgoCap		Real	Binocular Downward-Facing	-	-	-	8	30K
UnrealEgo		Synth.	Binocular Downward-Facing	-	-	30	17	450K
DIP-IMU		Real	-	Full-Body 3DoF×6	✓	15	10	330K
FreeDancing		Real	-	Full-Body 6DoF×3, 3DoF×3	✓	-	8	532.8K
Nymeria		Real	Binocular Forward-Facing	Upper-Body 6DoF×3	✓	20	264	260M
Ego-Exo4D (Ego Pose)		Real	Binocular Forward-Facing	Head 6DoF×1	-	-	-	9.6M
Ours		Real	Binocular Downward-Sloping	Full-Body 6DoF×3, 3DoF×2	✓	39	58	3.07M

Table 1. Comparison with existing egocentric motion datasets. EMHI is the first dataset that provides egocentric vision and full-body IMU signals captured by the real VR product suite, along with accurate SMPL annotations simultaneously.

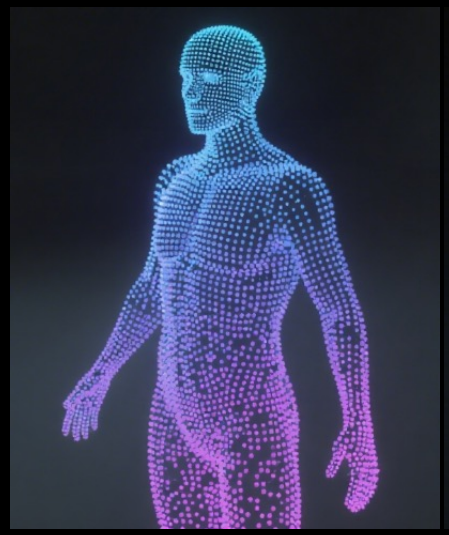


We open-sourced the EMHI dataset.

Fan, Zhen, et al. “EMHI: A Multimodal Egocentric Human Motion Dataset with HMD and Body-Worn IMUs”, AAAI 2025.

Motion alone isn't enough —
we need 3D appearance too.

2. Human Reconstruction: 3D Representation



Point Cloud

Mesh

Voxels

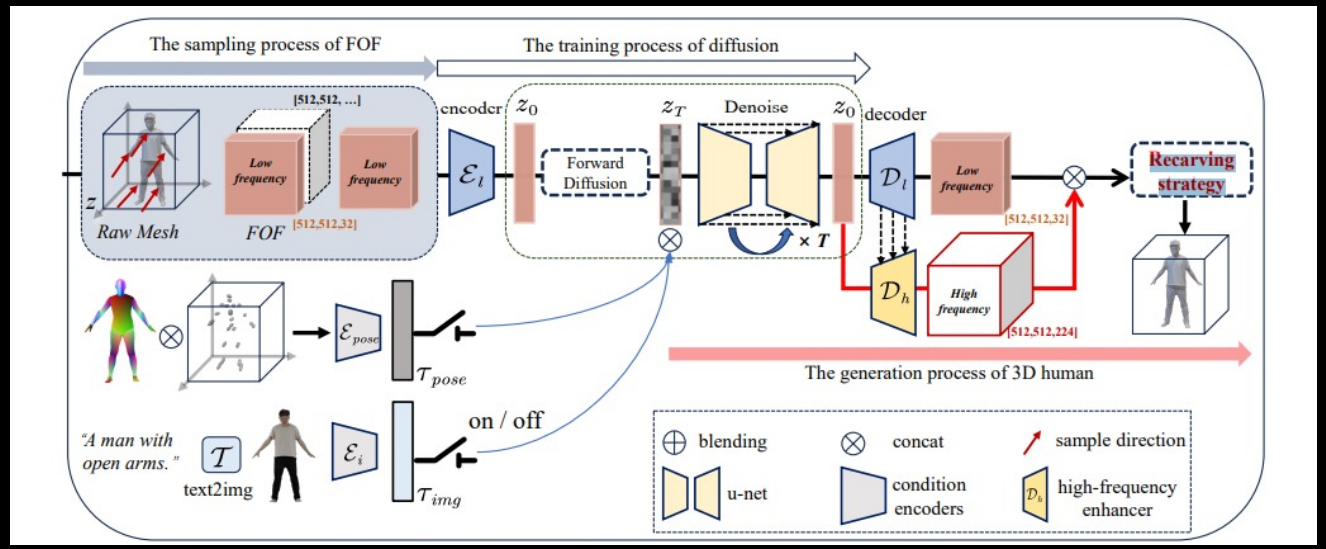
Gaussian Splatting

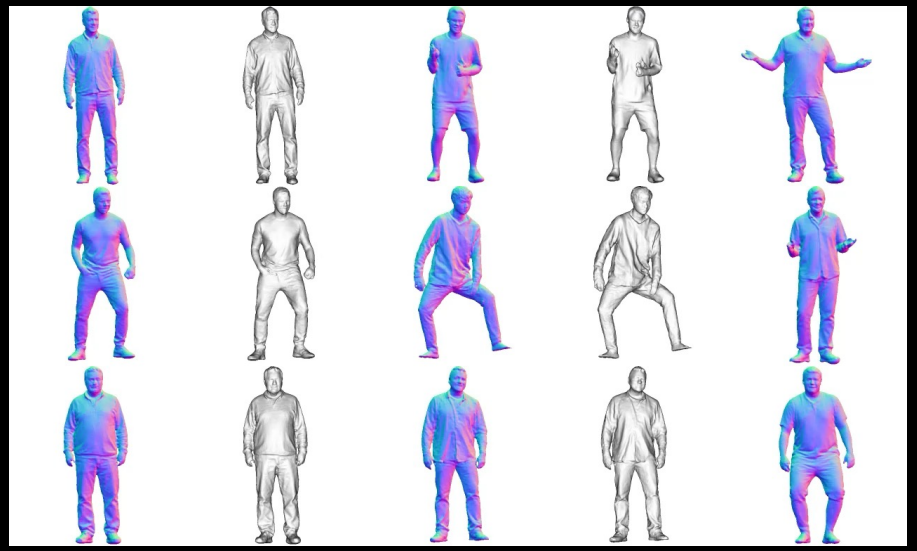
NeRF

Explicit **Implicit**

2. Shape Reconstruction

First to combine 2D diffusion and Fourier Occupancy Field for 3D generation.

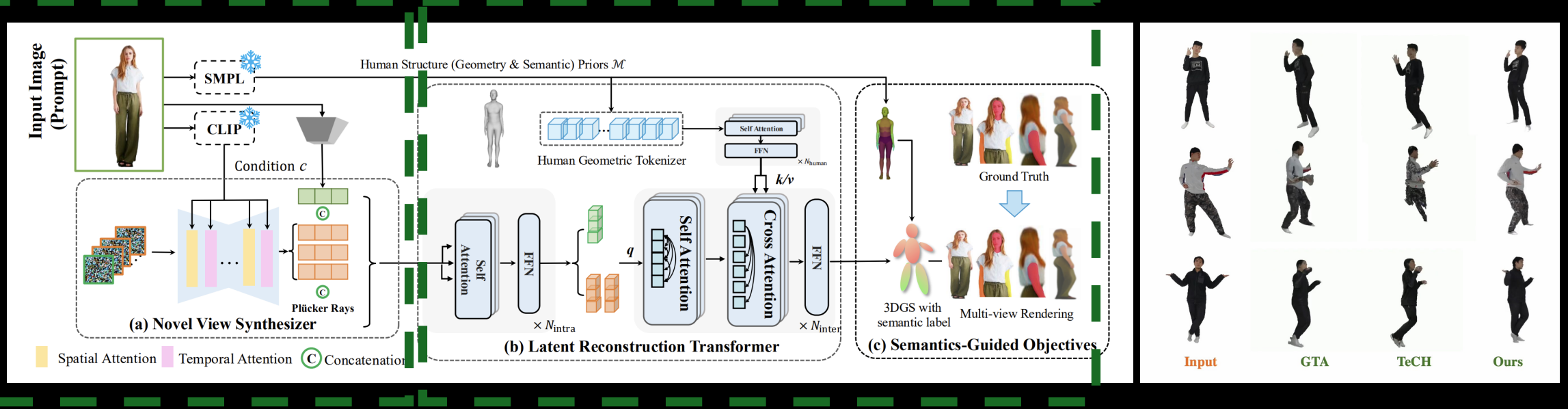




Muxin, Zhang, et al. “Joint2Human: High-quality 3D Human Generation via Compact Spherical Embedding of 3D Joints”, CVPR 2024.

2. Appearance Reconstruction

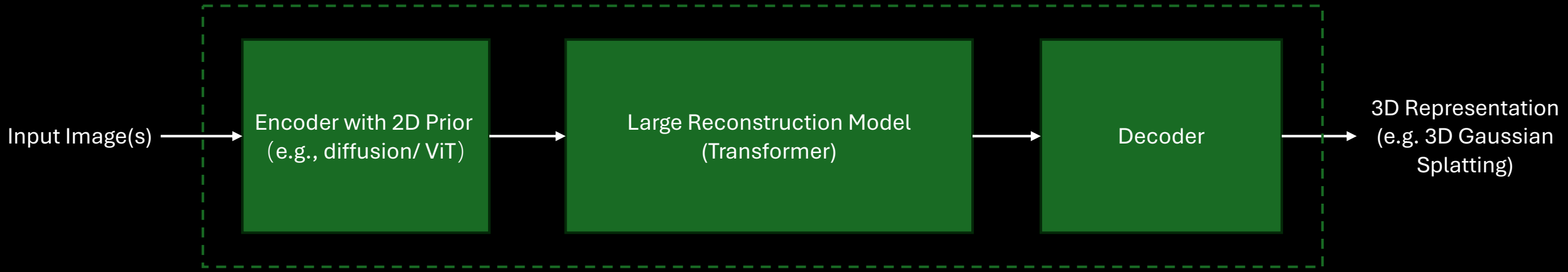
First to leverage 2D Video Generation and Large Reconstruction Model for single-image human Gaussian Splatting reconstruction.



Panwang, Pan, et al. “HumanSplat: Generalizable Single-Image Human Gaussian Splatting with Structure Priors”, NeuIPS 2024.

2. Human Reconstruction: Takeaways

- 2D priors enrich structural and texture details for better 3D reconstruction.
- Large reconstruction model enhance generalization across poses and appearances.
- Combining 2D and 3D modalities is crucial for quality and robustness.



- This framework is also suited for avatar generation.

Static Reconstruction isn’t enough — performance capture unifies motion and appearance over time.

3. Performance Capture: Early-Stage – Volumetric Capture

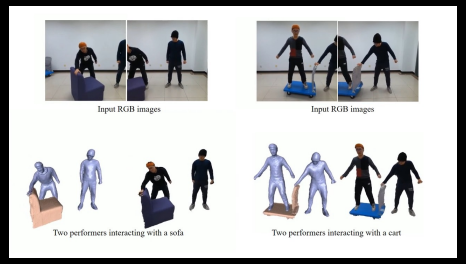
UnstructuredFusion

PAMI'19

Volumetric Capture + Non-rigid Warping

RobustFusion series

ECCV'20 / PAMI'22



+ Implicit Completion + Robust Tracking

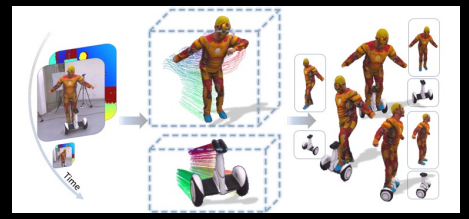
NeuralHOFusion

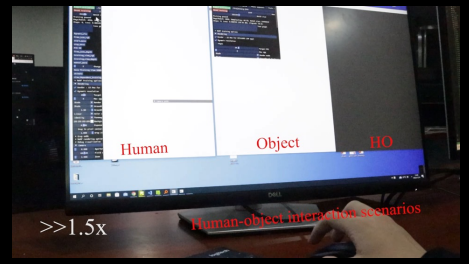
CVPR'22

+ Neural Blending-based Rendering

Instant-NVR

CVPR'23





+ Instant-NGP-based NeRF Rendering

Lan, Xu, et al. "UnstructuredFusion: Realtime 4D Geometry and Texture Reconstruction using Commercial RGBD Cameras", PAMI 2019.
Zhuo, Su, et al. "RobustFusion: Human Volumetric Capture with Data-driven Visual Cues using a RGBD Camera", ECCV 2020.
Zhuo, Su, et al. "Robust Volumetric Performance Reconstruction under Human-object Interactions from Monocular RGBD Stream", PAMI 2022.
Yuheng, Jiang, et al. "NeuralHOFusion: Neural Volumetric Rendering Under Human-Object Interactions", CVPR 2022.
Yuheng, Jiang, et al. "Instant-NVR: Instant Neural Volumetric Rendering for Human-object Interactions from Monocular RGBD Stream", CVPR 2023.

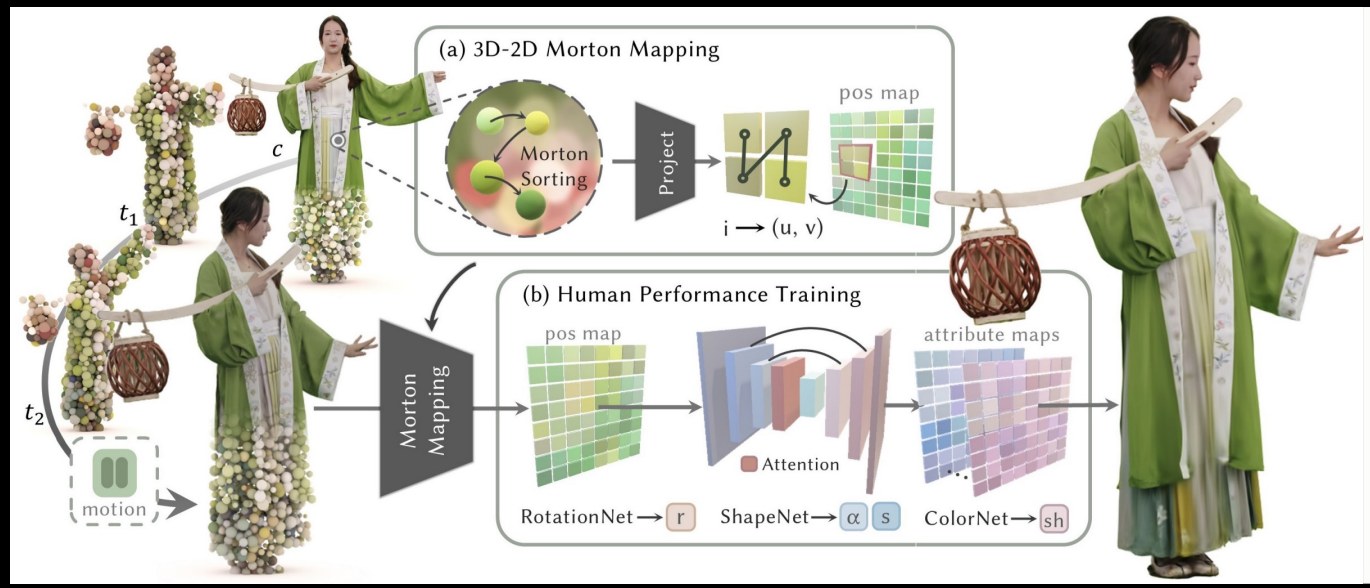
3. Performance Capture: Recent Advances – 4DGS

Bridging Gaussian Splatting and non-rigid tracking for compact volumetric video

Yuheng, Jiang, et al. “HiFi4G: High-Fidelity Human Performance Rendering via Compact Gaussian Splatting ”, CVPR 2024.

3. Performance Capture: Recent Advances – 4DGS

Unifying Playback and Re-Performance for Human-Centric Volumetric Video



Yuheng, Jiang, et al. “RePerformer: Immersive Human-centric Volumetric Videos from Playback to Photoreal Reperformance ”, CVPR 2025.

3. Performance Capture: Future work?

🔄 **More General** Human-Only → Human-Centric → General Dynamic Scene	⚡ **Faster** • Generalizable models with feed-forward inference • Real-time capture for holographic telepresence
🎥 **Fewer Sensors, Smarter Models** • Egocentric or monocular input • Learning-based priors for robust in-the-wild performance	🎯 **Task-Oriented** For downstream tasks such as: • High-quality volumetric video playback • Content generation and editing • Re-performer and motion transfer

Motion and appearance are in place — the last piece is animation. That’s where Avatar Creation comes in.

4. Avatar Creation: Definition

Performance Capture

Replays recorded motion and appearance

Avatar Creation

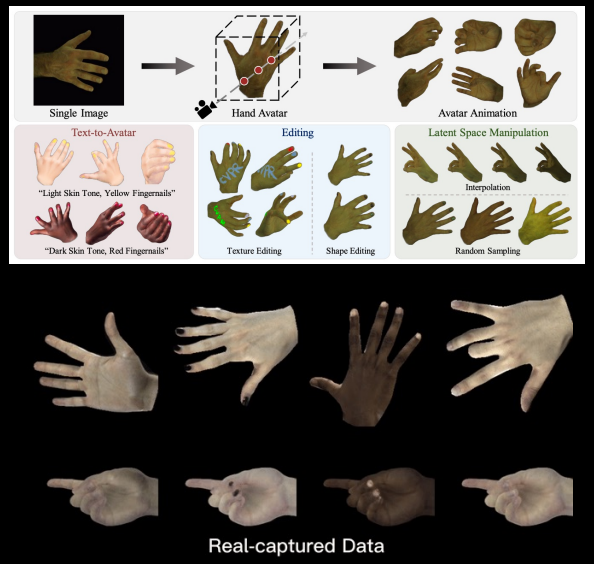
Builds animatable digital humans from limited input.

How do we build **controllable**, **generalizable**, and **realistic** avatars that go **beyond replay**, and support **animation**, **interaction**, and **immersion**?

4. Avatar Creation: A Series of Solutions

OHTA

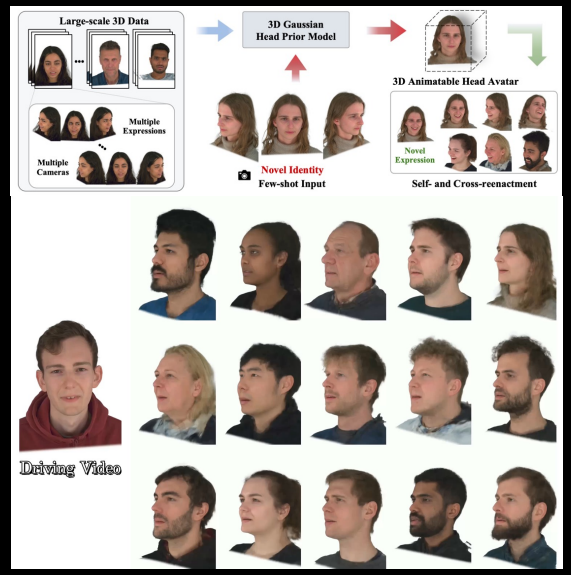
CVPR'24



One-shot hand avatar creation

HeadGap

3DV'25



Few-shot head avatar creation

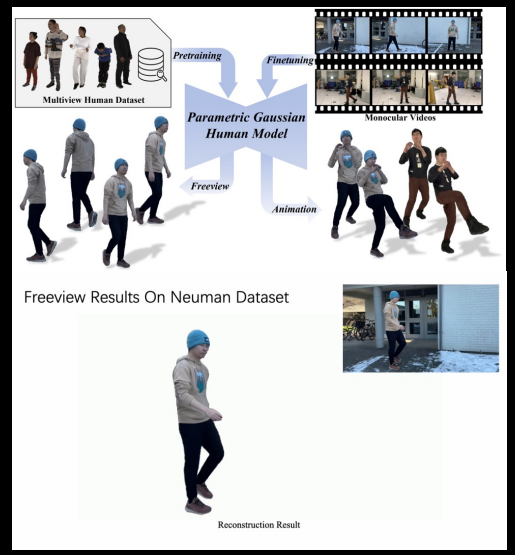
SEGA

arXiv'25

One-shot head avatar creation

PGHM

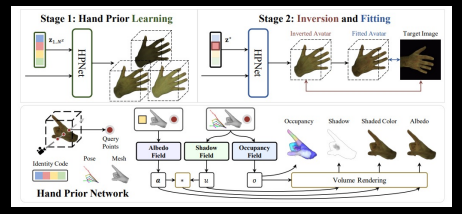
arXiv'25



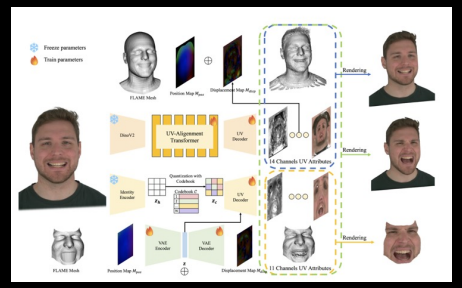
Generalizable full-body avatar creation

Xiaozheng, Zheng, et al. "OHTA: One-shot Hand Avatar via Data-driven Implicit Priors", CVPR 2024.
Xiaozheng, Zheng, et al. "HeadGAP: Few-shot 3D Head Avatar via Generalizable Gaussian Priors", 3DV 2025.
Chen, Guo, et al. "SEGA: Drivable 3D Gaussian Head Avatar from a Single Image", arXiv 2025.
Cheng, Peng, et al. "Parametric Gaussian Human Model: Generalizable Prior for Efficient and Realistic Human Avatar Modeling", arXiv 2025.

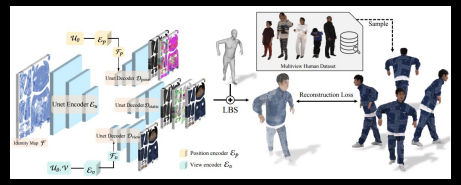
4. One/Few-shot Paradigm: Generalizable Prior Model



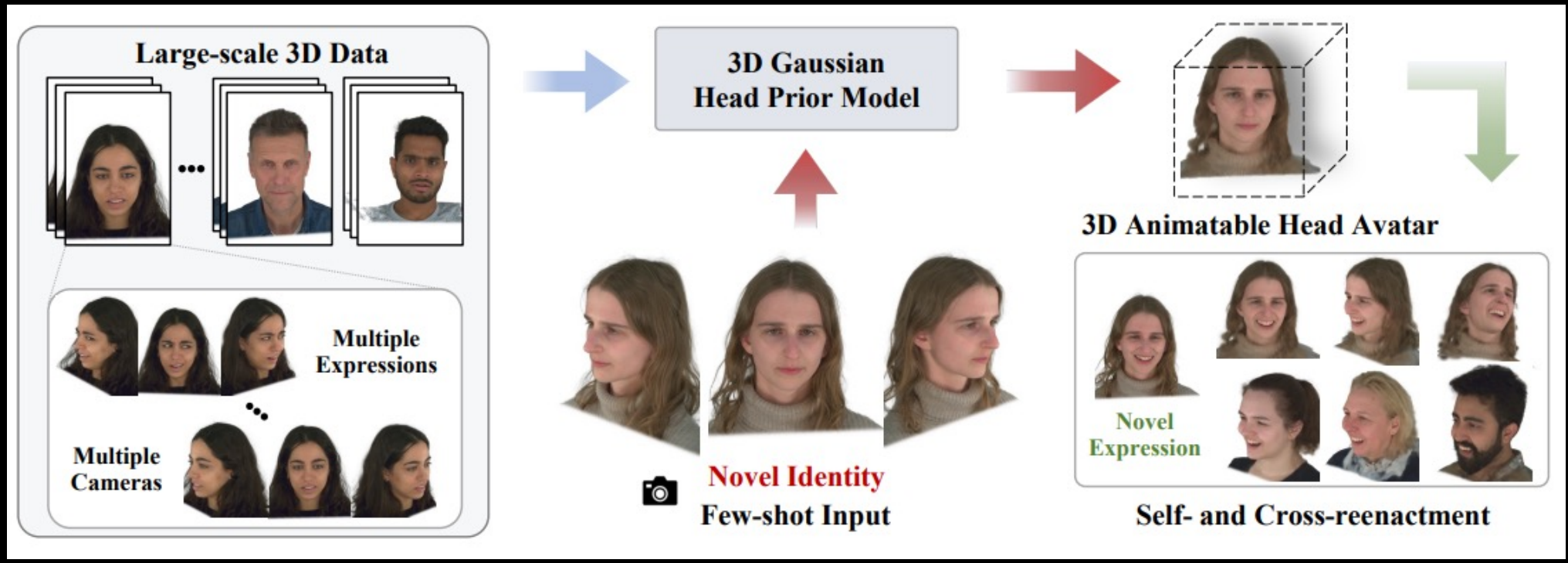
OHTA



SEGA



PGHM



HeadGap-style methods aim to generalize across identities and create one with one/few-shot adaptation.

4. Challenges and Opportunities

⚠️ **Bottlenecks in One/Few-Shot Paradigm**

- Training on expensive 3D data
- Limited Model Capacity and Scaling Bottlenecks: Quality plateaus as ID count increases
- Heavy Fine-tuning Required: Long per-identity adaptation process
- Poor Robustness & Generalization: Strong reliance on clean input; fails in complex scenes

🌱 **Inspiration: Shift by Video Foundation Models**

- Reframe Avatar Creation as a data-driven process leveraging large-scale, in-the-wild videos —
  making casual capture possible and robust.

- Could we bypass explicit 3D reconstruction for direct novel-view and novel-pose generation?

4. Avatar Creation: Trend?

So, Where Is 3D Avatar Headed?

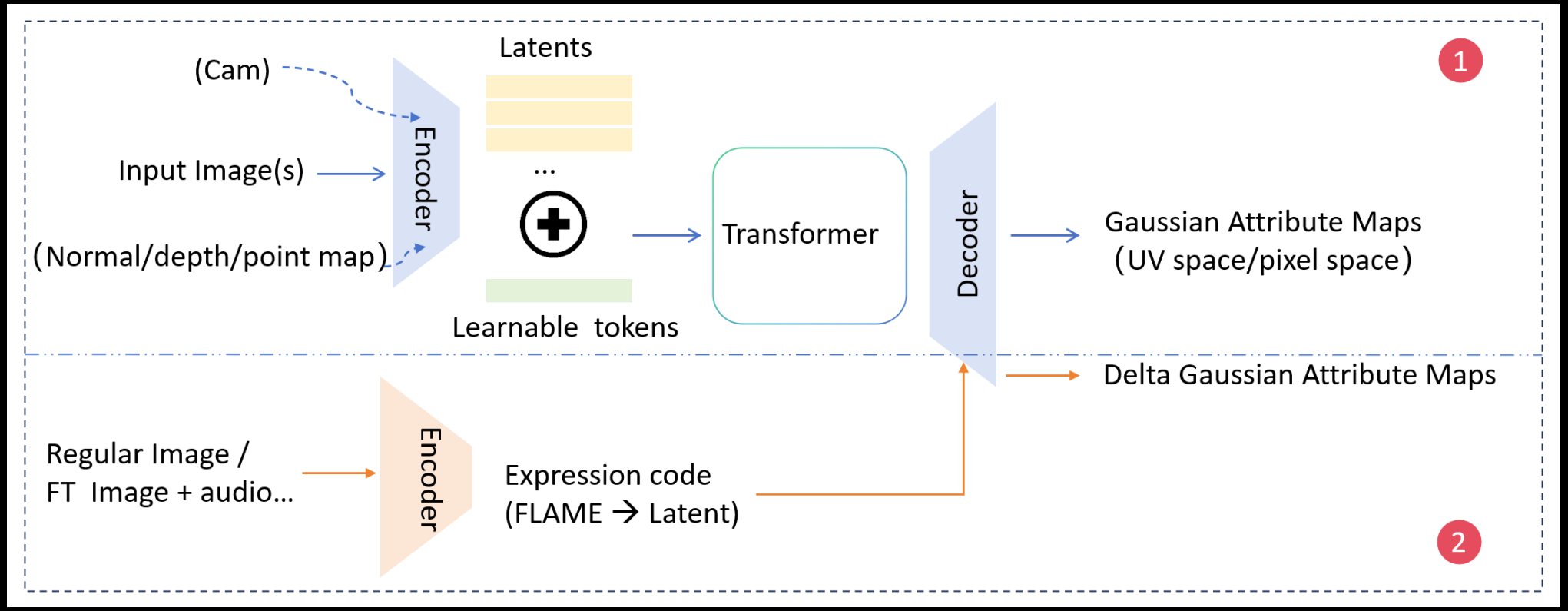
The answer is SCALING UP !!!

4. Avatar Creation: How to Scale Up

Component	Role in Scaling	Key Design Choice
Model	Framework to generate avatar from image(s)	Pretrained encoder (e.g., ViT) + LRM
Representation	Supports real-time inference compared with video generation	3D Representation (e.g., Gaussian Splatting)
Data	Fuel for model to learn generalization ability	High-quality 3D data + large-scale 2D data
Training	Fully use data of different quality for domain adaptation & scaling	Hybrid batches, pre-train + post-train

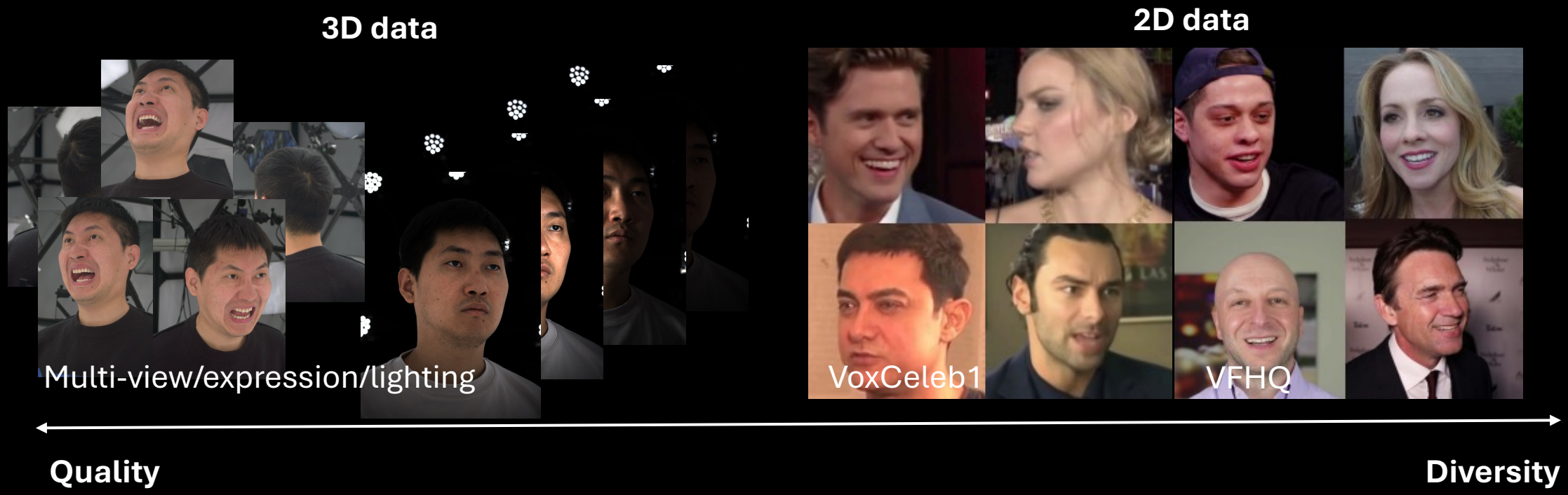
🧩 Scaling is not just about model size — it's about the right structure, supervision, and strategy.

4. Zero-Shot Paradigm: Model and Representation Examples



A generative framework using pretrained encoders and reconstruction models to efficiently create high-quality 3D avatars from 2D images via staged training.

4. Data & Training



Training Strategy: Hybrid batches to eliminate domain gap, pre-train on large scale 2D data + post-train on high-quality 3D data.

Summary: Building Human-Centric Pipelines for XR

We’ve explored the key components:

- Accurate motion capture from sparse observations
- High-fidelity reconstruction of geometry and appearance
- Performance capture that preserves expressivity and nuance.
- Avatar creation that unifies motion, appearance, and animation for scalable deployment.

Together, these pipelines bring real humans into virtual worlds — capturing not just how we move, but how we look, and express. That’s how we enable true presence.

Thank you!

SIGGRAPH 2025

Vancouver+ 10-14 August