Towards Deployable 3D Avatars: From Scalable Data Engine to Robust Modeling

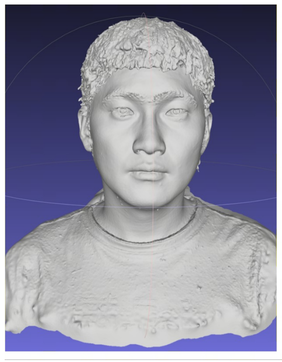
Su, Zhuo

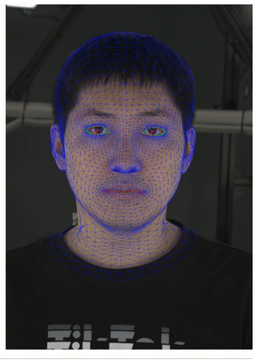
ByteDance 字节跳动

From Demo to Deployment: The Gap in Photorealistic 3D Avatars

Ideal Demo Setting:

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



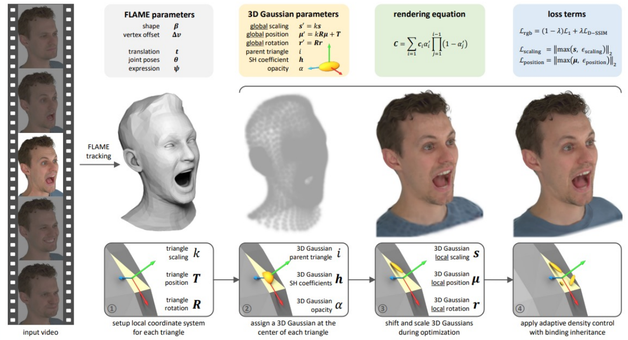


Real-World Deployment:

- **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.

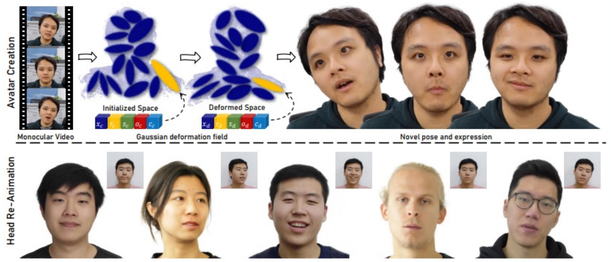
The Evolution & The Paradox of Avatar Modeling

Level 1: Multi-view Methods
Pros: Ultimate reconstruction quality.
Cons: Strict capture rigs, zero scalability.



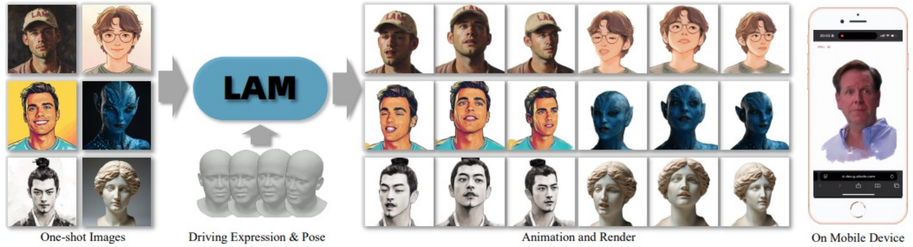
GaussianAvatars (Qian et al. / TUM & Toyota)

Level 2: Monocular Methods
Pros: No rigid setups needed.
Cons: Quality degrades with baked-in lighting.



Monocular Methods (Chen et al. / HIT & THU)

Level 3: Single/Few Image Methods
Pros: Low input barriers, highly scalable.
Cons: Degraded ID fidelity and quality ***—unless fueled by massive 3D training data.***



LAM (He et al. / Alibaba)

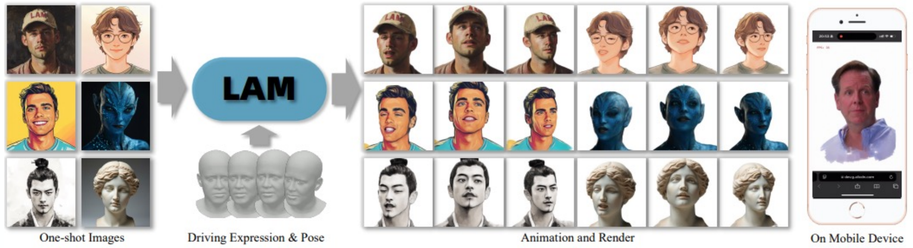
Paradox 1: The Input-to-Prior Trade-off
The less we demand from test-time inputs, the more desperately we depend on massive, high-quality training data.

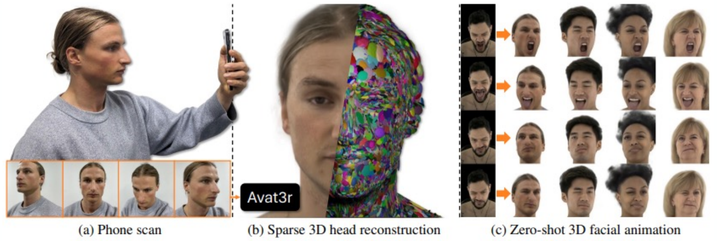
The Evolution & The Paradox of Avatar Modeling

2D-Only Methods
Pros: Good ID generalization.
Cons: Poor 3D consistency.

3D-Only Methods
Pros: Perfect 3D consistency.
Cons: Poor ID generalization.

Mixed Data Methods
Pros: Balances ID and geometry.
Cons: 2D data dominates with side-view collapse — **unless careful tuning of mixing ratios and training strategies (hybrid- to pre- & post-training).**





LAM (He et al. / Alibaba)

Avat3r (Tobias Kirschstein et al. / TUM & Meta)

FlexAvatar (Tobias Kirschstein et al. / TUM)

Paradox 2: The Quality-to-Diversity Dilemma
High-quality 3D data cannot scale, and scalable 2D wild data lacks 3D quality.

A Full-Stack Framework: Rethinking from Data to Model

To break the data paradox, we capture high-quality 3D/lighting data and unleash the power of generative foundation models to build a scalable data engine. Powered by this engine, we design a robust architecture to conquer unstructured real-world inputs.

Scalable Data Engine

Mission: Fueling robust generalization.

Action:

- **Acquisition:** Capturing scarce, high-quality 3D geometry and light-stage priors.
- **Expansion:** Leveraging generative models to synthesize infinite 3D and lighting data.

Feeds Infinite Pairs

Robust Modeling

Mission: Deploying in the chaos.

Action:

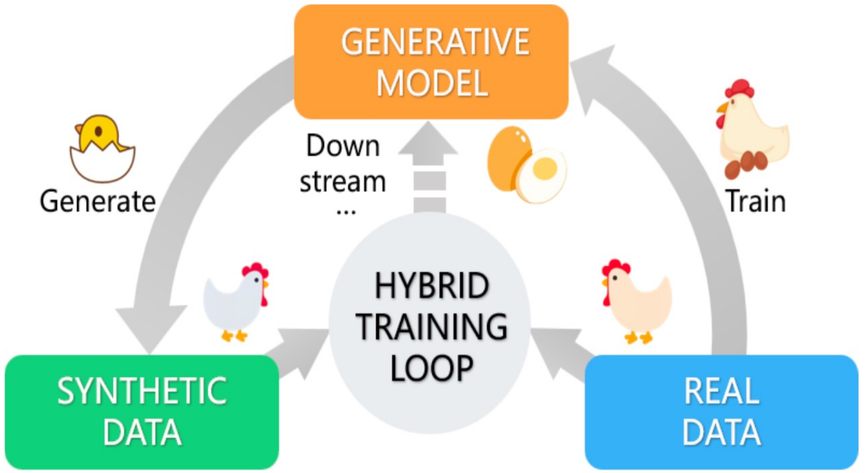
- **Pre-processing:** Normalizing complex inputs via image canonicalization and delighting.
- **Reconstruction:** Decoding unstructured sparse inputs via Flexible LRM.

Part 1: Scalable Data Engine

The Scalable Data Engine: Combining Real Data and Generative Loops

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

Physical-Captured Data：3D data (Multi-View Video) / Lighting data



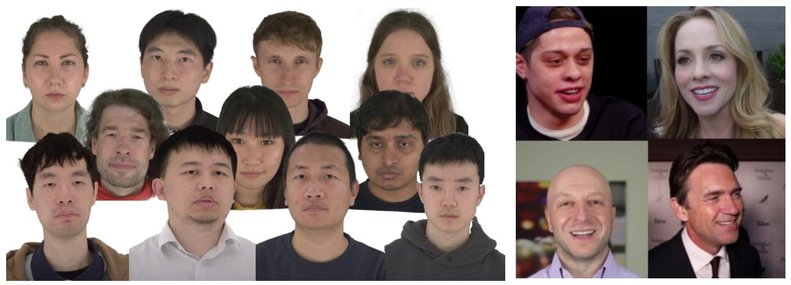
In-The-wild Data： Unknown Camera/Expression/Lighting

The Generative Loop: Real Data ↔ Synthetic Data

Quality **Diversity**

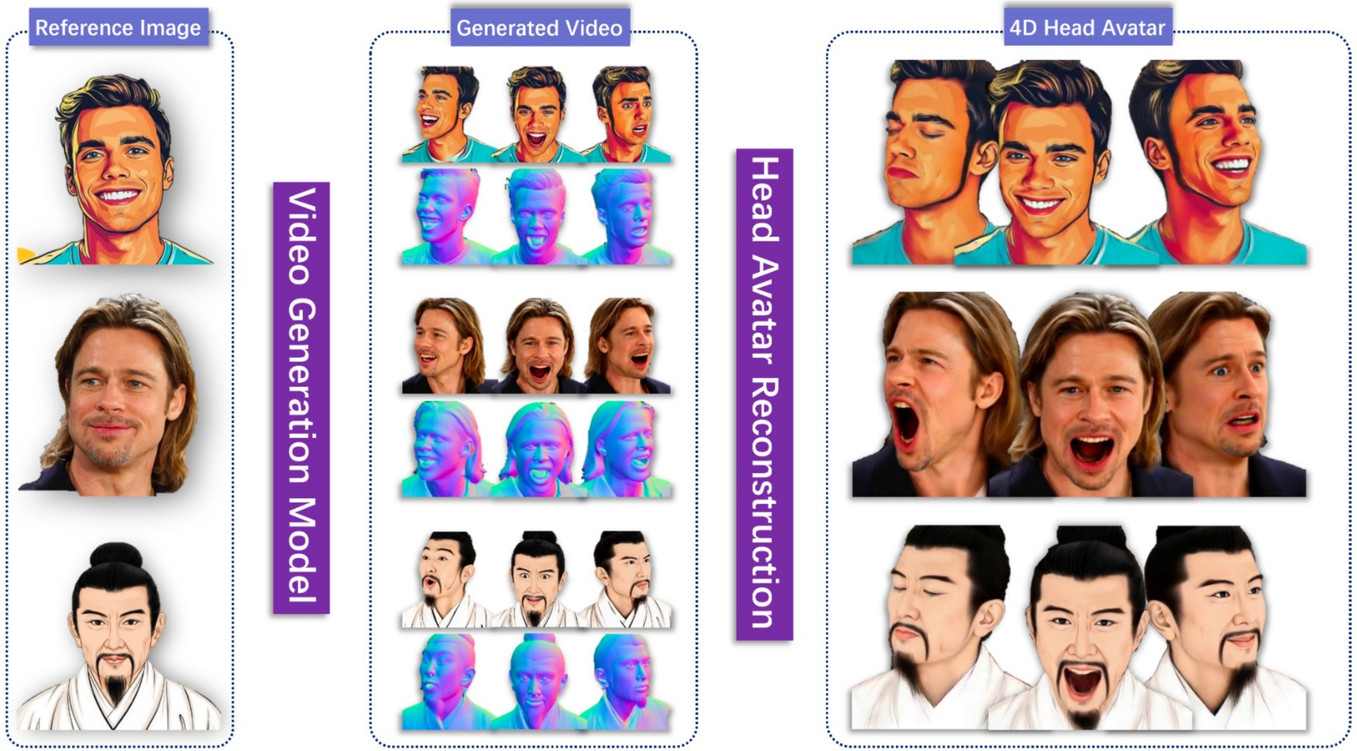
Scaling 3D Data via Generative Expansion

We use real-world 3D captures to fine-tune video generation models, creating massive synthetic assets.



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

Synthesized 4D Assets



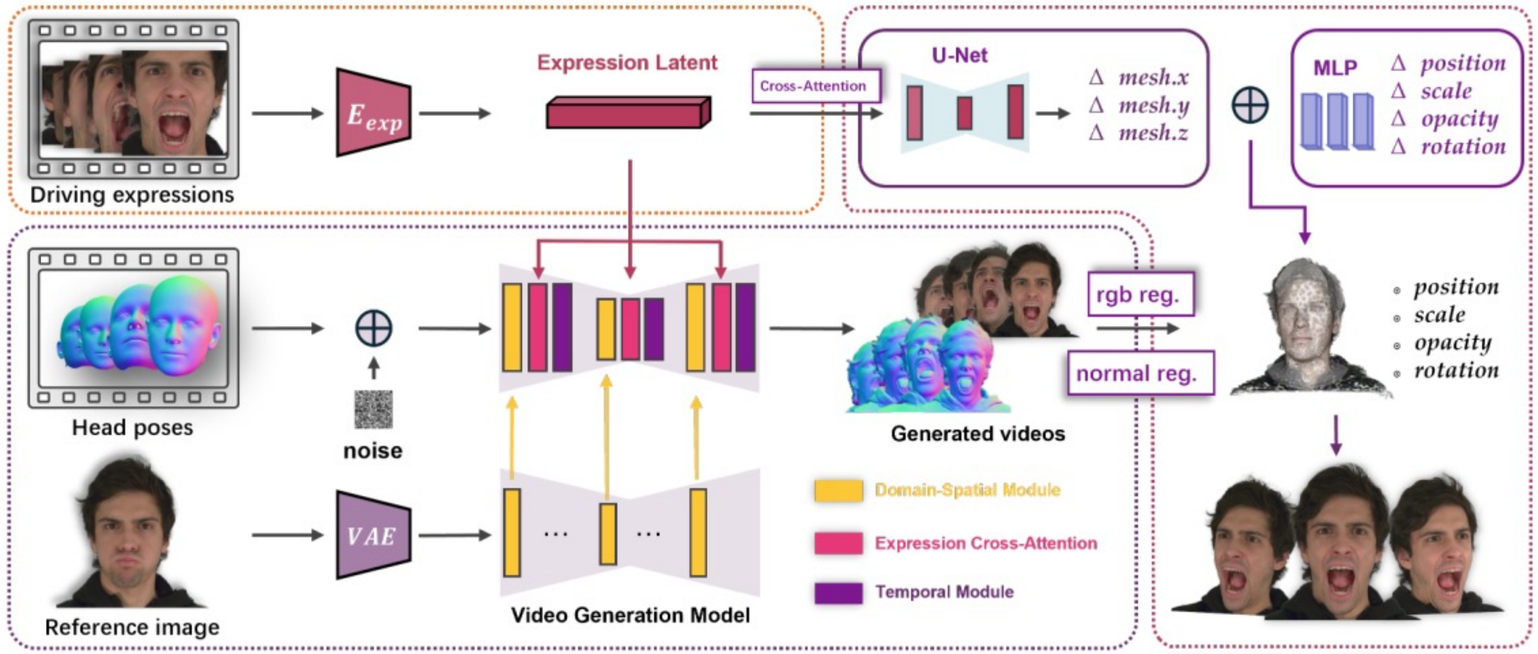
GeoDiff4D Engine: Jointly synthesizing video frames and normal for geometrically consistency

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

GeoDiff4D: Enforcing Geometry-Aware Consistency

The Pipeline for 3D Consistency:

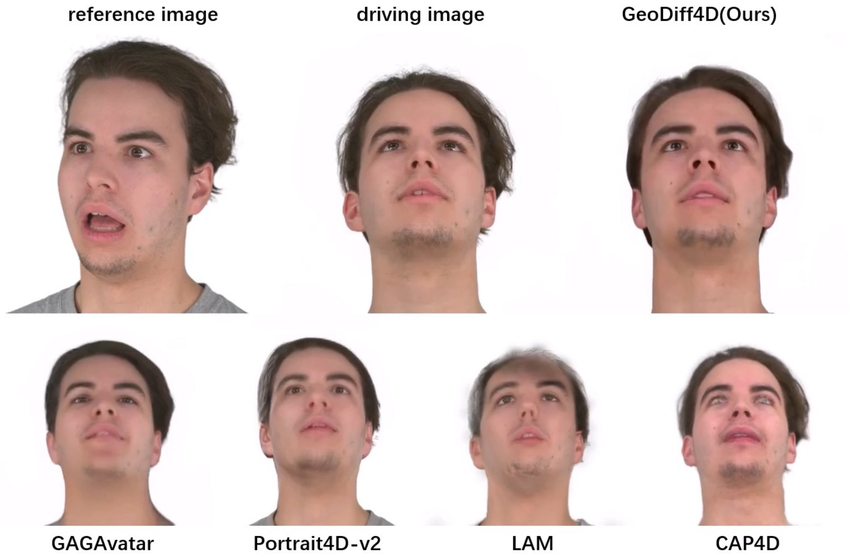
- **Decoupled Control:** Separate facial expressions from head poses to prevent spatial ambiguity.
- **Multi-View Synchronization:** Using cross-attention across views to share features during the diffusion process.
- **Joint RGB-Normal Synthesis (The Key):** Generate color and 3D normal at the same time, forcing 2D pixels to follow 3D rules.
- **Avatar Asset Creation:** Directly Distill the generated videos into 3D Gaussian Avatars.



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

Qualitative Results of GeoDiff4D

1. Self-Reenactment



3. Video Generation Model Ablation

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

Upgrading via Video Foundation Model

Upgrade the Data Engine's potential by replacing the base generation model with larger video foundation model, unlocking stronger dynamic priors while maintaining strict 3D consistency.

- **The Bottleneck:** Synthesis quality is strictly bounded by the original base model of GeoDiff4D (Stable Diffusion).
- **The Upgrade:** Replacing the backbone with Video Foundation Model (Wan2.2) to inherit its massive, unbounded video priors (natural motion, lighting).
- **The Constraint:** Applying the fine-tuning based on 3D captures to ensure strict 3D consistency.

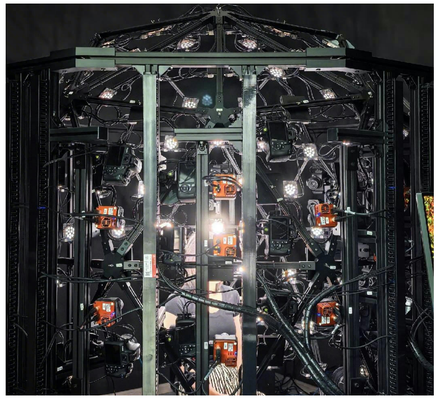
Breaking the data ceiling: generating realistic and annotated 4D assets

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

The Lighting Bottleneck

What is OLAT?

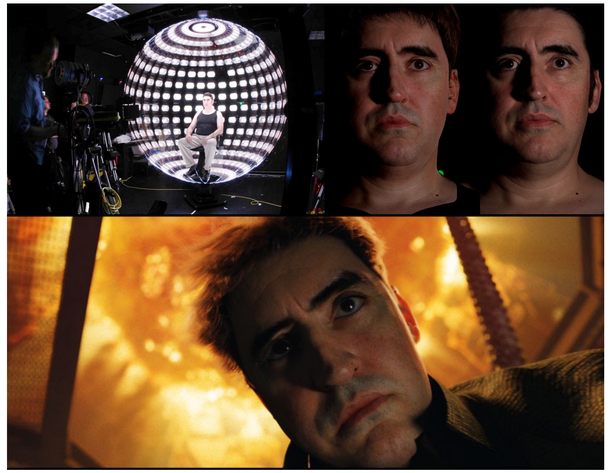
- OLAT stands for **One-Light-at-a-Time.**
- In a light stage, the subject is captured multiple times, with only one light turned on in each image.
- Each image records the **facial response to a single lighting direction.** These per-light responses can be combined to **synthesize portraits under new illumination conditions.**



Physical Capture: Capturing high-fidelity OLAT data in the lab.

Why Face OLAT data Matters?

- OLAT data is widely used for digital humans, relighting, and film production.
- OLAT data can be used for training Face De-lighting model, which can help normalize illumination chaos



Light-stage OLAT data enables realistic facial lighting effects in film production.

The Lighting Bottleneck

Meta dataset

Netflix dataset

Nvidia dataset

Disney dataset

High acquisition cost

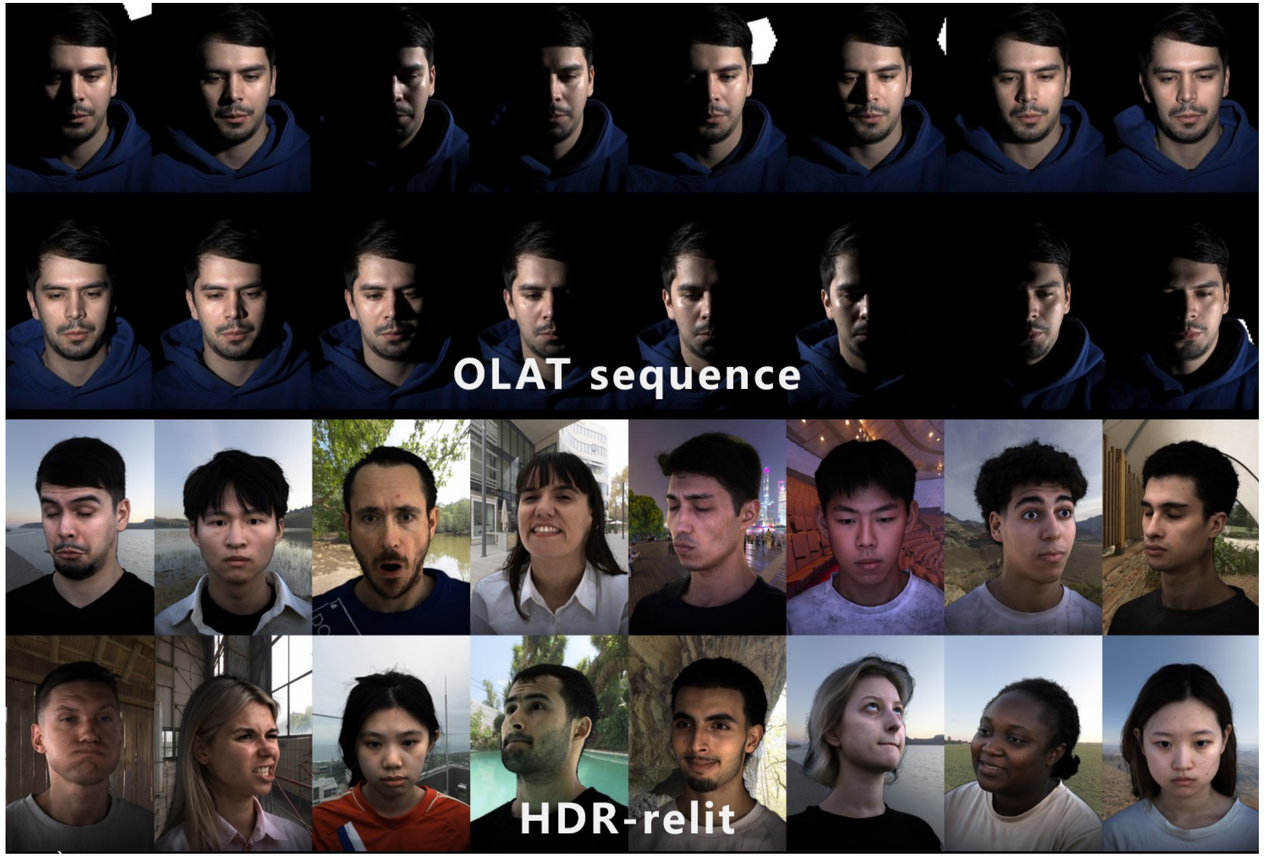
Limited public data

Build a high-quality open dataset

Generative model

Lower the cost of acquiring OLAT data

Breaking the Lighting Bottleneck via Physical Capture



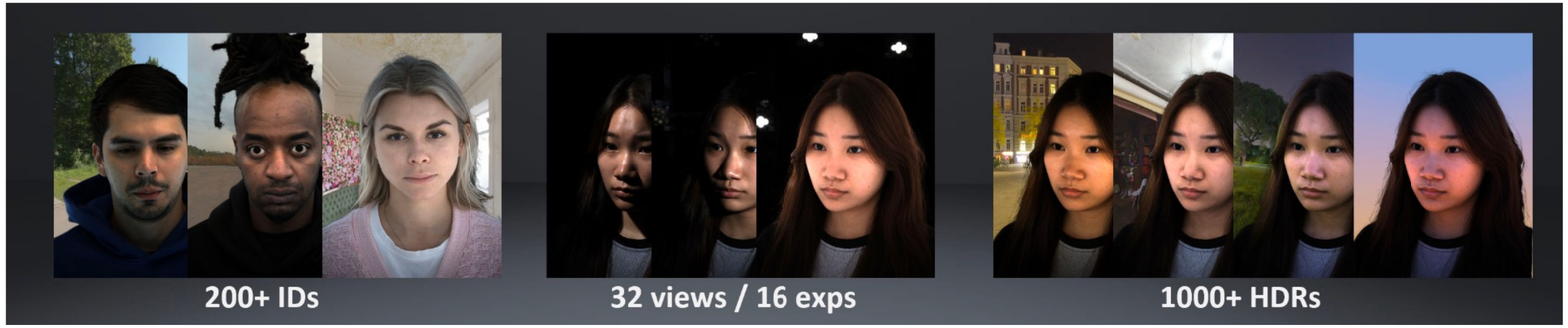
POLAR Dataset

- 220 subjects.
- 32 views
- 16 expressions
- 1000+ HDRs
- 28.8M frames
- 4K resolution
- OLAT + HDR-relit
- Open access

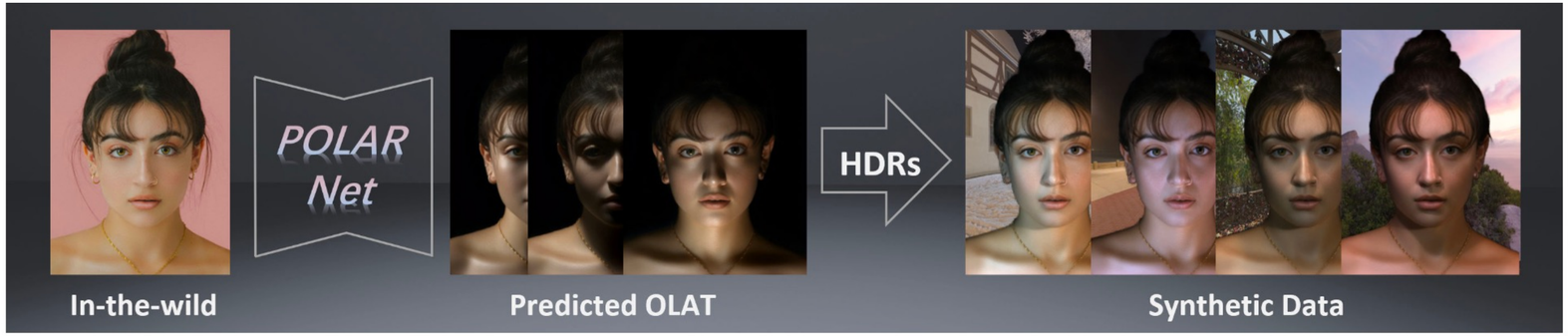
Chen, Zhuo, et al. "POLAR: A Portrait OLAT Dataset and Generative Framework for Illumination-Aware Face Modeling." CVPR 2026.

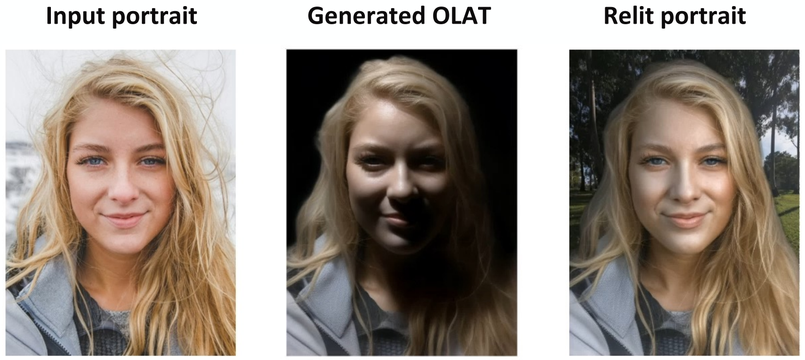
Lighting Data: From Physical Capture to Generative Expansion

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.





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

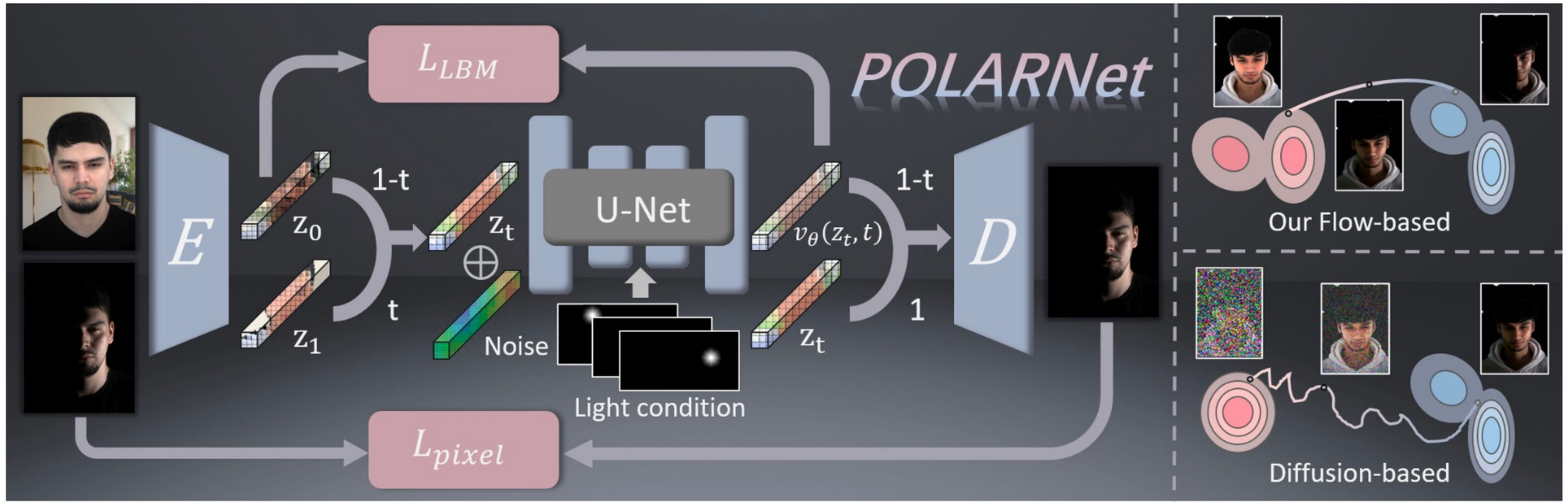
PolarNet: Scaling Lighting Data via Generative Expansion

Training: Learning the Path

- **Shared Space:** VAE maps inputs into a unified space.
- **Middle Step:** Mix states with noise.
- **Velocity Guide:** U-Net predicts the exact path to the target light.

Inference: One-Step Generation

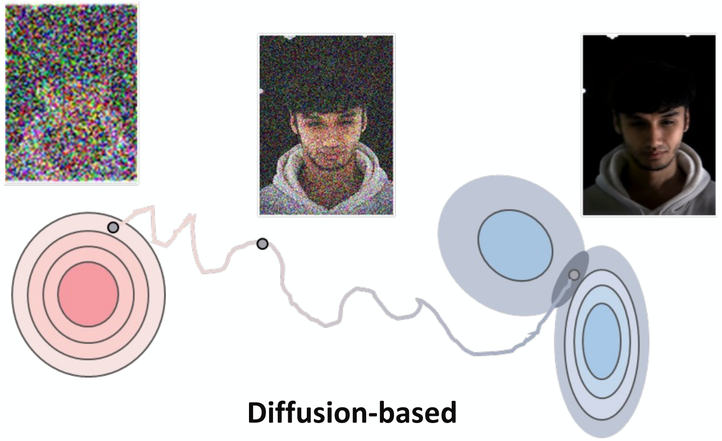
- **Follow the Path:** Model traces the learned trajectory.
- **Result:** Generate OLAT images in just 1 step.



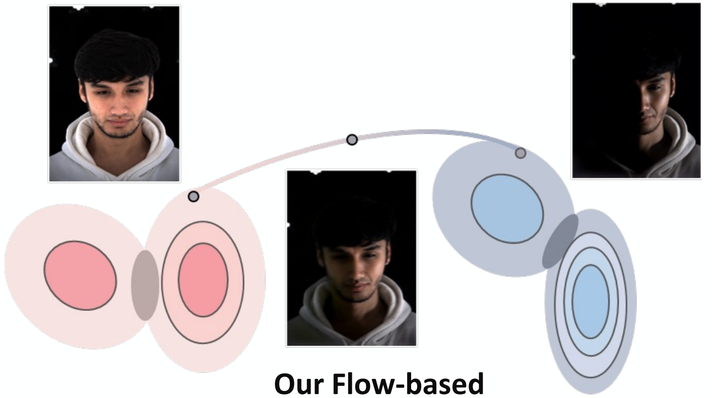
Chen, Zhuo, et al. "POLAR: A Portrait OLAT Dataset and Generative Framework for Illumination-Aware Face Modeling." CVPR 2026.

PolarNet: Scaling Lighting Data via Generative Expansion

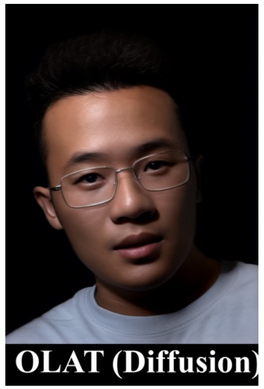
Flow-based approach directly follows a smooth illumination trajectory.



Diffusion-based



Our Flow-based



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

Comparison with Baseline Methods

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

Generated Relighting Data From In-the-wild inputs

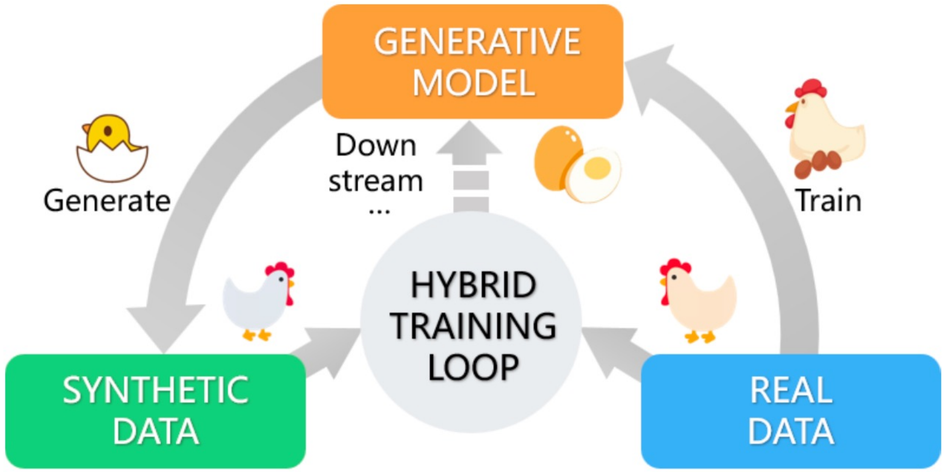
Peng, Cheng, et al. "FlexAvatar: Flexible Large Reconstruction Model for Animatable Gaussian Head Avatars with Detailed Deformation." CVPR 2026.

Part 2: Robust Modeling

Entering the Chaos: The Burden on the Model

Our Battlefield - Real-World Inputs

Our Fuel - Data Engine



The Modeling Checklist (Surviving the Chaos):

- ☐ **Illumination Chaos:** Must handle complex lighting and occlusion.
- ☐ **Input Chaos:** Must handle unstructured, unposed, casually-captured images.
- ☐ **Computational Constraint:** Must achieve real-time driving and rendering on edge devices.

Pre-processing: Normalizing Wild Inputs

Bridging the domain gap for extreme wild images via a two-step pre-processing: geometric canonicalization and generative delighting.

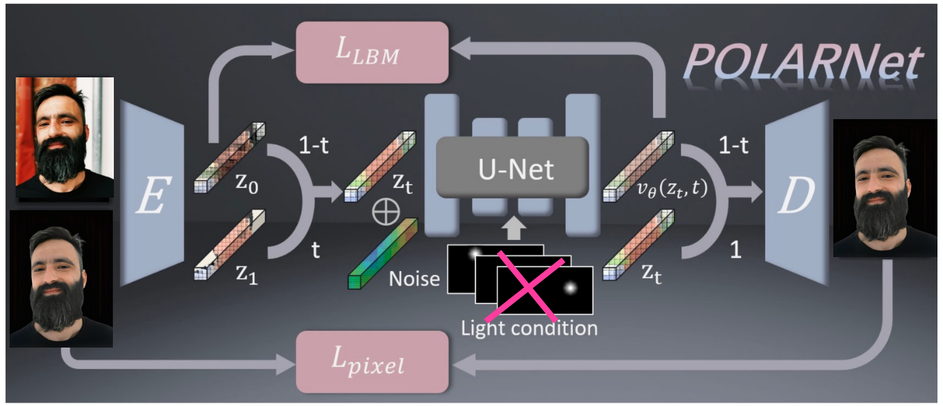


Image Delighting

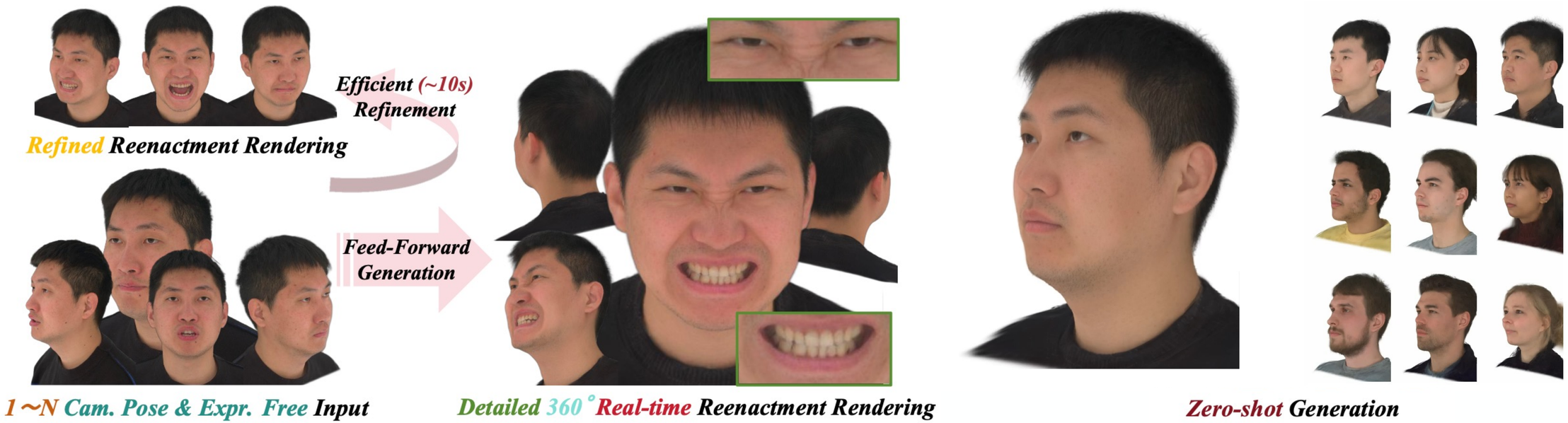
- **Method:** Condition-Free POLARNet (fueled by the lighting Data Engine).
- **Function:** Map complex lighting into a uniform space to secure a clean input for reconstruction.

Image Canonicalization (Optional)

- **Method:** Image Editing Models, such as Seedream.
- **Function:** Remove extreme occlusions and frontalize head poses (for single input) to align with the training domain.

Backbone: Decoding Unstructured Inputs via FlexAvatar

We propose FlexAvatar, A Flexible Large Reconstruction Model for Animatable Gaussian Head Avatars with Detailed Deformation



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:

- ***Feed-forward full-3D generation with detailed zero-shot realism.***

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

Flexible LRM: Decoding Unstructured Inputs

Step 1: Feature Extraction (Image Encoder)

Extracts dense 2D features from the variable-length inputs.

Step 2: Unified Spatial Mapping (Transformer Backbone)

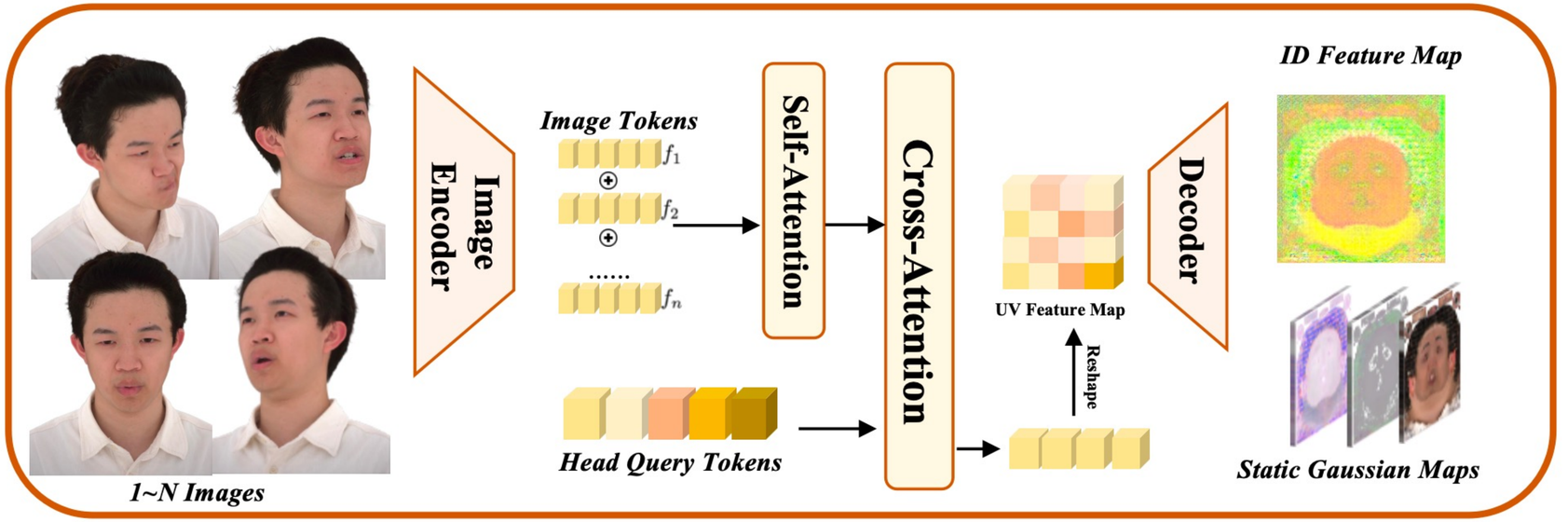
Global Context: Uses Self-Attention to aggregate the variable-length features.

Canonical Anchoring: Employs learnable Head Query Tokens in a canonical space.

Alignment: Uses Cross-Attention to map unstructured features onto a structured UV dimension.

Step 3: 3D GS Reconstruction (UV Decoder)

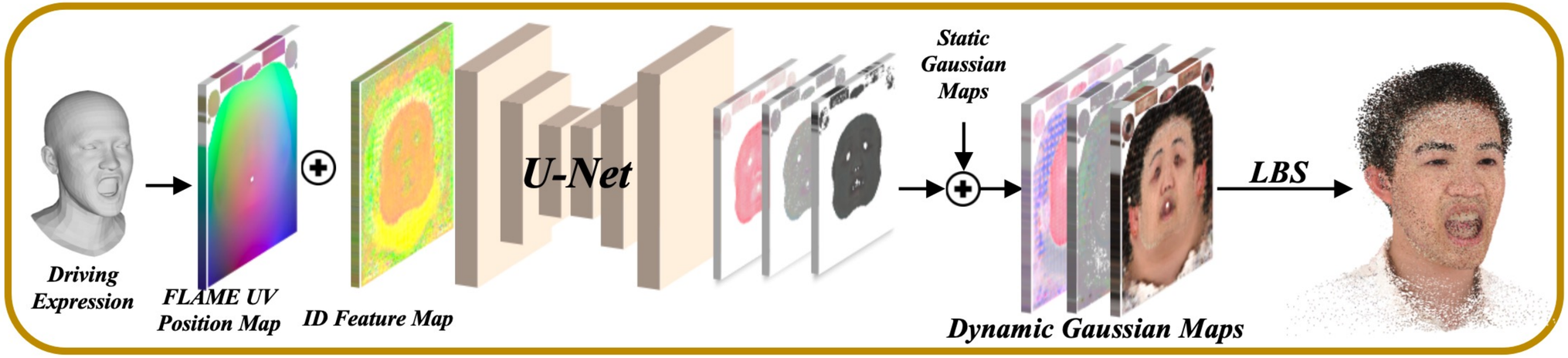
Decodes the structured UV features directly into static 3D Gaussian Splatting parameters.



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

Efficient Driving 3DGS via Lightweight UNet

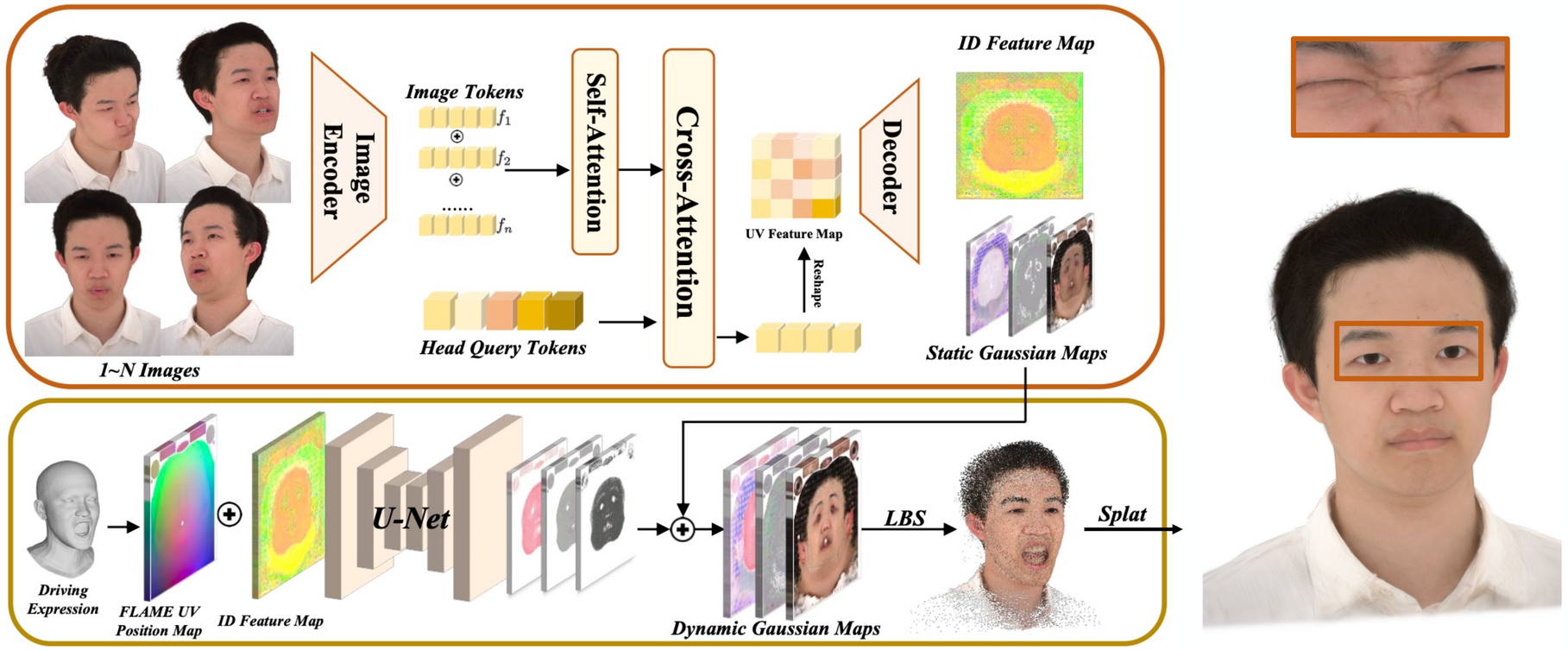
Driving Signal: Position Map (3DMM projected into UV space).
• **Mechanism:** Combine Position Map with identity features inside the UV space.
• **Impact:** Precisely guides the model to learn expression-driven local deformations.
Driving Module: Lightweight 2D U-Net
• **Vs. MLPs:** U-Net uses convolutions to capture the **local neighborhood** in UV space, perfectly reconstructing micro-dynamics (deep wrinkles).
• **Vs. Cross-Attention:** Drops heavy operators to achieve real-time animation.



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

FlexAvatar: The Complete Pipeline

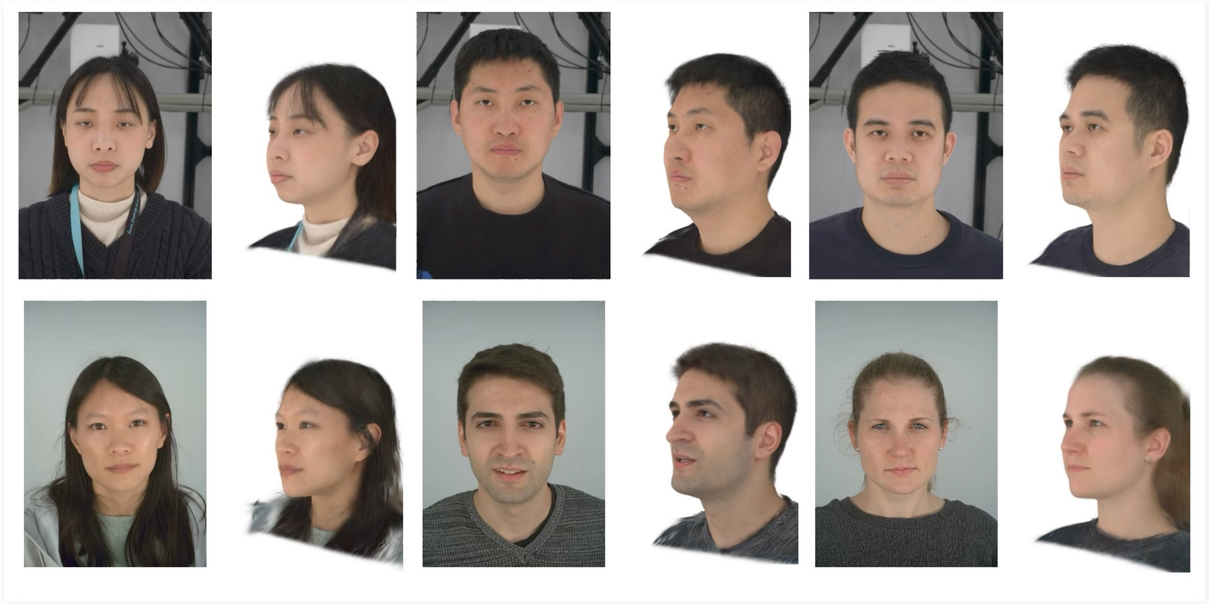
By mapping unstructured inputs into a structured UV space for fast cloud reconstruction, and driving it via a 2D U-Net for real-time mobile inference, we unify robust generalization with efficiency.



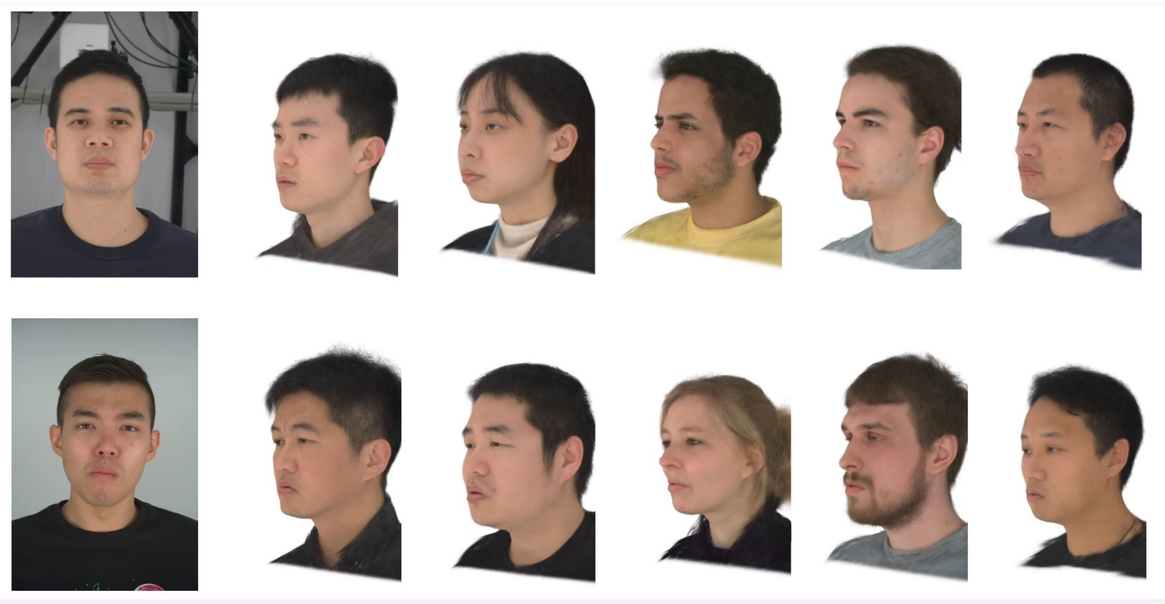
Peng, Cheng, et al. "FlexAvatar: Flexible Large Reconstruction Model for Animatable Gaussian Head Avatars with Detailed Deformation." CVPR 2026.

Qualitative Results of FlexAvatar

Self-Reenactment



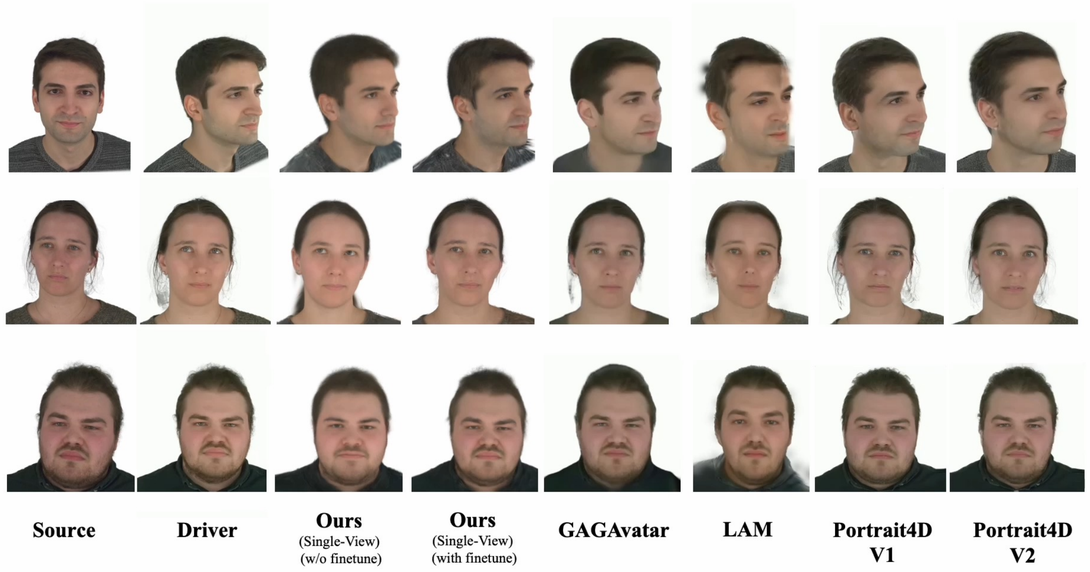
Cross-Reenactment



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

Comparison with Baseline Methods

Self-Reenactment



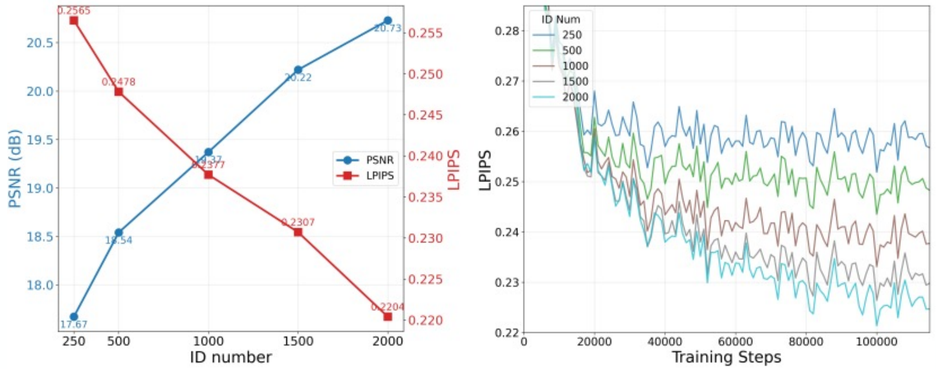
Method	Self Reenactment						Cross Reenactment		
	PSNR ↑	SSIM ↑	LPIPS ↓	CSIM ↑	AKD ↓	AED ↓	CSIM ↑	AKD ↓	AED ↓
LAM	17.8277	0.8031	0.2730	0.8213	5.7647	2.8077	0.8278	8.3807	4.8838
Portrait4D-v1	19.3746	0.8121	0.2410	0.8390	4.6719	2.3877	0.8399	8.0663	4.1993
Portrait4D-v2	19.6279	0.8184	0.2360	0.8385	4.6910	2.4325	0.8389	8.0622	4.2359
GAGAvatar	19.1667	0.8283	0.2567	0.8474	3.8730	2.1175	0.8479	8.2597	4.0454
Ours(feed-forward)	21.1516	0.8335	0.2193	0.8490	3.6518	2.0502	0.8501	7.8726	3.6415
Ours(with finetune)	22.6313	0.8491	0.1833	0.8532	3.4437	1.9127	0.8549	7.2426	3.5879

Table 1. **Quantitative comparisons on Nersemble test dataset.** In feedforward evaluations, our method surpasses other Gaussian-based single-image avatar generation approaches on every tested metric. Moreover, applying a subsequent finetuning stage yields additional gains in reconstruction fidelity and perceptual quality, further improving the realism and accuracy of generated outputs.

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

The Baseline FlexAvatar: Data Scarcity and Fast Refinement

Limited Training Data: At the time of paper submission, the baseline model was trained on only 2,000 real 3D scans.
Unconverged Potential: The zero-shot scaling curve clearly shows the model has not yet converged—it is still data-starved.



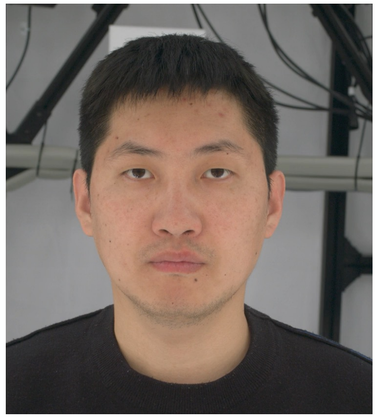
(a) Feed-forward test results on FaceCap.

(b) Model convergence speed.

Optional Fast Refinement: To handle out-of-domain inputs and recover high-frequency facial details, we introduced a fast test-time optimization.

Ours Feed-forward

Ours Refinement



Reference Video

While fast refinement is an effective patch for limited data, it is not the fundamental solution. To achieve true zero-shot perfection, we must return to our Phase 1 Data Engine.

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

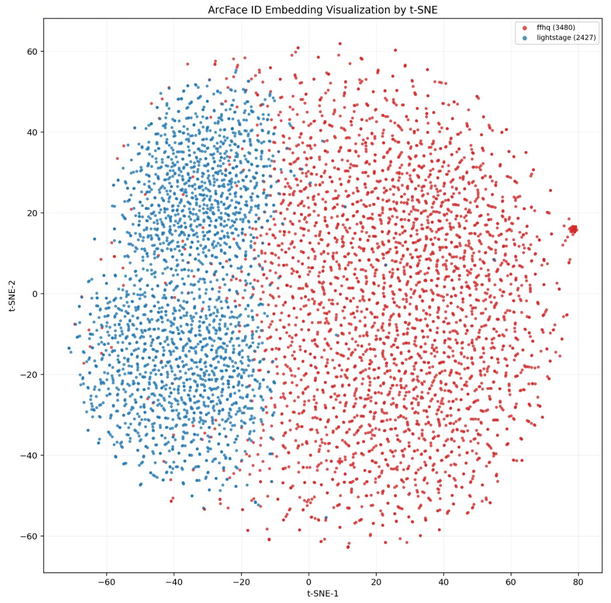
Achieving Robust Modeling in the Wild

For extreme in-the-wild scenarios, we tackling the domain gap through a dual strategy: standardizing extreme inputs by pre-processing and expanding the model's internal priors by augmenting training with generated data.

Image Pre-processing based on the lighting data engine: Rescuing severe out-of-domain images

Baseline FlexAvatar

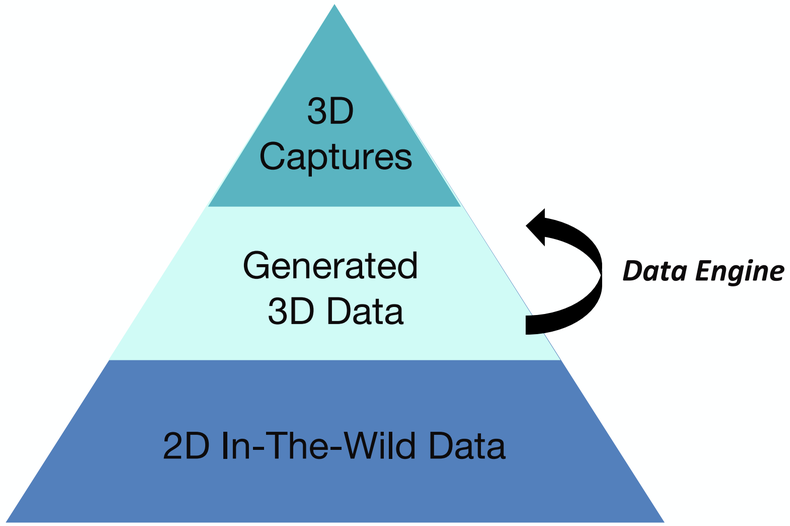
New FlexAvatar



Augmenting Training based on the 3D data engine: expanding the data distribution (t-SNE)

The Ultimate Strategy: Maximizing the Scaling Law

As the Data Engine scales, we must adapt our training strategy to respect the data hierarchy



Context: The Data Engine has just begun.

Massive data explosion causes hybrid-training to collapse

Phase 1: Pre-training (Bottom + Middle)

Data: 2D Wild + Synthetic 3D

Goal: Learn extreme diversity.

- **Phase 2: Post-training** (Top)

  Data: Real 3D Captures

  Goal: Lock in perfect 3D geometry.

Towards Better Avatars: Data, Model, and Beyond

1. DATA: The 3D Scaling Law

- ***Continuous Scaling:*** *Scaling to 10K+ real captures and expanding to 1M+ synthetic identities via our Data Engine.*
- ***The Goal:*** *Paving the way for true 3D avatar foundation models.*

2. CAPABILITY: Ultimate Controllability

- ***Relightable & Editable:*** *Developing inherently disentangled architectures.*
- ***The Goal:*** *Enabling fine-grained, real-time control over complex relighting, makeup, and aging.*

3. ARCHITECTURE: A Paradigm Shift to Generative Models

- ***LRM to DiT:*** *Transitioning from single-step regression to multi-step diffusion for better photorealism.*
- ***Beyond Explicit 3D:*** *Eliminating explicit 3D representations in favor of pure stereo-video generation pipelines.*

THANKS