FastAvatar: Towards Unified and Fast 3D Avatar Reconstruction with Large Gaussian Reconstruction Transformers

Anonymous authors
Paper under double-blind review

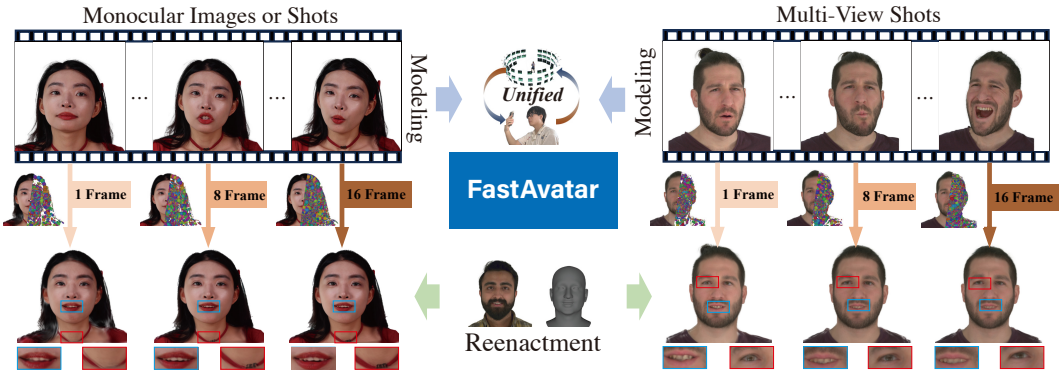


Figure 1: Unlike existing 3D Avatar methods that can only process fixed-length data, FastAvatar achieves incremental reconstruction. It can strike a good balance between modeling quality and inference speed based on available data volume, delivering high-quality models with sufficient data while providing viable reconstruction results at high speed even with limited data.

Abstract

Despite significant progress in 3D avatar reconstruction, it still faces challenges such as high time complexity, sensitivity to data quality, and low data utilization. We propose **FastAvatar**, a feedforward 3D avatar framework capable of flexibly leveraging diverse daily recordings (e.g., a single image, multi-view observations, or monocular video) to reconstruct a high-quality 3D Gaussian Splatting (3DGS) model within seconds, using only a single unified model. The core of FastAvatar is a Large Gaussian Reconstruction Transformer (LGRT) featuring three key designs: First, a 3DGS transformer aggregating multi-frame cues while injecting initial 3D prompt to predict the corresponding registered canonical 3DGS representations; Second, multi-granular guidance encoding (camera pose, expression coefficient, head pose) mitigating animation-induced misalignment for variable-length inputs; Third, incremental Gaussian aggregation via landmark tracking and sliced fusion losses. Integrating these features, FastAvatar enables incremental reconstruction, i.e., improving quality with more observations without wasting input data as in previous works. This yields a quality-speed-tunable paradigm for highly usable 3D avatar modeling. Extensive experiments show that FastAvatar has a higher quality and highly competitive speed compared to existing methods.

1 Introduction

Creating photorealistic 3D avatar reconstruction is one of the fundamental problems in computer vision and graphics. Contemporary methods Kirschstein et al. (2025); He et al. (2025); Chen et al. (2024b); Qian et al. (2024a); Pan et al. (2024); Wen et al. (2024); Qian et al. (2024b); Jiang et al. (2024); Hu et al. (2024); Qiu et al. (2025) for 3D avatars have made significant advancements in 3D representation and modeling quality. However, these approaches commonly suffer from drawbacks

such as data sensitivity (e.g., requiring richly expressive data), high time complexity, and low data utilization efficiency. These issues, pose severe challenges to the low-cost application of 3D avatars.

Three factors hinder existing 3D avatar methods from addressing the aforementioned challenges: 1) *Inability to Leverage Prior Knowledge*: Although contemporary 3D avatars have widely adopted efficient representations like 3DGS Kerbl et al. (2023), they still primarily rely on per-scene optimization. This approach fails to utilize "experience" from similar scenes, preventing the acquisition of good initial values to accelerate optimization. Furthermore, since all model information stems solely from the input observations, missing data cannot be reconstructed, resulting in a heavy dependence on complete 3D observations. Daily captures, however, often contain significant information gaps. 2) *Low Accuracy in Observation Alignment*: 3D avatar methods typically depend entirely on parametric proxy models (e.g., 3DMM/FLAME Blanz & Vetter (1999); Li et al. (2017)) for coarse view alignment. The precision of this alignment is critical for effective modeling; For instance, GaussianAvatars Qian et al. (2024a) even requires the proxy model to provide detailed meshes for hair. However, these parametric models are susceptible to limitations in representational capacity (e.g., blendshapes from 3D scan databases), lighting conditions, and data quality, often failing to produce highly accurate proxy 3D models. Using this proxy without refinement leads to poor robustness, hindering unified adaptation to diverse data sources (e.g., light fields, smartphones, DSLR cameras). 3) *Inadequate Handling of Variable-Length Data*: Optimization-based 3D avatar methods typically require input data of a minimum specific length (typically at least 30 seconds at 25fps). Insufficient data often leads to modeling failure, resulting in severely limited capability to process few-shot data (e.g., 1 frame, 4 frames). Meanwhile, recently proposed feedforward-style methods Kirschstein et al. (2025); He et al. (2025) are usually designed for fixed-length inputs for training convenience. For instance, LAM He et al. (2025) can only process single-frame input, and Avat3r Kirschstein et al. (2025) is fixed to handling exactly 4 frames. However, real-world data can consist of any arbitrary number of frames. The inability to process inputs of variable lengths will result in wasted valuable observation data, consequently limiting modeling quality.

To pursue data-efficient, high-quality, and fast 3D avatar reconstruction, we propose *FastAvatar*. It enables direct feedforward reconstruction of animatable avatars within seconds from arbitrary-length input frames and can incrementally leverage additional observation data. The core of FastAvatar is a **L**arge **G**aussian **R**econstruction **T**ransformer (LGRT). It can align and aggregate variable-length facial inputs based on head pose and camera pose, then generate high-quality Gaussian model groups using coarse 3D positional prompts. Finally, these groups can be flexibly fused into a single 3DGS avatar model according to quality requirements and computational resources. Notably, compared to LAM He et al. (2025) and Avat3r Kirschstein et al. (2025), FastAvatar handles variable-length observation data with greater model flexibility and higher data utilization efficiency. Unlike VGGT Wang et al. (2025a), FastAvatar is capable of directly generating 3DGS avatars and can achieve granular 3D model aggregation. Benefiting from the successful architecture of VGGT, LGRT is designed as a variant of the VGGT structure. We replace the unstable Dense Prediction Transformer (DPT) Ranftl et al. (2021) with an MLP that directly predicts canonical 3DGS models, while adopting 3D parametric model (e.g., FLAME) mesh vertices as positional prompts for the output. These improvements maximize adaptability for the prediction of the 3DGS avatar. Instead of relying solely on single camera pose encoding, 3D avatar reconstruction demands higher requirements for input data alignment. Therefore, we additionally incorporate expression coefficients and head pose as positional encoding for input observations, enabling more precise cross-frame information aggregation. Critically, we propose a landmark tracking loss and sliced fusion loss to efficiently supervise the model for enhanced aggregation accuracy while enabling incremental 3DGS models fusion. Integrating these key designs, our model pioneers incremental 3D avatar reconstruction, meaning it can continuously ingest new observational data to progressively refine modeling quality.

Extensive experiments demonstrate that our model achieves highly competitive 3D reconstruction quality compared to state-of-the-art methods. It uniquely accomplishes incremental 3D avatar reconstruction, currently unattainable by existing approaches, and holds promise for delivering favorable solutions in the quality-speed trade-off paradigm.

2 Related Work

2.1 3D-based Head Avatar Reconstruction

FLAME-based Li et al. (2017); Feng et al. (2021); Daněček et al. (2022); Ma et al. (2024); Cudeiro et al. (2019) techniques utilize a parametric model in head reconstruction, allowing for effective expression control but struggle to represent details (e.g., eyes, teeth) and limited to single-view. Since neural radiance fields have demonstrated strong ability to synthesis photo-realistic images, some method Zielonka et al. (2023); Shao et al. (2023); Athar et al. (2023); Müller et al. (2022) have adopted NeRF Mildenhall et al. (2021) for head reconstruction, which perform higher fidelity, particularly in modeling fine-scale details like hair. However, NeRF-based approaches Athar et al. (2022); Guo et al. (2021); Liu et al. (2022) often suffer from a significant issue with head rendering speed limitations and extensive training data. Recently, 3DGS Kerbl et al. (2023) has demonstrated superior performance surpassing NeRF in both novel view synthesis quality and rendering speed. Approaches Qian et al. (2024a); Chen et al. (2024b); Xu et al. (2024); Wang et al. (2025b) generate photorealistic head avatars that allow full control over expressions and poses using multi-view videos. Another line of research places explicit emphasis on identity preservation Gerogiannis et al. (2025); Zheng et al. (2025); Zielonka et al. (2025). Despite 3DGS's impressive performance, it requires multi-frame data for identity-specific training and lacks flexibility, necessitating separate models for single-view and multi-view scenarios. In contrast, our FastAvatar achieves ultra-fast 3D head avatar reconstruction with a unified model.

2.2 Feed-forward Reconstruction Model

Traditionally, 3D reconstruction and view synthesis, depending on optimization-based approaches such as Structure-from-Motion Schonberger & Frahm (2016) and Multi-View Stereo Schönberger et al. (2016), are often computationally intensive, slow to converge, and reliant on precisely calibrated dataset, limiting their applications in real-world scenarios. Recently, series of research Wang et al. (2024); Chen et al. (2024a); Liu et al. (2024); Ye et al. (2024); Hong et al. (2023); Charatan et al. (2024); Szymanowicz et al. (2024); Tang et al. (2024); Jin et al. (2024); Zhang et al. (2024a); Jiang et al. (2025) initiate a new research paradigm termed Feed-forward 3D reconstruction model. DUSt3R Wang et al. (2024) introduces a method for dense and unconstrained stereo 3D reconstruction, operating without prior camera calibration or viewpoint poses. VGGT Wang et al. (2025a) uses a large feed-forward transformer to effectively predict all key 3D attributes from a single image or multiple images. While feed-forward networks excel in generic 3D reconstruction, their application to 3D head avatar reconstruction is still nascent and warrants systematic exploration. LAM He et al. (2025) introduces a feed-forward framework that reconstructs an animatable gaussian head from a single image, allowing animation and rendering without additional post-processing. Avat3r Kirschstein et al. (2025) regresses animatable 3D head avatar from just a few input images, reducing compute requirements during inference. A key challenge in Feed-forward Head Reconstruction Model is the absence of a unified framework to handle diverse real-world inputs, including monocular videos, sparse multi-view captures. To address this gap, we propose a VGGT-style transformer architecture to jointly process different observation resource, achieving state-of-the-art performance.

3 Methodology

Daily observations are diverse and varied, such as single selfies, multi-angle selfies, video vlogs, etc. In summary, they are variable-length. Existing optimization-based 3D Avatar methods Kirschstein et al. (2025); Chen et al. (2024b); Qian et al. (2024a) cannot effectively handle overly short data (typically a single image). FastAvatar was specifically designed to address this scenario.

3.1 Problem Definition and Notation

FastAvatar $\mathcal{G}(\cdot)$ is a feed-forward avatar reconstruction framework designed to take any number of input observations and output a high-quality animatable 3DGS avatar:

$$\mathcal{A} \leftarrow \mathcal{G}(I, \pi, z_{exp}, z_{pose}), \tag{1}$$

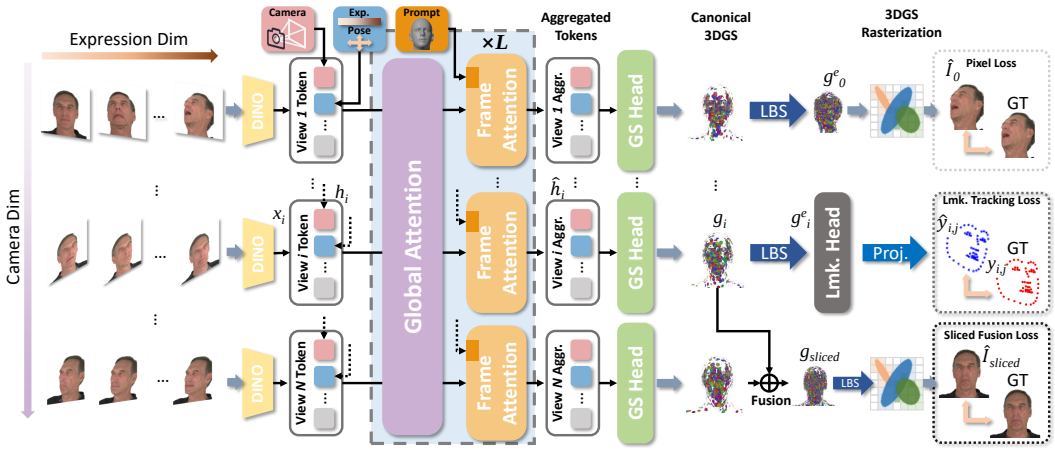


Figure 2: The core of FastAvatar is a Large Gaussian Reconstruction Transformer (LGRT), which can flexibly process input data with varying expressions, poses, and camera angles, aggregating them into a high-precision 3DGS avatar model. This capability is enabled by several key designs: the interleaving of global attention and frame attention to register complex input data while encoding 3D positional prompts; multi-granular positional information encoding; and the use of landmark tracking loss and sliced fusion loss, allowing the model to smoothly and incrementally fuse additional input data.

where $(I_i)_{i=1}^N$ denotes an unordered sequence of N RGB observations, with each $I_i \in \mathbb{R}^{3\times H\times W}$. N does not exceed the maximum GPU capacity, typically $1 \sim 16$ frames. The corresponding facial expression and head pose are represented by z_{exp} and z_{pose}, respectively. π denotes the camera parameters and poses. The output 3D Gaussian avatar representation g, including color $c \in \mathbb{R}^3$, opacity $o \in \mathbb{R}$, per-axis scale factors $s \in \mathbb{R}$, rotation $R \in \mathbb{R}^4$, importance score $m \in \mathbb{R}$, and points offest $O \in \mathbb{R}^3$, can be driven by any desired expression and pose from an arbitrary viewpoint using a differentiable rasterizer Ψ Kerbl et al. (2023).

3.2 Large Gaussian Reconstruction Transformer

The core of FastAvatar is a Large Gaussian Reconstruction Transformer (LGRT). The LGRT is redesigned specifically for 3D avatar tasks, which demand finer granularity compared to SLAM-based environmental reconstruction. Moreover, the human subjects captured for 3D avatars cannot remain perfectly static and often exhibit rich dynamic characteristics (i.e., expressions, poses, etc.). The LGRT comprises 6 stages: facial feature extraction, face encoding, face aggregation and registration, 3DGS attribute generation, canonical 3DGS model fusion, and 3DGS rasterization.

Face Encoding FastAvatar encodes each face observation I_i to a set of token x_i through DINOv2 Oquab et al. (2023). These tokens vary from head poses, facial expressions to camera poses, and undifferentiated aggregation would lead to over-smoothing and aliasing effects. Therefore, FastAvatar introduces three critical encodings to label distinct facial tokens, facilitating subsequent aggregation. This process can be formulated as:

$$h_i = \mathcal{U}\left(x_i, \mathrm{MLP}([\pi_i, z_i^{exp}, z_i^{pose}])\right), \tag{2}$$

where $\mathcal{U}(\cdot)$ denotes concatenation along the dimensional axis. h_t denotes the encoded face tokens. π_i, z_i^{pose}, and z_i^{exp} represent the camera pose, head pose, and expression coefficients of x_i respectively. These are processed through a lightweight MLP layer for feature alignment and dimensionality alignment. We obtain rough initial estimates of the three parameters through multi-view FLAME tracking.

Face Aggregation and Registration The core enabling component for constructing a dense 3D avatar from variable-length input data lies in the aggregation and registration of face tokens. The purpose of aggregation is to extract intra-token features while incorporating initialized 3D positional

prompts. These positional prompts provide the LGRT with initial 3DGS positions, thereby accelerating 3D reconstruction. As illustrated in Figure 2, aggregation is implemented through frame attention, composed of dual-stream DiT blocks Labs et al. (2025); Labs (2024), which aggregates intra-token information while fusing 3D positional prompts. Face token registration serves as the fundamental operation for fusing multiple inputs. In Figure 2, this is achieved via global attention, which aligns encoded face tokens to achieve 3D spatial registration and fusion. To enhance quality and accelerate convergence, we initialize our frame attention using weights from LAM's blocks. Global attention and frame attention are interleaved in a cascaded architecture, with a total of L pairs employed to process face tokens, ultimately yielding tokens suitable for generating the 3DGS representation $\{\hat{h}_0, \cdots, \hat{h}_N\}$.

Canonical 3DGS Model Fusion Following the aggregation and registration through global attention and frame attention, we obtain processed tokens $\hat{h}_i$ corresponding to each frame. These features are then fed into a GS Head (i.e., a two-layer MLP with shared weights across tokens) to predict the target 3DGS attributes g_i. The point cloud g_i^e derived by driving g_i through Linear Blend Skinning (LBS) expression deformation is then rendered via Gaussian splatting to obtain the reconstructed face $\hat{I}_i$. Our approach extends beyond g_i; we further aggregate all $g_0, \cdots, g_N$:

$$g_f = \mathcal{U}(g_1, g_2, \cdots, g_N). \tag{3}$$

The fused g_f integrates unique information from all perspectives (e.g., multi-view observations, diverse expressions), achieving optimal reconstruction quality. However, naive fusion would cause point cloud misalignment, ghosting artifacts, and color discrepancies. To address this, we introduce Landmark Tracking Loss and Sliced Fusion Loss to explicitly encourage proper alignment of Gaussian point clouds during aggregation and registration stages.

3DGS Pruning Although the 3DGS method achieves high-quality and real-time rendering, it often suffers from redundant memory consumption due to its explicit structure and tends to be more prone to overfitting because of the lack of smoothness bias in the neural network. This is especially problematic in our incremental reconstruction scenarios, where the number of GS points increases linearly with the number of input frames, leading to inefficient resource usage and limiting rendering speed. To address this, inspired by LP-3DGS Zhang et al. (2024b) and MaskGaussian Liu et al. (2025) we apply Gumbel-Softmax Jang et al. (2017) to sample one differentiable category, denoted as $\mathcal{M}_i \in \{0, 1\}$. Then we integrate masks directly within the rasterization framework, effectively decoupling Gaussian presence from attributes such as opacity and shape. This mechanism prunes over 50% of the GS primitives without degrading reconstruction quality, further improving rendering efficiency. To prune redundant 3D Gaussian primitives, we apply an L1 regularization term to the trainable mask, encouraging it to become sparse, formulated as $\mathcal{L}_{mask} = \frac{1}{N}\sum_{i=1}^{N}|m|$.

3.3 Training Strategy

Sliced Fusion Loss To enable the model to take advantage of the richer information provided by multiple inputs, we introduce *Sliced Fusion Loss*, allowing $\mathcal{G}$ to handle arbitrary numbers of input frames. Specifically, during training, we randomly sample one frame from the input to obtain a single frame-wise Gaussian representation g_i. In parallel, we randomly select N_{sliced} frames from the input, where N_{sliced} is less than the total number of input frames for memory efficiency, and fuse them to construct a multi-frame Gaussian representation g_{sliced}. Both g_i and g_{sliced} are rendered into RGB images using the camera parameters, expression coefficients, and head poses of all input and target frames, and the corresponding losses are computed.

$$\hat{I}_i = \Psi\left(g_i, \pi_i, z_i^{\text{exp}}, z_i^{\text{pose}}\right), \tag{4}$$

$$\hat{I}_{\text{sliced}} = \Psi\left(g_{\text{sliced}}, \pi_i, z_i^{\text{exp}}, z_i^{\text{pose}}\right). \tag{5}$$

The overall loss function consists of two components: one supervises the rendering quality of the constructed 3D Gaussian head, and the other supervises the combination of frame-wise Gaussian representations to ensure consistency across frames.

Method	1 View					
	PSNR ↑	SSIM ↑	LPIPS ↓	Identity ↓	FPS ↑	Modeling Time (s) ↓
LAM	17.30	0.773	0.149	0.135	125	**0.31**
MonoGaussianAvata	11.83	0.631	0.620	0.432	<10	>100
GaussianAvatars	16.35	0.740	0.332	0.299	<10	>100
FastAvatar	**20.08**	**0.860**	**0.143**	**0.116**	**240.17**	1.33
Method	4 Views					
	PSNR ↑	SSIM ↑	LPIPS ↓	Identity ↓	FPS ↑	Modeling Time (s) ↓
LAM*	16.69	0.743	0.204	0.167	45	**0.39**
MonoGaussianAvatar	12.71	0.798	0.437	0.368	<10	>100
GaussianAvatars	17.52	0.802	0.340	0.278	<10	>100
FastAvatar	**22.12**	**0.880**	**0.094**	**0.098**	**101.62**	4.22
Method	8 Views					
	PSNR ↑	SSIM ↑	LPIPS ↓	Identity ↓	FPS ↑	Modeling Time (s) ↓
LAM*	16.59	0.718	0.235	0.206	24	**0.43**
MonoGaussianAvatar	13.11	0.650	0.493	0.298	<10	>100
GaussianAvatars	20.35	0.820	0.320	0.252	<10	>100
FastAvatar	**22.19**	**0.880**	**0.093**	**0.097**	**52.28**	8.56
Method	16 Views					
	PSNR ↑	SSIM ↑	LPIPS ↓	Identity ↓	FPS ↑	Modeling Time (s) ↓
LAM*	16.49	0.697	0.265	0.238	13	**0.69**
MonoGaussianAvatar	15.81	0.721	0.406	0.202	<10	>100
GaussianAvatars	21.48	0.873	0.281	0.185	<10	>100
FastAvatar	**22.29**	**0.881**	**0.092**	**0.095**	**17.65**	26.06

Table 1: The quantitative comparison among FastAvatar, LAM He et al. (2025), MonoGaussianAvatar Chen et al. (2024b), and GaussianAvatars Qian et al. (2024a) includes 3 critical metrics: Reconstruction quality (PSNR, SSIM, LPIPS); Modeling time: Duration required to reconstruct the 3DGS model; Inference speed: Animation rendering FPS of the output 3DGS model.

Pixel Losses The rendered RGB images are supervised using photometric losses against the corresponding ground truth target images:

$$\mathcal{L}_{rgb} = \left\| \hat{I}_i, I^{gt} \right\|_1 + \left\| \hat{I}_{\text{sliced}}, I^{gt} \right\|_1, \tag{6}$$

$$\mathcal{L}_{ssim} = \text{SSIM}(\hat{I}_i, I^{gt}) + \text{SSIM}(\hat{I}_{\text{sliced}}, I^{gt}), \tag{7}$$

We also incorporate perceptual losses to encourage the emergence of more high-frequency details:

$$\mathcal{L}_{lpips} = \text{LPIPS}(\hat{I}_i, I^{gt}) + \text{LPIPS}(\hat{I}_{\text{sliced}}, I^{gt}). \tag{8}$$

Landmark Tracking Loss Unlike novel view synthesis, canonical space registration is supervised directly on the input frames. The landmark tracking loss is introduced to encourage precise localization of facial landmarks throughout the input images:

$$\mathcal{L}_{track} = \sum_{j=1}^{M} \sum_{i=1}^{N} \| y_{j,i} - \hat{y}_{j,i} \| . \tag{9}$$

Our total loss is defined as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{rgb} + \lambda_2 \mathcal{L}_{ssim} + \lambda_3 \mathcal{L}_{lpips} + \lambda_4 \mathcal{L}_{track} + \lambda_5 \mathcal{L}_{mask}, \tag{10}$$

with $\lambda_1 = 0.8$, $\lambda_2 = 0.1$, $\lambda_3 = 0.1$, $\lambda_4 = 0.1$ and $\lambda_5 = 0.0005$.

4 Experiments

4.1 Training

We train FastAvatar on a multi-task dataset derived from NeRSemble Kirschstein et al. (2023), which contains multi-person, multi-camera videos with a wide range of facial expressions. To encourage

adaptability to diverse input settings, the dataset includes both monocular and multi-view subsets. Input-output pairs are constructed by sampling 16 frames each, either from a single video or from 12 camera views of the same subject. For each pair, a random subset of 1 to 16 input frames is further selected, enabling the model to robustly handle scenarios with sparse or varying numbers of input frames—such as real-world recordings with incomplete or irregular camera captures.

4.2 Experimental Setups

Task. We evaluate the model's ability to generate a 3D head avatar for unseen subjects from various types of input, including a single image, an unordered and arbitrary number of monocular video frames, and multi-view frames.

Metrics. We employ three paired-image metrics to measure the quality of individual rendered images: Peak Signal-to-Noise (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). We also evaluate identity preservation by computing similarity using ArcFace Deng et al. (2019) features.

Baselines. We compare FastAvatar with recent state-of-the-art systems for 3D head avatar generation across various tasks, including reconstruction from a single image, monocular video frames and multi-view frames. LAM He et al. (2025) A model that generates one-shot animatable Gaussian heads using a canonical Transformer with point-cloud representation and multi-scale cross-attention, enabling real-time, expression-consistent avatar animation and editing from a single image. Avat3r Kirschstein et al. (2025) A system for regressing animatable 3D head avatars from limited multi-view images by combining large reconstruction and foundation models with cross-attention layers to effectively model 3D facial dynamics and generalize across diverse data. MonoGaussianAvatar Chen et al. (2024b) and GaussianAvatar Qian et al. (2024a) are two representative optimization-based 3D Avatar methods, both of which use FLAME as a proxy 3D model, similar to ours.

We conduct all comparison experiments on the same 48GB Nvidia RTX 4090 GPU. To ensure a fair comparison, we slightly modify the LAM renderer for single-shot input, enabling it to fuse information from multiple frames like FastAvatar, while keeping all other components consistent with the official repository. We retain the original model weights and confirm that its performance faithfully reflects the official version. We refer to this variant as LAM* in the following. Notably, with only 1 input frame, LAM* automatically reverts to the official LAM, which is designed for single-frame inputs, ensuring fair comparison.

4.3 Qualitative Comparison

Comparative results against LAM, MonoGaussian, and GaussianAvatar are presented in Figure 3. We evaluate 4 distinct input configurations (1, 4, 8, and 16 views) by progressively increasing the number of input frames. Key observations indicate that while LAM yields performance roughly on par under single-view conditions, it fails to benefit from additional input views due to the lack of registration. Conversely, optimization-based methods exhibit significant performance degradation with sparse inputs (e.g., 1 view), though their reconstruction quality improves progressively as more views become available. FastAvatar consistently outperforms the baseline across all view settings (1$\sim$16 views), while further enhancing its ability to capture fine-grained details—such as teeth gaps, wrinkles, and acne—as the number of views increases.

4.4 Quantitative Comparison

Table 1 presents a quantitative comparison of the four methods. All approaches demonstrate substantial improvements in reconstruction metrics with increasing input frames, with both MonoGaussianAvatar and GaussianAvatar exhibiting gains in both subjective assessments (i.e., LPIPS) and objective metrics (i.e., PSNR, SSIM). This reaffirms the critical importance of richer input data for high-fidelity reconstruction. However, LAM shows an inverse relationship: as its input views increase, quantitative performance degrades, which can also be illustrated in Figure 4. Although LAM achieves impressive visual quality with single image (LPIPS: 0.149), its generative bias introduces pose and expression artifacts that compromise objective measurements. FastAvatar achieves highly

Figure 3: We benchmark FastAvatar against representative optimization-based methods (MonoGaussian Avatar Chen et al. (2024b), GaussianAvatar Qian et al. (2024a)) and feedforward approaches (LAM He et al. (2025)). Our results demonstrate the performance evolution across methods as the number of input views (referring to input images number) increases. Please zoom in for a better view.

competitive growth across both subjective and objective dimensions, validating our core hypothesis. Through architectural and training innovations, FastAvatar establishes an optimal equilibrium between generative capability (hallucinating plausible details under sparse inputs) and reconstruction fidelity (strict adherence to observed data given sufficient views).

4.5 Incremental Reconstruction

FastAvatar enables incremental improvement by accepting input sequences of any length and order. As more observations are added, the reconstruction quality continues to improve. In contrast, existing methods often require a fixed number of input frames, which reduces flexibility and may result in data loss. FastAvatar's incremental design thus ensures both versatility and efficient data usage.

As illustrated in Figure 4, our method achieves superior expressiveness and rendering fidelity in avatar generation compared to the baselines, and demonstrates robust performance even for subjects wearing accessories.

Moreover, increasing the number of input views further improves the reconstruction of fine-grained details, such as hair and teeth gaps. This incremental reconstruction allows the model to overcome the limitations of restricted viewpoints by leveraging more informative input frames—a capability that is difficult to achieve for models with fixed input forms. For example, in Figure 4, the character's left-ear earring is not sufficiently observed with a small number of input views, but it is reliably reconstructed as the number of views increases.

Unlike the sparse and randomly sampled data used in our main experiments, real-world sequences are highly continuous. Processing all frames with Global Attention imposes prohibitive computational and memory costs. A naive solution is to uniformly sample 16 frames, but this risks missing important information present in the remaining frames. Inspired by FramePack Zhang et al. (2025), we retain the 16 uniformly sampled frames as sparse inputs and compress all remaining frames into an aggregated token representation, which is then treated as an additional input. This preserves complementary details while keeping computation tractable, enabling FastAvatar to process hundreds of frames in a single feed-forward pass. Figure 5 shows that adding the compressed frames improves fine-grained details.

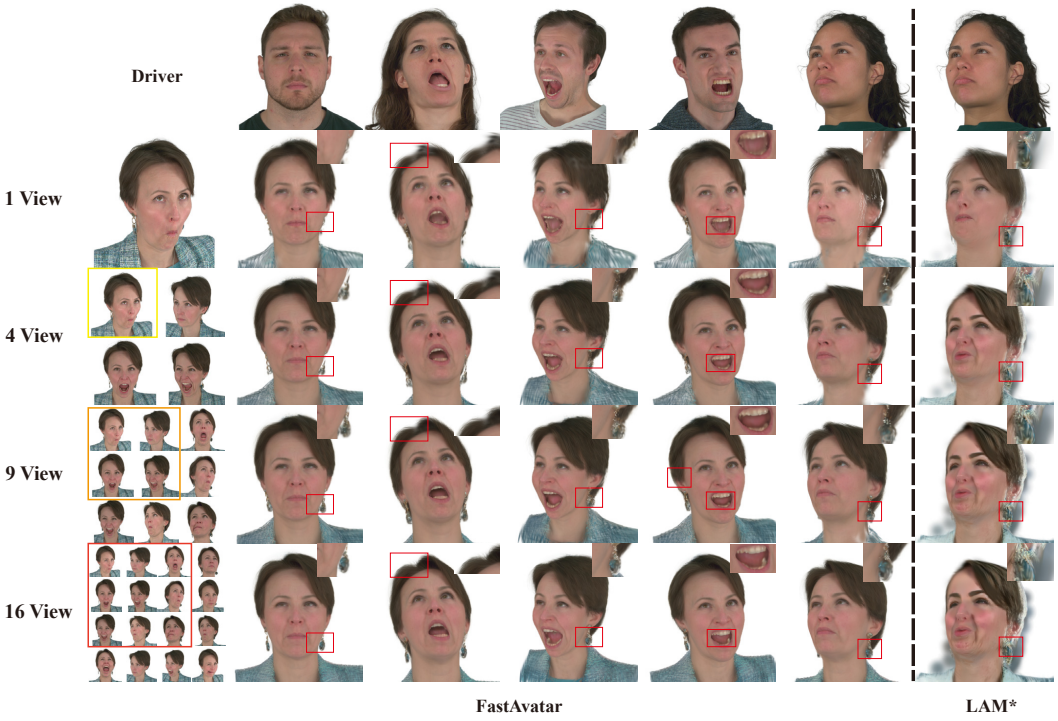


Figure 4: Reconstruction quality as the number of input observations increases. More observations improve reconstruction quality.

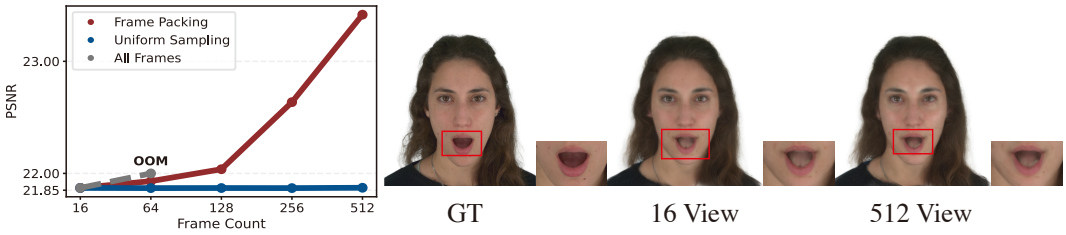


Figure 5: Performance on longer input sequences. Starting from strong reconstructions using only the 16 sparse input frames, incorporating the compressed additional frames further enhances fine-grained details (e.g., the oral cavity, which is absent in most frames). While uniform sampling fails to achieve this improvement, feeding all frames leads to OOM.

4.6 Multi-view Observations Reconstruction

To further evaluate FastAvatar's performance on the Multi-view Observations Reconstruction task, we create multi-view few-shot 3D head avatars for subjects from the Ava256 Martinez et al. (2024) dataset, which was not used during training. We compare our results with those of the state-of-the-art method Avat3r Kirschstein et al. (2025) and use the results provided in its original paper to ensure a fair comparison (since its implementation is not publicly available, we ensure fairness by directly using results from the Avat3r paper). The qualitative results are shown in Figure 6. Note that we only used images from Ava256 for tracking and to obtain FLAME parameters, without utilizing the provided informed encodings. Nevertheless, FastAvatar still achieves highly competitive results and benefits from additional multi-view inputs, producing more detailed reconstructions. Close inspection reveals that Avat3r fails to preserve facial identity accurately, reconstructing consistently wider facial structures than observed in source inputs.

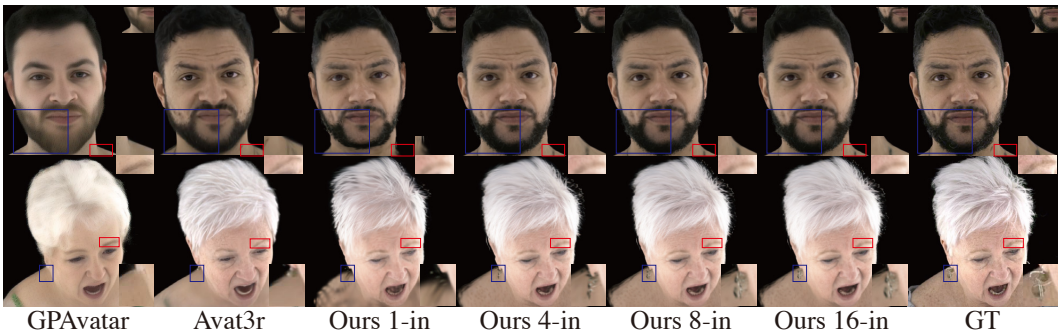


Figure 6: Visual comparison with Avat3r and GPAvatar Chu et al. (2024). Please zoom in for a clearer view.

Method	1 View						4 View					
	L1 ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Identity ↓	# GS (K) ↓	L1 ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Identity ↓	# GS (K) ↓
w/o sliced fusion loss	0.0373	20.47	0.861	**0.123**	0.158	20.7	0.0345	21.69	0.857	0.131	0.138	82.8
w/o tracking loss	**0.0372**	**20.93**	**0.863**	0.140	0.172	13.0	0.0322	21.64	0.866	0.120	0.128	41.2
w/o global attention	0.0922	15.40	0.760	0.238	0.400	**10.3**	0.0462	19.49	0.828	0.162	0.270	40.1
w/o GS fusion	0.0379	20.31	0.857	0.136	**0.138**	12.8	0.0467	18.94	0.838	0.157	0.185	**12.5**
w/o GS pruning	0.0380	20.32	0.857	0.137	0.144	21.7	0.0322	21.67	0.868	0.112	0.130	86.7
Ours full	0.0379	20.31	0.857	0.136	0.148	12.8	**0.0303**	**21.86**	**0.871**	**0.107**	**0.118**	42.8
Method	8 View						16 View					
	L1 ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Identity ↓	# GS (K) ↓	L1 ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Identity ↓	# GS (K) ↓
w/o sliced fusion loss	0.0386	21.12	0.849	0.144	0.151	165.4	0.0417	20.62	0.839	0.159	0.180	330.5
w/o tracking loss	0.0320	21.61	0.867	0.119	0.124	78.6	0.0322	21.66	0.865	0.123	0.129	164.2
w/o global attention	0.0413	19.97	0.835	0.156	0.223	78.0	0.0405	20.06	0.830	0.167	0.210	155.7
w/o GS fusion	0.0574	17.44	0.823	0.179	0.227	**12.5**	0.0682	16.25	0.811	0.196	0.259	**12.4**
w/o GS pruning	0.0326	21.63	0.868	0.110	0.130	173.4	0.0327	21.61	0.867	0.110	0.136	346.8
Ours full	**0.0297**	**21.95**	**0.871**	**0.103**	**0.118**	77.0	**0.0293**	**22.04**	**0.876**	**0.102**	**0.118**	138.9

Table 2: Ablation studies on key components of FastAvatar. The appendix visualizations are strongly recommended for better understanding.

4.7 Ablation Study

To validate the effectiveness of each component in our method, we conduct both quantitative and qualitative ablation studies. The qualitative visualizations are provided in the appendix for space considerations. As shown in Table 2, Global Attention is crucial for coordinating inter-frame information, while GS Fusion aggregates the GS points from each frame into a unified representation. Sliced Fusion Loss and Tracking Loss supervise GS registration, enforcing structural consistency and temporal coherence. Without these components, newly introduced frames fail to provide reliable information, leading to degraded registration and blurred outputs as the number of frames increases. Meanwhile, Gaussian Pruning removes redundant primitives, slightly improving performance while substantially accelerating rendering. Together, these mechanisms ensure effective inter-frame coordination, accurate registration, and efficient rendering.

5 Conclusion

In this paper, we present FastAvatar, a feed-forward 3D avatar reconstruction framework capable of constructing a high-quality animatable 3DGS avatar within seconds. Distinct from existing approaches, FastAvatar demonstrates a unique capability for incremental avatar reconstruction – flexibly leveraging incoming observations to progressively enhance reconstruction quality. We contend this represents a promising research trajectory. Three pivotal innovations enable this functionality: Alternating Attention, augmented with fine-grained expression and pose encodings, achieves high-precision registration of unordered data; The proposed Landmark Tracking Loss and Sliced Fusion Loss facilitate robust fusion of multiple 3DGS representations for superior modeling fidelity. Experimental validation confirms FastAvatar's potential in these dimensions.

REFERENCES

ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. Rignerf: Fully controllable neural 3d portraits. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 20364–20373, 2022.

ShahRukh Athar, Zhixin Shu, and Dimitris Samaras. Flame-in-nerf: Neural control of radiance fields for free view face animation. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–8. IEEE, 2023.

Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In Warren N. Waggenspack (ed.), *SIGGRAPH 1999*, pp. 187–194. ACM, 1999.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19457–19467, 2024.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pp. 370–386. Springer, 2024a.

Yufan Chen, Lizhen Wang, Qijing Li, Hongjiang Xiao, Shengping Zhang, Hongxun Yao, and Yebin Liu. Monogaussianavatar: Monocular gaussian point-based head avatar. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–9, 2024b.

Xuangeng Chu, Yu Li, Ailing Zeng, Tianyu Yang, Lijian Lin, Yunfei Liu, and Tatsuya Harada. Gpavatar: Generalizable and precise head avatar from image(s). In *ICLR 2024*. OpenReview.net, 2024.

Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10101–10111, 2019.

Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20311–20322, 2022.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021.

Dimitrios Gerogiannis, Foivos Paraperas Papantoniou, Rolandos Alexandros Potamias, Alexandros Lattas, and Stefanos Zafeiriou. Arc2avatar: Generating expressive 3d avatars from a single image via id guidance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10770–10782, 2025.

Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5784–5794, 2021.

Yisheng He, Xiaodong Gu, Xiaodan Ye, Chao Xu, Zhengyi Zhao, Yuan Dong, Weihao Yuan, Zilong Dong, and Liefeng Bo. Lam: Large avatar model for one-shot animatable gaussian head. In *SIGGRAPH*, 2025.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023.

Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 634–644. IEEE, 2024.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=rkE3y85ee.

Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, Dahua Lin, and Bo Dai. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *CoRR*, abs/2505.23716, 2025. doi: 10.48550/ARXIV.2505.23716. URL https://doi.org/10.48550/arXiv.2505.23716.

Yuheng Jiang, Zhehao Shen, Penghao Wang, Zhuo Su, Yu Hong, Yingliang Zhang, Jingyi Yu, and Lan Xu. Hifi4g: High-fidelity human performance rendering via compact gaussian splatting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 19734–19745. IEEE, 2024.

Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias. *arXiv preprint arXiv:2410.17242*, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Trans. Graph.*, 42(4), jul 2023. ISSN 0730-0301. doi: 10.1145/3592455. URL https://doi.org/10.1145/3592455.

Tobias Kirschstein, Javier Romero, Artem Sevastopolsky, Matthias Nießner, and Shunsuke Saito. Avat3r: Large animatable gaussian reconstruction model for high-fidelity 3d head avatars, 2025. URL https://arxiv.org/abs/2502.20220.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

Minghua Liu, Chong Zeng, Xinyue Wei, Ruoxi Shi, Linghao Chen, Chao Xu, Mengqi Zhang, Zhaoning Wang, Xiaoshuai Zhang, Isabella Liu, et al. Meshformer: High-quality mesh generation with 3d-guided reconstruction model. *Advances in Neural Information Processing Systems*, 37:59314–59341, 2024.

Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. Semantic-aware implicit neural audio-driven video portrait generation. In *European conference on computer vision*, pp. 106–125. Springer, 2022.

Yifei Liu, Zhihang Zhong, Yifan Zhan, Sheng Xu, and Xiao Sun. Maskgaussian: Adaptive 3d gaussian representation from probabilistic masks. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 681–690, June 2025.

Haoyu Ma, Tong Zhang, Shanlin Sun, Xiangyi Yan, Kun Han, and Xiaohui Xie. Cvthead: One-shot controllable head avatar with vertex-feature transformer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6131–6141, 2024.

Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, Chenghui Li, Shih-En Wei, Rohan Joshi, Wyatt Borsos, Tomas Simon, Jason Saragih, Paul Theodosis, Alexander Greene, Anjani

Josyula, Silvio Mano Maeta, Andrew I. Jewett, Simon Venshtain, Christopher Heilman, Yueh-Tung Chen, Sidi Fu, Mohamed Ezzeldin A. Elshaer, Tingfang Du, Longhua Wu, Shen-Chi Chen, Kai Kang, Michael Wu, Youssef Emad, Steven Longay, Ashley Brewer, Hitesh Shah, James Booth, Taylor Koska, Kayla Haidle, Matt Andromalos, Joanna Hsu, Thomas Dauer, Peter Selednik, Tim Godisart, Scott Ardisson, Matthew Cipperly, Ben Humberston, Lon Farr, Bob Hansen, Peihong Guo, Dave Braun, Steven Krenn, He Wen, Lucas Evans, Natalia Fadeeva, Matthew Stewart, Gabriel Schwartz, Divam Gupta, Gyeongsik Moon, Kaiwen Guo, Yuan Dong, Yichen Xu, Takaaki Shiratori, Fabian Prada, Bernardo R. Pires, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, and Yaser Sheikh. Codec Avatar Studio: Paired Human Captures for Complete, Driveable, and Generalizable Avatars. *NeurIPS Track on Datasets and Benchmarks*, 2024.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Panwang Pan, Zhuo Su, Chenguo Lin, Zhen Fan, Yongjie Zhang, Zeming Li, Tingting Shen, Yadong Mu, and Yebin Liu. Humansplat: Generalizable single-image human gaussian splatting with structure priors. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *CVPR 2024*, pp. 20299–20309. IEEE, 2024a.

Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 5020–5030. IEEE, 2024b.

Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, and Liefeng Bo. LHM: large animatable human reconstruction model from a single image in seconds. *CoRR*, abs/2503.10625, 2025.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV 2021*, pp. 12159–12168. IEEE, 2021.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European conference on computer vision*, pp. 501–518. Springer, 2016.

Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16632–16642, 2023.

Stanislaw Szymanowicz, Chrisitian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10208–10217, 2024.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2024.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.

Jie Wang, Jiu-Cheng Xie, Xianyan Li, Feng Xu, Chi-Man Pun, and Hao Gao. Gaussianhead: High-fidelity head avatars with learnable gaussian derivation. *IEEE Transactions on Visualization and Computer Graphics*, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024.

Jing Wen, Xiaoming Zhao, Zhongzheng Ren, Alexander G. Schwing, and Shenlong Wang. Gomavatar: Efficient animatable human modeling from monocular video using gaussians-on-mesh. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 2059–2069. IEEE, 2024.

Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1931–1941, 2024.

Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. *arXiv preprint arXiv:2410.24207*, 2024.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024a.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Zhaoliang Zhang, Tianchen Song, Yongjae Lee, Li Yang, Cheng Peng, Rama Chellappa, and Deliang Fan. Lp-3dgs: Learning to prune 3d gaussian splatting. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 122434–122457. Curran Associates, Inc., 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/dd51dbce305433cd60910dc5b0147be4-Paper-Conference.pdf.

Xiaozheng Zheng, Chao Wen, Zhaohu Li, Weiyi Zhang, Zhuo Su, Xu Chang, Yang Zhao, Zheng Lv, Xiaoyuan Zhang, Yongjie Zhang, et al. Headgap: Few-shot 3d head avatar via generalizable gaussian priors. In *2025 International Conference on 3D Vision (3DV)*, pp. 946–957. IEEE, 2025.

Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4574–4584, 2023.

Wojciech Zielonka, Stephan J Garbin, Alexandros Lattas, George Kopanas, Paulo Gotardo, Thabo Beeler, Justus Thies, and Timo Bolkart. Synthetic prior for few-shot drivable head avatar inversion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10735–10746, 2025.

A LLM USAGE

Large Language Models (LLMs) were used solely to assist in refining the manuscript's language, improving readability, and ensuring clarity, including sentence rephrasing and grammar checking. The LLM did not contribute to the research ideas, methodology, experiments, or data analysis. The authors take full responsibility for the scientific content and confirm that all LLM-assisted text adheres to ethical guidelines, with no contribution to plagiarism or misconduct.

B ETHICS STATEMENT

The proposed FastAvatar framework follows the same data assumptions and usage boundaries as existing 3D avatar reconstruction and neural rendering methods. It does not introduce new mechanisms that lower the barrier to unauthorized identity reconstruction, nor does it relax requirements on input data quality. In practice, the method still relies on clean and identity-consistent multi-frame input, which inherently limits large-scale or covert misuse.

All experiments are conducted on publicly available datasets with appropriate licenses, and no private or sensitive data are collected. The intended applications of FastAvatar lie in areas such as AR/VR telepresence, digital content creation, and human–computer interaction, where user consent is typically explicit. We emphasize that responsible deployment requires ensuring that the method is not applied to reconstruct individuals without consent or to generate deceptive or impersonating content.

C REPRODUCIBILITY

In this section, we provide more implementation details of FastAvatar, including data preparation and model architecture. Furthermore, our code will be released after the paper is accepted.

C.1 DATA PREPARATION

Our training utilizes the Nersemble dataset. Initially, FLAME tracking is applied to extract FLAME parameters and camera poses, which serve as inputs for subsequent training stages. From the original Nersemble data, we extract 521 distinct video clips, and sample the frames at 15 FPS. Cameras with poor face tracking quality were excluded, remaining 12 cameras. The processed data was sampled twice to construct the final dataset: first, sampling frames within the same video sequence, and second, performing random sampling across all shots of the same action sequence. These two sampling strategies collectively support training the unified task. To enhance the stability of training and testing, we randomly assign the processed figures' backgrounds to black, white, or gray. Note that, to enhance the model's generative capability, all expression parameters, poses, and related data used during inference differ from the input data.

	Hyperparameter	Value
Encoder	DINOv2 patch size	14×14
	#expression token MLP layers	2
	#camera token MLP layers	2
	Expression Token MLP activation	GELU
	Camera Token MLP activation	GELU
	Output dimension	1024
	Input image resolution	504×504
AA	#Frame Attn Layers	10
	#Global Attn Layers	10
	Hidden dimension	1024
	Order	[Global, Frame]

Table 3: Hyperparameters. Where AA represents Alternation Attention

Noise	L1↓	PSNR↑	SSIM↑	LPIPS↓	Identity↓
1 px	0.0264	**22.50**	**0.873**	**0.096**	0.100
4 px	0.0268	22.38	0.872	0.098	**0.099**
8 px	0.0273	22.22	0.870	0.098	0.103
16 px	0.0277	22.10	0.869	0.099	0.102
32 px	0.0280	22.02	0.869	0.099	0.105
Ours	**0.0263**	**22.50**	0.872	**0.096**	**0.099**

Table 4: Ablation studies on FLAME tracking. We evaluate the robustness of FastAvatar under varying levels of landmark perturbation during FLAME tracking.

The accuracy of FLAME tracking primarily depends on the precision of detected facial landmarks, as the FLAME parameters are typically estimated by optimizing the model to fit these landmarks. However, the reliability of such proxy models (e.g., FLAME and other 3DMMs) is inherently constrained by factors such as limited representational capacity and sensitivity to landmark quality. To assess how these factors affect our method, we include an additional ablation experiment that injects controlled noise into the facial landmarks. The results (Table 4) demonstrate that FastAvatar remains robust under reasonable perturbations, indicating that strict accuracy in FLAME tracking is not required.

C.2 Training

In table 3, we present the most important hyperparameters for training FastAvatar. We train the model by optimizing the training loss with the AdamW optimizer for 150K iterations. We use a cosine learning rate scheduler with learning rate of 4e-5. The training runs on 8 H100 GPUs over 14 days. The substantial accumulation of Gaussian points across multiple input frames leads to high GPU memory consumption during training. To address this, we adopt bfloat16 precision and gradient checkpointing for improved memory and computational efficiency.

D More Results

In this section, we present additional results of FastAvatar, including its performance on a broader set of video sequences and its generalization to real-world daily-captured data.

Figure 7: Qualitative results on the INSTA dataset. LPIPS scores: Ours (**0.1332**), LAM (0.1479), Arc2Avatar (0.4585), SynShot (0.1523). Identity scores: Ours (**0.076**), LAM (0.124), Arc2Avatar (0.411), SynShot (0.115).

More Comparison We provide additional qualitative results comparing FastAvatar and baseline models in both self-reenactment and cross-reenactment settings. As shown in Figure 15, Figure 16, and Figure 14. FastAvatar outperforms the baselines. Optimization-based 3D avatar methods fail to achieve satisfactory results with sparse inputs, while LAM often exhibits unrealistic details and significant pose inaccuracies. The advantage of FastAvatar becomes even more evident in the cross-reenactment setting, where the subject's identity and camera pose exhibit large discrepancies. We further evaluate FastAvatar against additional competitive methods. The results are presented in Figure 7 and Figure 8.

Generalization to Wide-Range Viewpoints The Nersemble training set contains only 12 constrained camera views. To evaluate the robustness of our method, we test it on a much wider range

Figure 8: Qualitative results on the Nersemble dataset. LPIPS scores: Ours (**0.1267**), LAM (0.1608), Arc2Avatar (0.4665), HeadGap (0.1592). Identity scores: Ours (**0.070**), LAM (0.097), Arc2Avatar (0.308), HeadGap (0.101).

Figure 9: Comparison of visual effects of model reconstruction after removing the key components.

of viewpoints. As shown in Figure 12, the model maintains high-fidelity reconstruction across all novel views, demonstrating strong wide-range generalization. For comparison, we include the results of LAM in Figure 13. The results demonstrate that FastAvatar outperforms the state-of-the-art across a wide range of viewpoints.

Ablation Study We further highlight the role of the key components in incremental reconstruction. As illustrated in Figure 9, removing these components prevents fine details from being properly aligned, leading to noticeable artifacts and blurred regions. Although the landmark tracking loss only supervises 68 facial landmark points, it still provides strong guidance for Gaussian registration, effectively assisting the model in aligning new frames during incremental updates. Together with the Sliced Fusion Loss, it ensures that additional observations can be accurately fused, enabling consistent refinement of the reconstructed avatar. Global Attention enables the model to leverage inter-frame dependencies, integrating complementary features from multiple frames; without it, information remains localized, and cross-frame consistency cannot be achieved. GS Fusion consolidates per-frame Gaussians into a coherent representation, allowing the model to maintain consistency across frames. Finally, Gaussian Pruning removes redundant primitives, slightly improving performance while significantly accelerating rendering, enabling efficient incremental updates even for long sequences.

Streaming Incremental Reconstruction FastAvatar is also capable of streaming incremental reconstruction, meaning the model continuously updates and refines the 3DGS representation as new video frames are received. To achieve this, we adopt a sliding-window approach, where each window contains 16 frames with a 4-frame overlap for registering incoming frames. Thanks to the Alternating Attention design, new frames can build upon the Gaussians predicted by the previous model to produce more informative reconstructions. Figure11 demonstrates this: starting from a single-view image, as additional views are provided (excluding the test view), reconstruction quality improves overall, including at previously unseen angles, thus realizing streaming incremental reconstruction.

Limitations First, our method relies on LBS and FLAME-based encodings to drive 3D head avatar motion, which limits the representation of complex facial muscle dynamics. As a result, the model has difficulty reproducing fine-grained muscle-dependent details such as wrinkles and also cannot accurately capture eye-gaze movements, often defaulting to an average direction. Furthermore, because the Gaussians are anchored to FLAME vertices, the model is unable to represent structures outside the FLAME topology, including the tongue. Figure 10 presents representative failure cases.

Figure 10: Typical failure cases: FastAvatar relying on LBS and FLAME-based encodings, struggles with complex facial muscle dynamics, fine-grained details (e.g., wrinkles), eye-gaze movements, and structures outside the FLAME topology such as the tongue.

Figure 11: As the streaming input is progressively incorporated, the reconstruction of the oral cavity—largely invisible in most chunks—gradually improves while structural consistency is maintained in other regions, enabling incremental reconstruction.

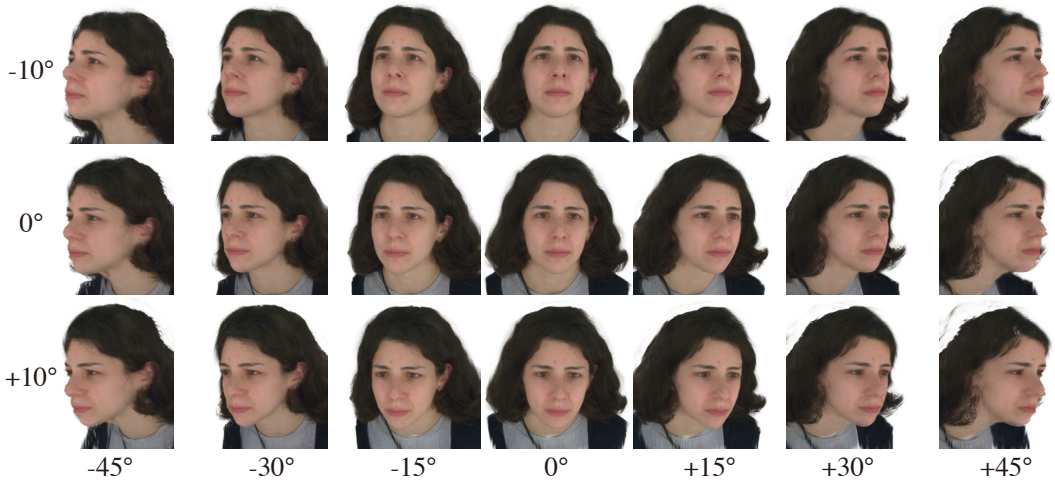


Figure 12: Generalization to wide-range viewpoints. FastAvatar achieves high-fidelity reconstruction across 14 novel viewpoints that are entirely outside the training set.

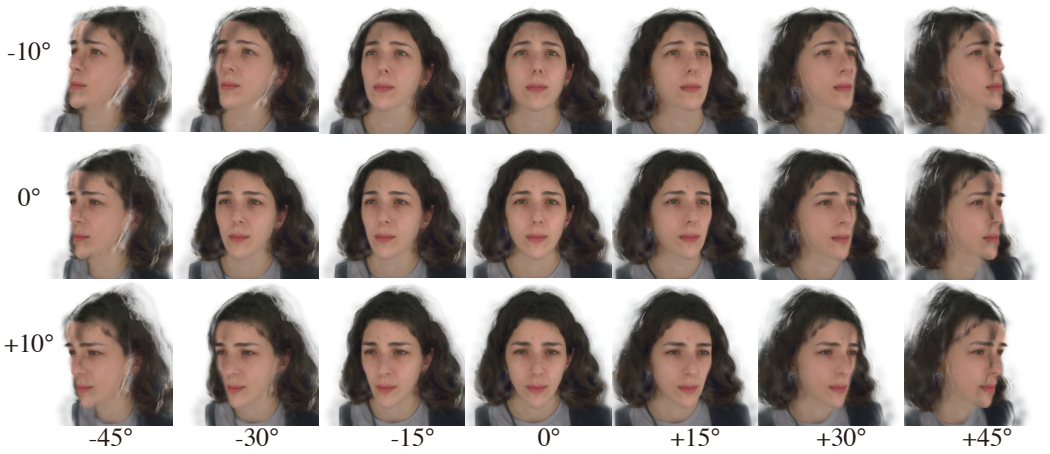


Figure 13: The performance of LAM on wide-range viewpoints.

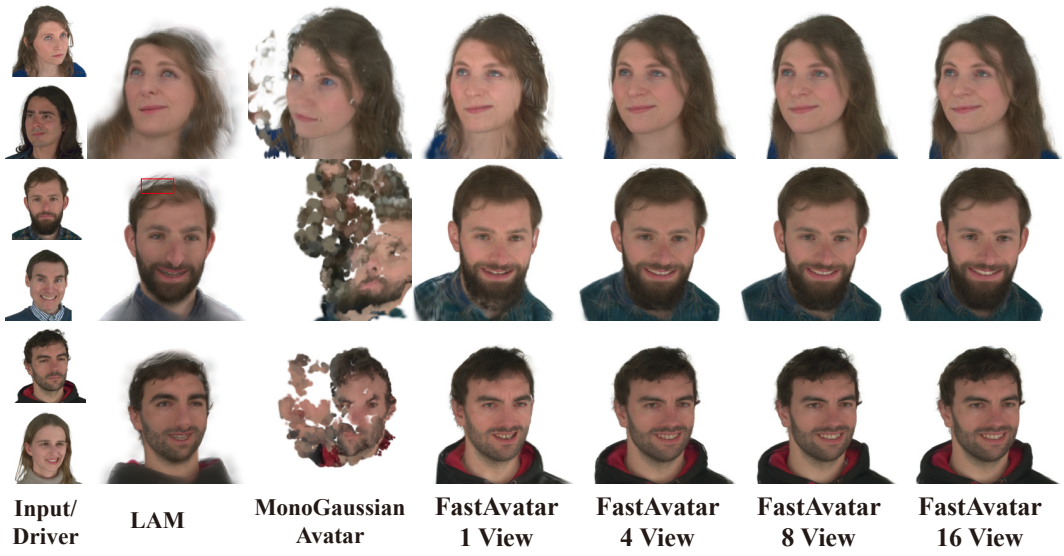
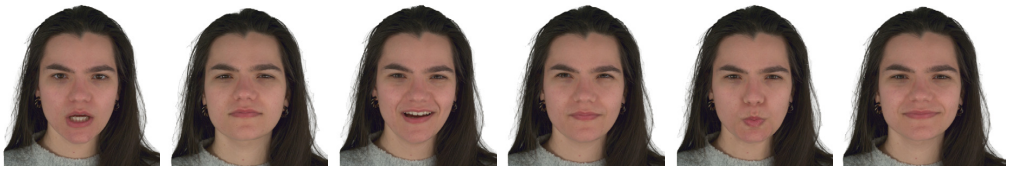


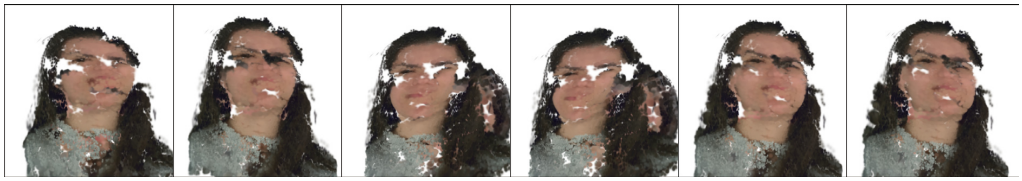
Figure 14: Additional Comparisons with Baseline Methods (cross-reenactment).

GT

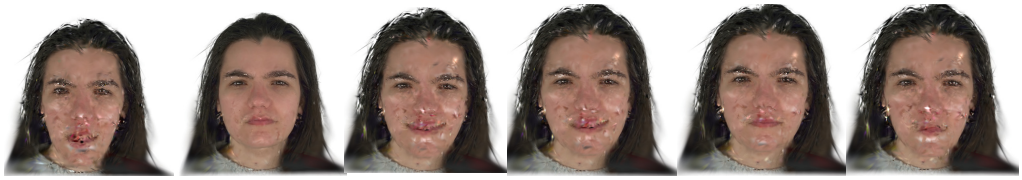


LAM

MonoGaussianAvatar



GaussianAvatars



FastAvatar

Figure 15: Additional Comparisons with Baseline Methods (self-reenactment part 1).

GT

LAM

MonoGaussianAvatar

GaussianAvatars

FastAvatar

Figure 16: Additional Comparisons with Baseline Methods (self-reenactment part 2).