Towards Deployable High-Fidelity Avatars: A 3D Modeling Framework Powered by Scalable Real-World Data

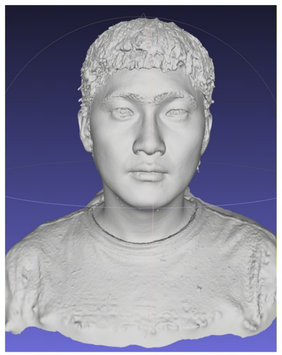
Zhuo Su 苏卓

ByteDance 字节跳动

From Demo to Deployment: The Gap in Photorealistic Digital Humans

Ideal Demo Setting:

- **Illumination:** Controlled light-stages.
- **Acquisition:** Dense multi-view rigs.
- **Annotations:** Precisely labeled camera poses and expressions.


Real-World Deployment:

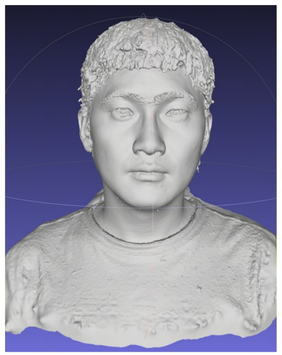
- **Illumination:** Uncontrolled indoor/outdoor environments and harsh shadows.
- **Annotations:** Unknown camera poses and random, unlabeled expressions.
- **Acquisition:** Sparse inputs limited to 1–*N* casual monocular images.



From Demo to Deployment: The Gap in Photorealistic Digital Humans

Ideal Demo Setting:

- **Illumination:** Controlled light-stages.
- **Acquisition:** Dense multi-view rigs.
- **Annotations:** Precisely labeled camera poses and expressions.


Real-World Deployment:

- **Illumination:** Uncontrolled indoor/outdoor environments and harsh shadows.
- **Annotations:** Unknown camera poses and random, unlabeled expressions.
- **Acquisition:** Sparse inputs limited to 1–*N* casual monocular images.



Conclusion: Unstructured sparse inputs lead to pseudo-3D artifacts. To cross this gap, we must rethink the foundation from Data to Model.

Scalable Data Engine: Anchoring robust generalization. → **Delighting Pre-processing:** Removing real-world lighting artifacts. → **Flexible LRM:** Adapting to unstructured sparse inputs.

Data Engine: Bridging the Reality Gap with Targeted Data Design

3D captures anchor quality, 2D wild data scales diversity, and generative synthesis bridges the gap—together forming the ultimate data engine for our Avatar backbone.

3D Data： Multi-View/Expression/Lighting

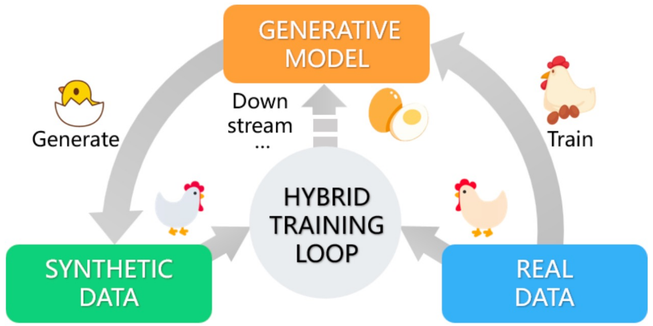
2D Data： Unknown Camera/Expression/Lighting

Quality → **Diversity**

Data Engine: Bridging the Reality Gap with Targeted Data Design

3D captures anchor quality, 2D wild data scales diversity, and generative synthesis bridges the gap—together forming the ultimate data engine for our Avatar backbone.

3D Data： Multi-View/Expression/Lighting



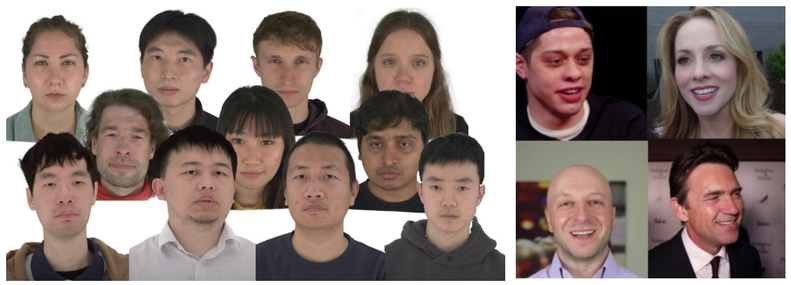
The Generative Loop: Real Data ↔ Synthetic Data

2D Data： Unknown Camera/Expression/Lighting

Quality **Diversity**

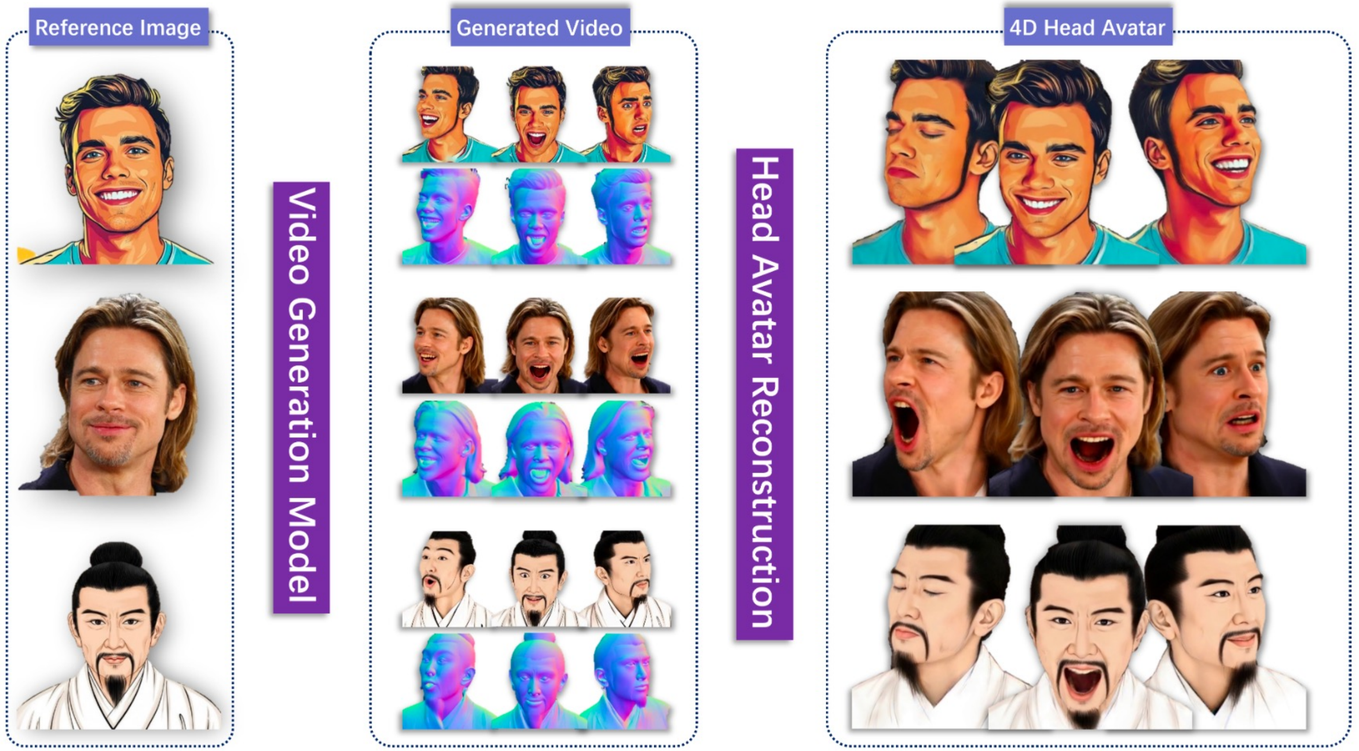
Foundation Data: Unifying Real-World Priors and 4D Synthesis

We learn a geometry-aware 4D synthesis engine from real-world priors, and then unify both real and synthesized assets to build a massive foundation for Avatar training.



Real-World Priors:（3D Dome + 2D Wild Data)

Synthesized 4D Assets

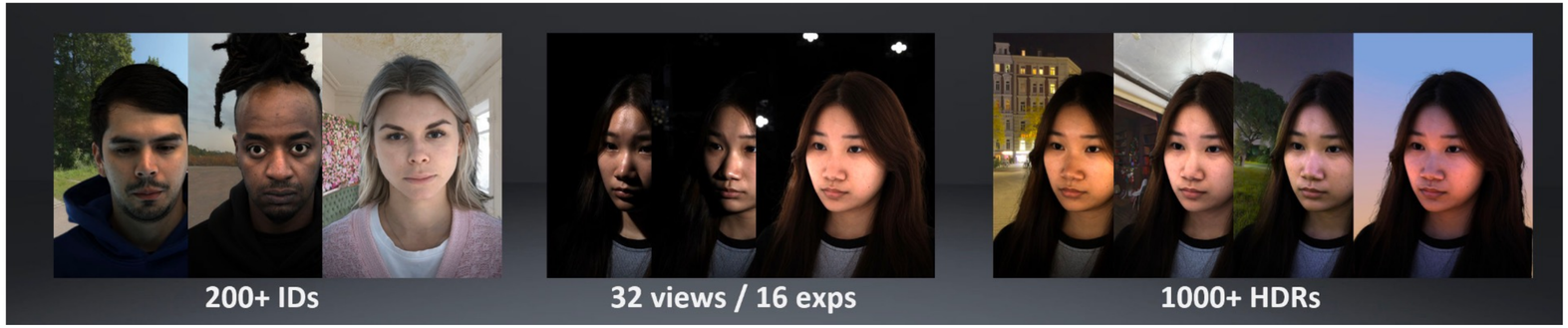


GeoDiff4D Engine: Jointly synthesizing video frames and normal for geometrically consistency

Xu, Chao, et al. "GeoDiff4D: Geometry-Aware Diffusion for 4D Head Avatar Reconstruction." CVPR 2026.

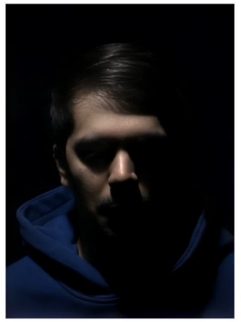
Lighting Data: From Physical Capture to Generative Expansion

Physical Capture: Capturing high-fidelity One-Light-at-a-Time (OLAT) data in the lab.

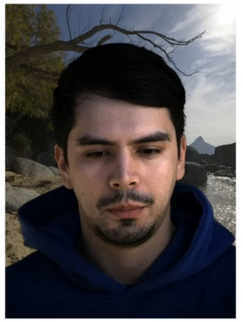


GT OLAT

Generated OLAT



Relit portrait from generated OLAT

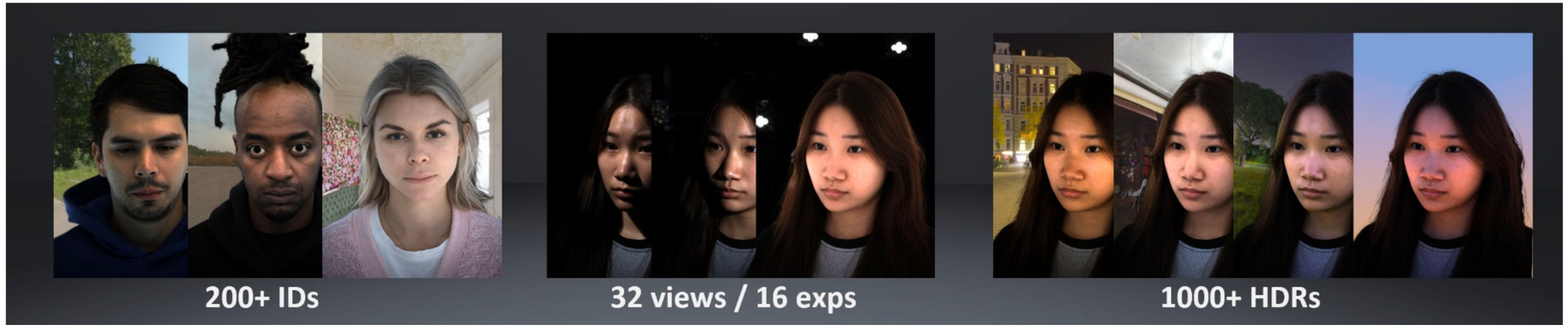


ByteDance字节跳动

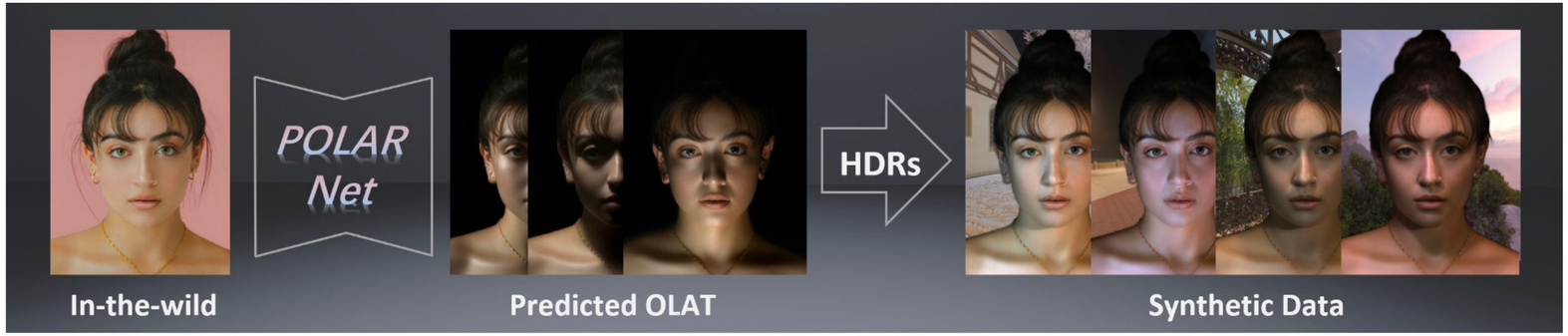
Chen, Zhuo, et al. "POLAR: A Portrait OLAT Dataset and Generative Framework for Illumination-Aware Face Modeling." CVPR 2026.

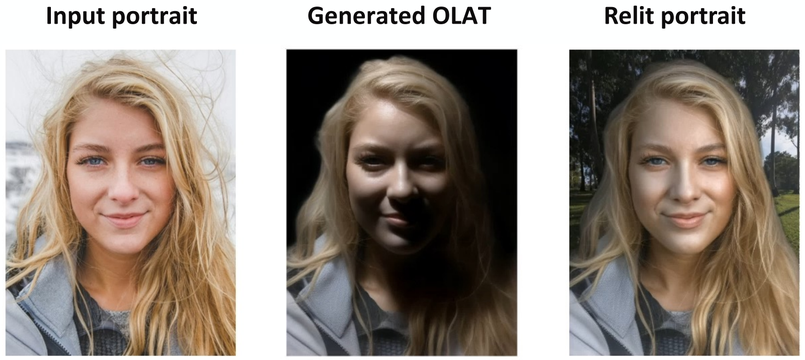
Lighting Data: From Physical Capture to Generative Expansion

Physical Capture: Capturing high-fidelity One-Light-at-a-Time (OLAT) data in the lab.



Generative Expansion: Scaling lab captures to in-the-wild identities via POLARNet.

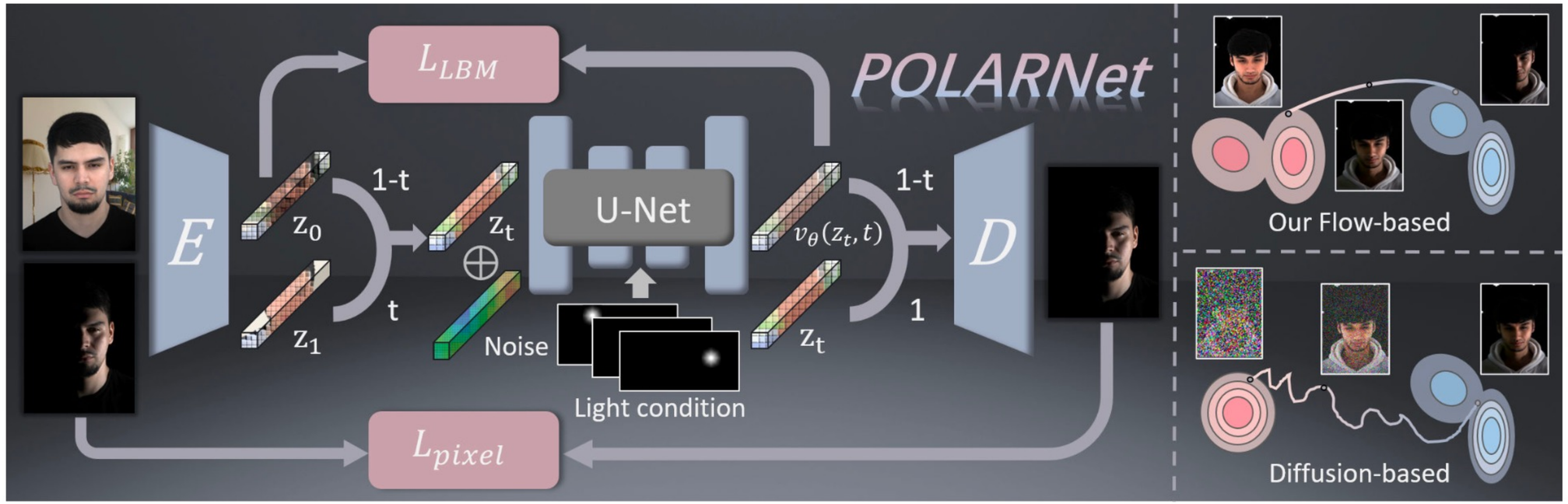




ByteDance字节跳动

Chen, Zhuo, et al. "POLAR: A Portrait OLAT Dataset and Generative Framework for Illumination-Aware Face Modeling." CVPR 2026.

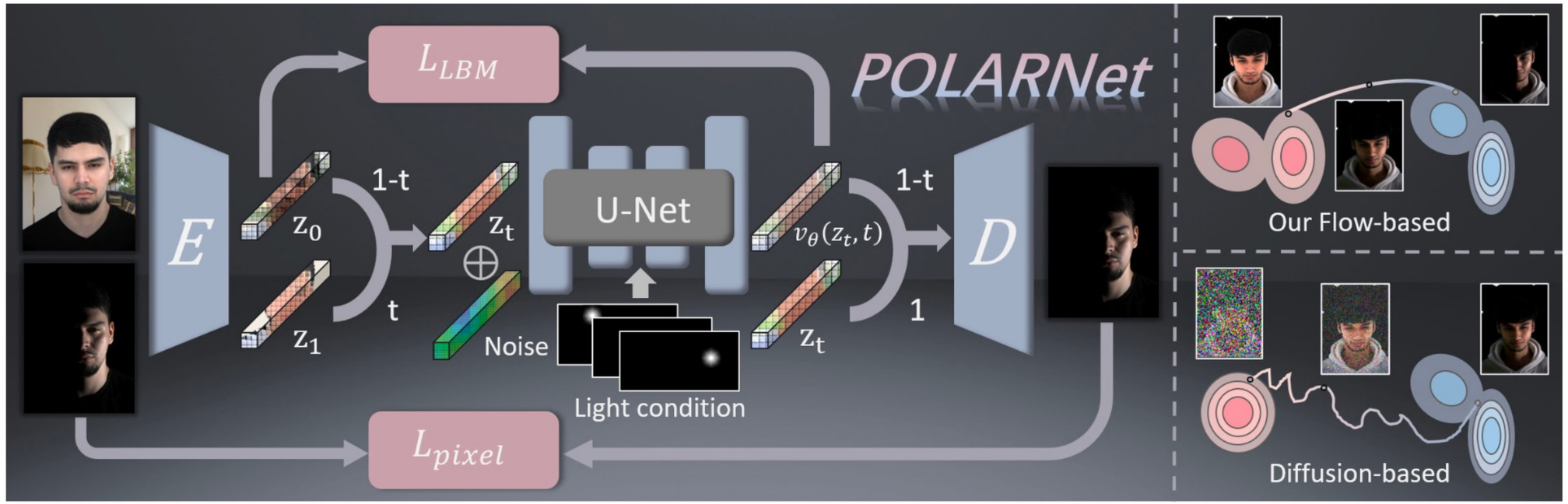
From Lighting Synthesis to Delighting Pre-processing



- **Flow-Based Synthesis:** Proposing POLARNet to leverage latent bridge matching for ensuring physically consistent and identity-preserving relighting.

Chen, Zhuo, et al. "POLAR: A Portrait OLAT Dataset and Generative Framework for Illumination-Aware Face Modeling." CVPR 2026.

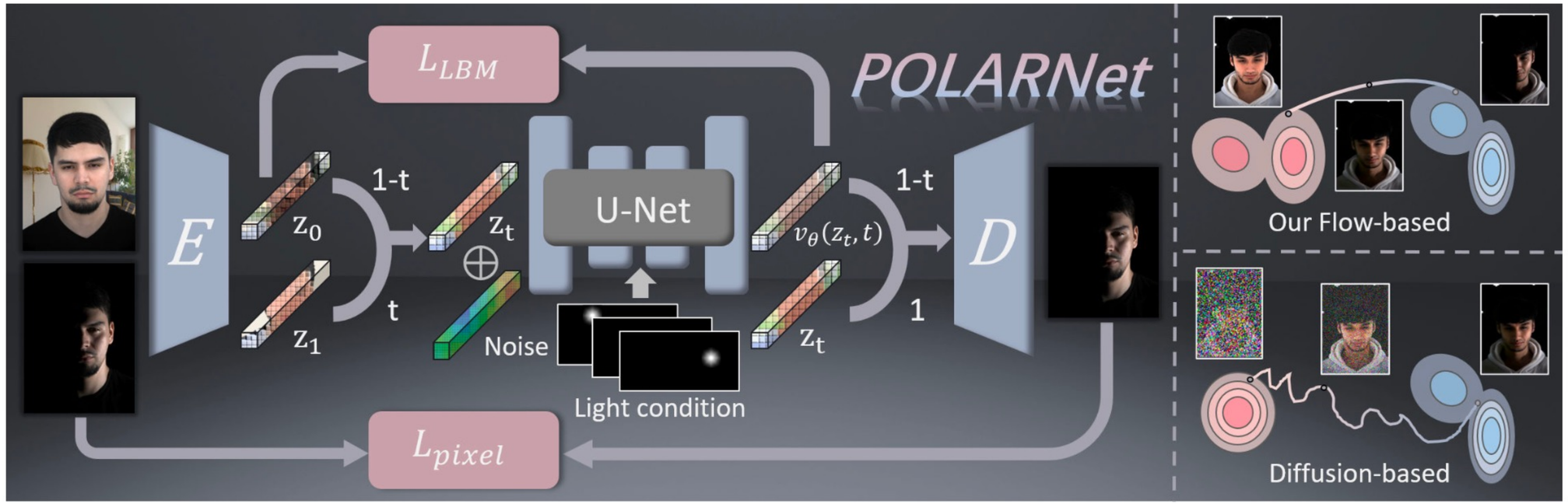
From Lighting Synthesis to Delighting Pre-processing



- **Flow-Based Synthesis:** Proposing POLARNet to leverage latent bridge matching for ensuring physically consistent and identity-preserving relighting.
- **Infinite Paired Data:** Using POLARNet to synthesize massive (Complex Lighting ↔ Uniform Light) pairs for arbitrary in-the-wild identities.

Chen, Zhuo, et al. "POLAR: A Portrait OLAT Dataset and Generative Framework for Illumination-Aware Face Modeling." CVPR 2026.

From Lighting Synthesis to Delighting Pre-processing

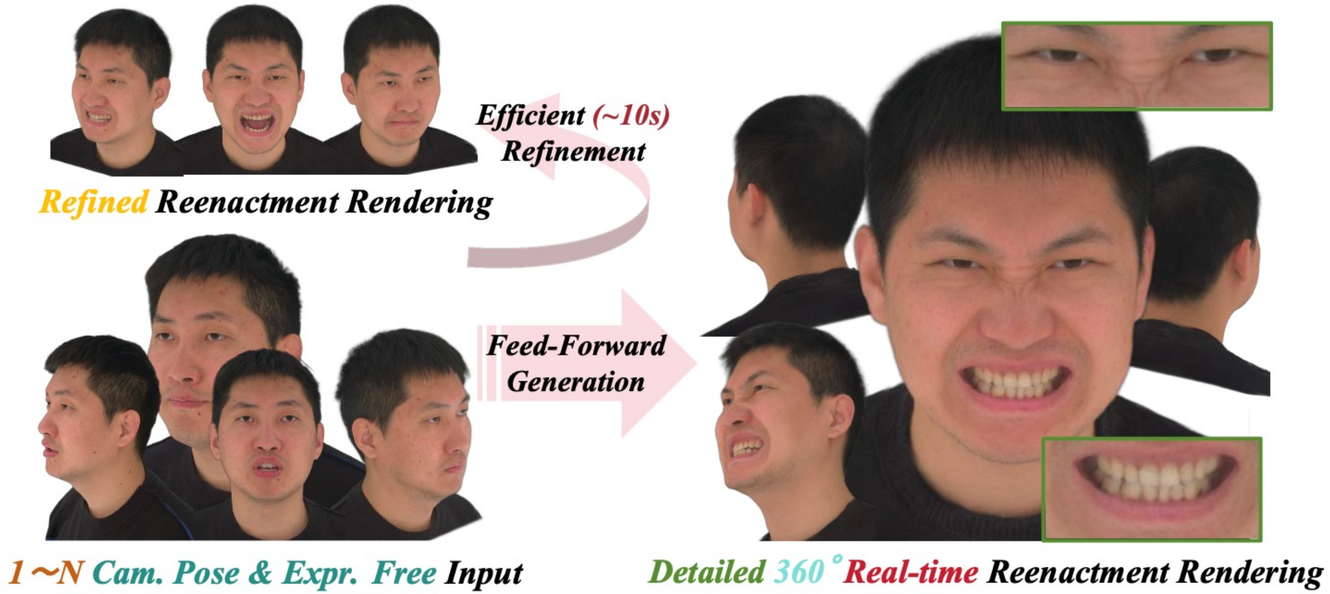


Delighting

- **Flow-Based Synthesis:** Proposing POLARNet to leverage latent bridge matching for ensuring physically consistent and identity-preserving relighting.
- **Infinite Paired Data:** Using POLARNet to synthesize massive (Complex Lighting ↔ Uniform Light) pairs for arbitrary in-the-wild identities.
- **Delighting Pre-processing:** Adapting POLARNet to normalize chaotic illumination into a uniform space, securing clean inputs for Avatar reconstruction.

Chen, Zhuo, et al. "POLAR: A Portrait OLAT Dataset and Generative Framework for Illumination-Aware Face Modeling." CVPR 2026.

FlexAvatar: The Philosophy of Ultimate Flexibility



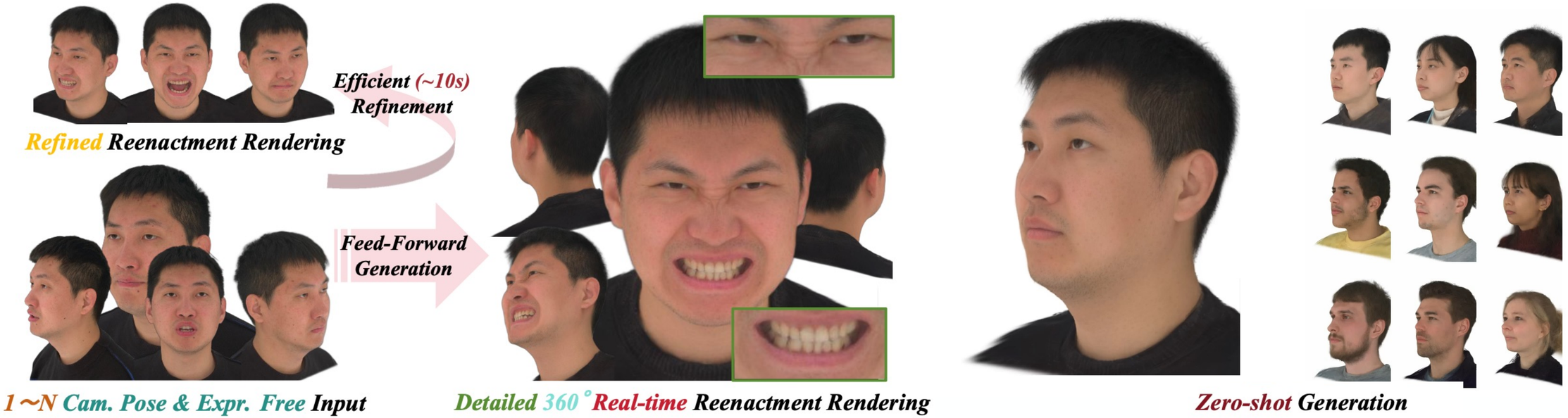
Why "Flex"? Breaking Input Barriers:

- ***Count-Free****: Reconstructs robustly from highly sparse inputs (1 to N).*
- ***Pose-Free****: Eliminates the need for precise camera initialization.*
- ***Expression-Free****: Handles random, chaotic, and unlabeled in-the-wild expressions.*

ByteDance字节跳动

Peng, Cheng, et al. "FlexAvatar: Flexible Large Reconstruction Model for Animatable Gaussian Head Avatars with Detailed Deformation." CVPR 2026.

FlexAvatar: The Philosophy of Ultimate Flexibility



Why "Flex"? Breaking Input Barriers:

- ***Count-Free****: Reconstructs robustly from highly sparse inputs (1 to N).*
- ***Pose-Free****: Eliminates the need for precise camera initialization.*
- ***Expression-Free****: Handles random, chaotic, and unlabeled in-the-wild expressions.*

The Output Promise:

- ***Feed-forward instant 360° generation with detailed zero-shot realism.***

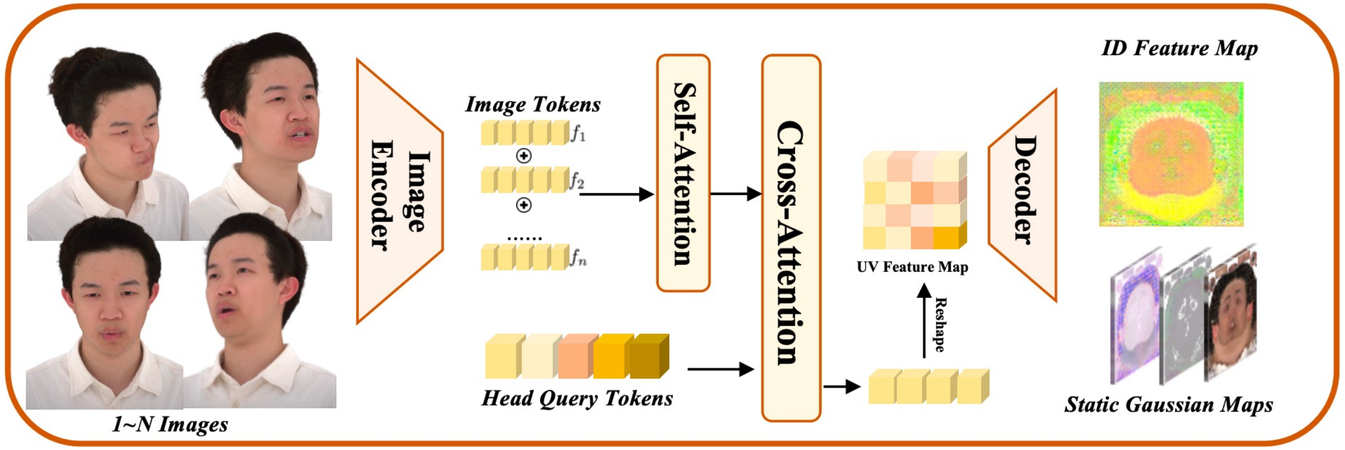
ByteDance字节跳动

Peng, Cheng, et al. "FlexAvatar: Flexible Large Reconstruction Model for Animatable Gaussian Head Avatars with Detailed Deformation." CVPR 2026.

Under the Hood: A Flexible Large Reconstruction Model

Fast Reconstruction: Flexible LRM

- **Mechanism:** Maps 1~N unposed inputs without expression annotation into unified **Structured Head Queries** via a robust transformer backbone.
- **Why Flexible:** Unifies arbitrary inputs into a structured latent space, completely bypassing rigid calibrations and expression priors.



Peng, Cheng, et al. "FlexAvatar: Flexible Large Reconstruction Model for Animatable Gaussian Head Avatars with Detailed Deformation." CVPR 2026.

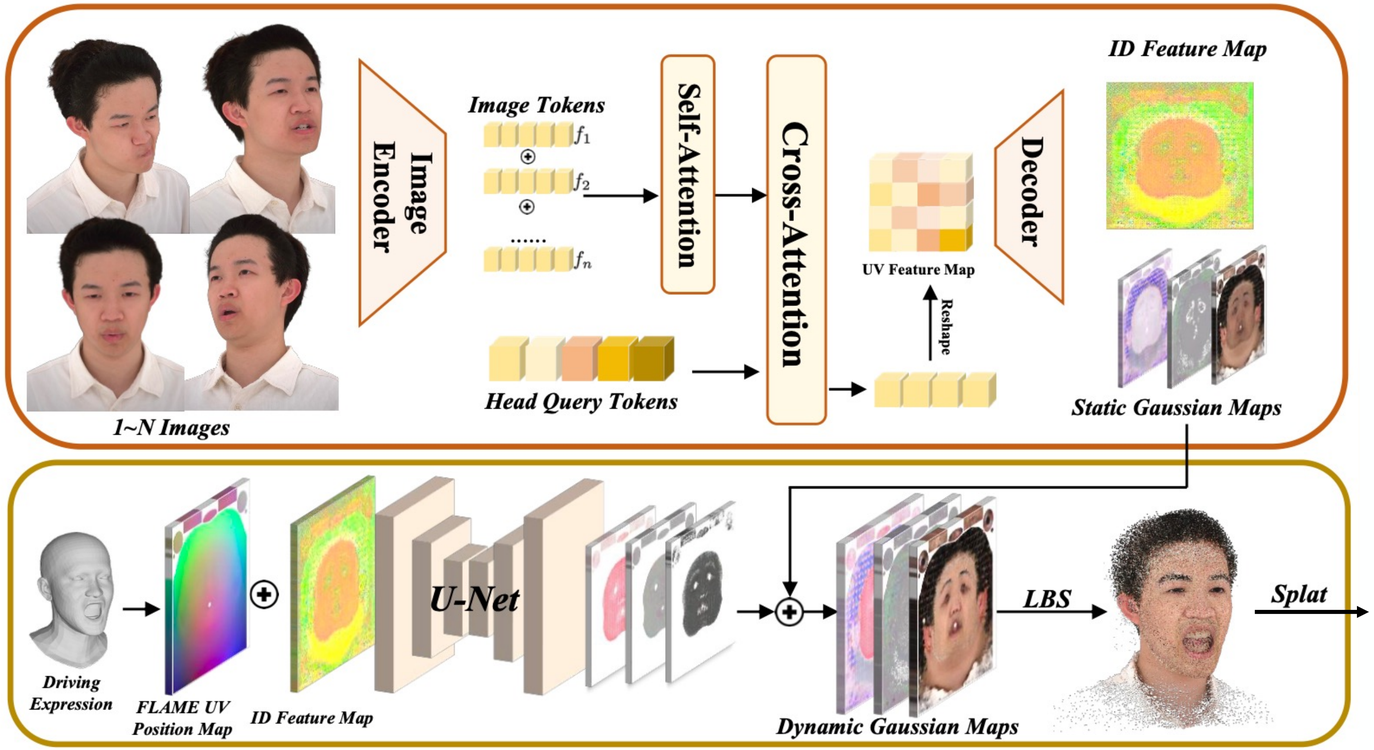
Under the Hood: A Flexible Large Reconstruction Model

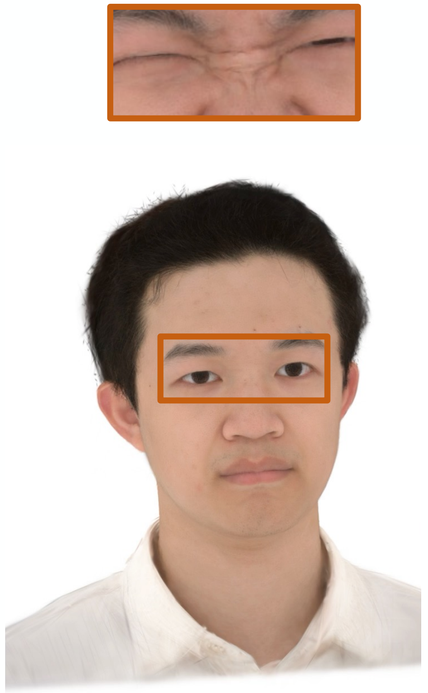
Fast Reconstruction: Flexible LRM

- **Mechanism:** Maps 1~N unposed inputs without expression annotation into unified **Structured Head Queries** via a robust transformer backbone.
- **Why Flexible:** Unifies arbitrary inputs into a structured latent space, completely bypassing rigid calibrations and expression priors.

Real-time Rendering: Fast U-Net

- **Mechanism:** Decodes 3D features directly in the 2D UV space via UNet.
- **Why Fast:** Hardware-friendly 2D CNNs bypass 3D bottlenecks to ensure real-time mobile inference.

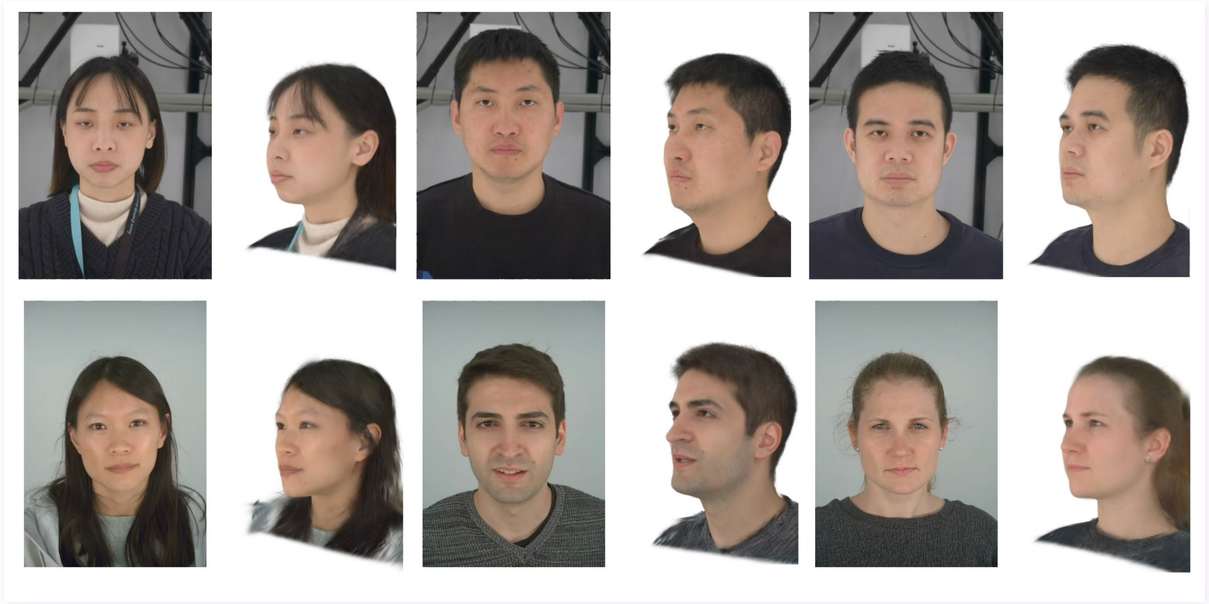




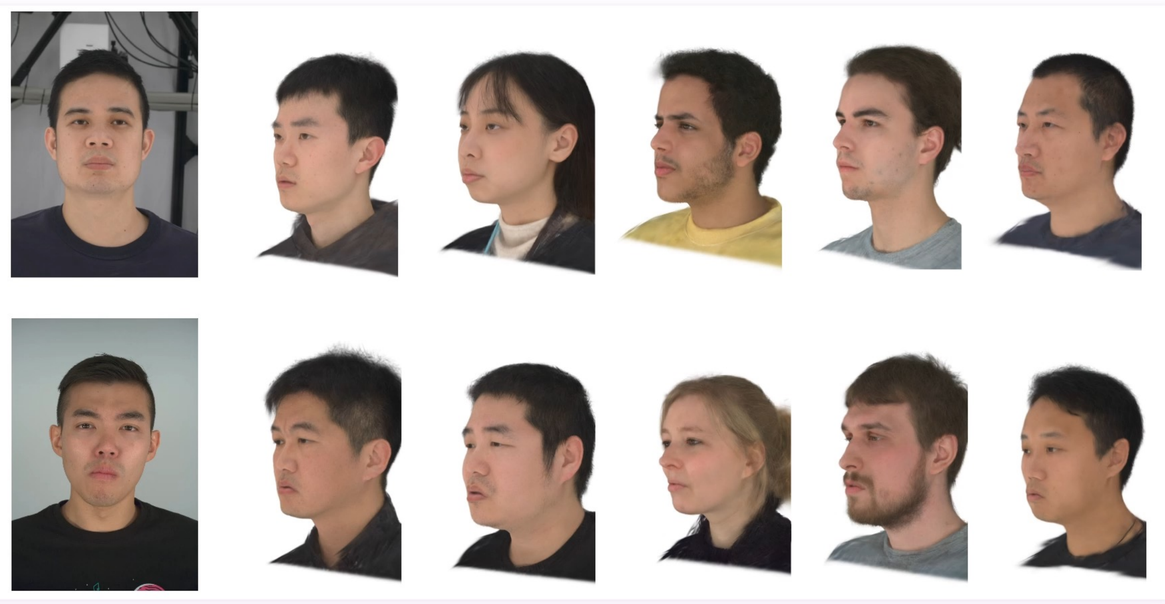
Peng, Cheng, et al. "FlexAvatar: Flexible Large Reconstruction Model for Animatable Gaussian Head Avatars with Detailed Deformation." CVPR 2026.

FlexAvatar: High-Fidelity Reenactment Results

Self-Reenactment



Cross-Reenactment



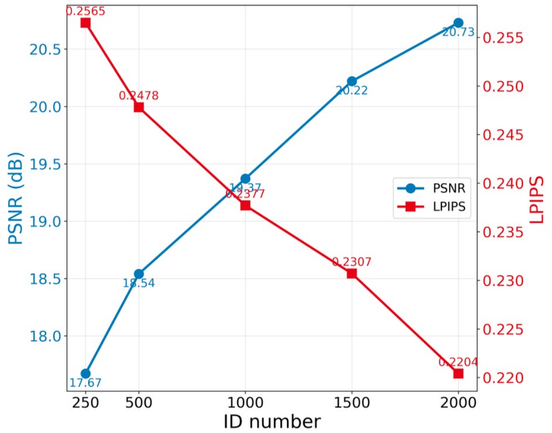
Deployable Pipeline: 📸 1~4 Casually Captured Inputs + ⚡ 10s Fast Refinement

ByteDance 字节跳动

Peng, Cheng, et al. "FlexAvatar: Flexible Large Reconstruction Model for Animatable Gaussian Head Avatars with Detailed Deformation." CVPR 2026.

Towards Better Avatars: Data, Model, and Beyond

Data: The 3D Scaling Law

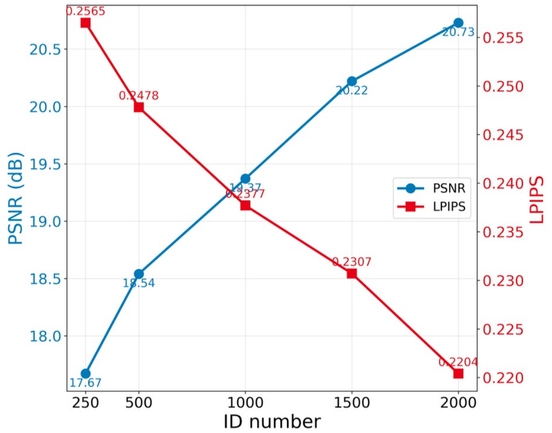


Continuous Scaling: *Training foundation models via 100K+ real-world identities.*

Chen, Zhuo, et al. "Relightable and Animatable Gaussian Head Avatar from Monocular Videos." TVCG 2026.

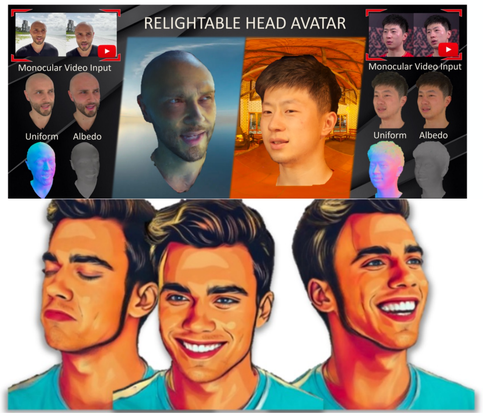
Towards Better Avatars: Data, Model, and Beyond

Data: The 3D Scaling Law



Continuous Scaling: *Training foundation models via 100K+ real-world identities.*

Capability: Ultimate Controllability

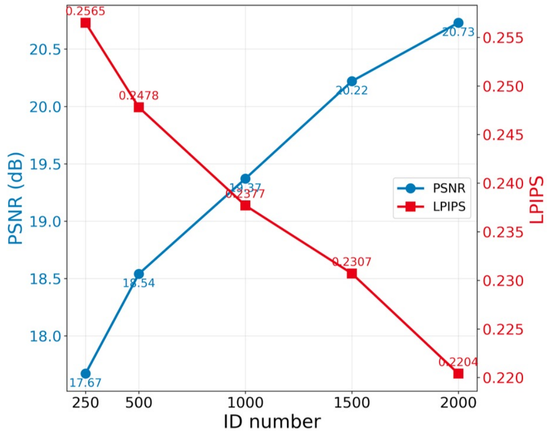


Relightable & Editable: *Enabling fine-grained, real-time control over relighting, makeup, and aging.*

Chen, Zhuo, et al. "Relightable and Animatable Gaussian Head Avatar from Monocular Videos." TVCG 2026.

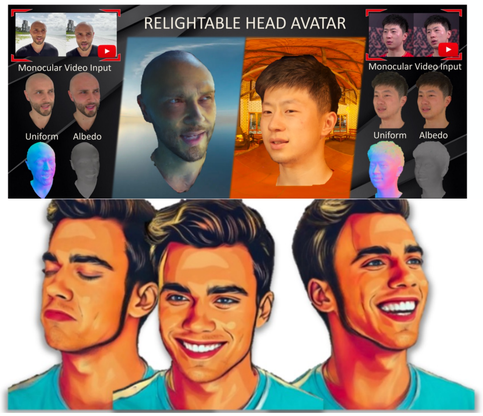
Towards Better Avatars: Data, Model, and Beyond

Data: The 3D Scaling Law



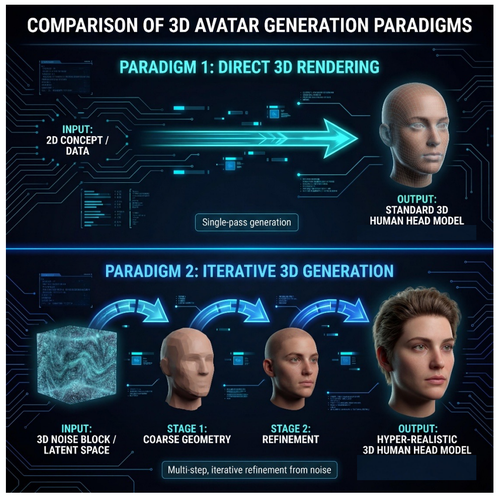
Continuous Scaling: *Training foundation models via 100K+ real-world identities.*

Capability: Ultimate Controllability

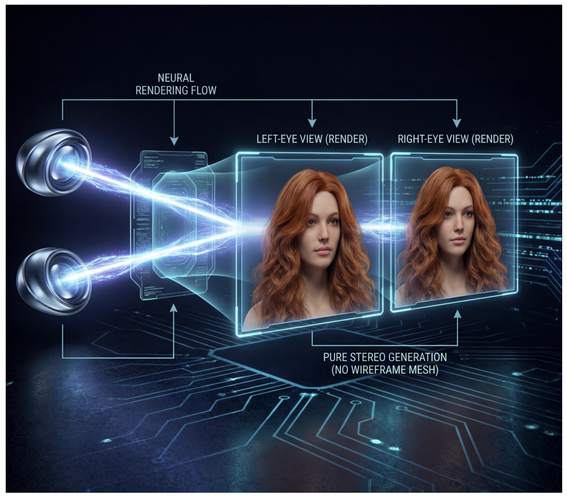


Relightable & Editable: *Enabling fine-grained, real-time control over relighting, makeup, and aging.*

Architecture: Paradigm Shift to Generative Models



LRM to DiT Paradigm: *Using multi-step diffusion for ultimate photorealism.*



Beyond Explicit 3D: *Eliminating 3D representations via pure stereo video pipelines.*

Chen, Zhuo, et al. "Relightable and Animatable Gaussian Head Avatar from Monocular Videos." TVCG 2026.

THANKS