GLASSHEAD: LEARNING 3D EYEGLASS-WEARING HEADS FROM A SINGLE PORTRAIT VIA PSEUDO-MULTIVIEW SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-image 3D reconstruction of eyeglass-wearing heads is challenging because important structures, particularly the temples and their attachments around the ears, are poorly observed from a frontal portrait. Learning to recover these regions requires diverse viewpoint supervision, yet large-scale multiview captures with varied eyeglasses are expensive and difficult to obtain. We address this data bottleneck by using generative video as a scalable source of pseudo-multiview supervision. We introduce a scalable pipeline that leverages a video generation model to transform frontal portraits into view-diverse pseudo-multiview observations, recover their camera parameters, and construct the Eyeglass Portrait Pseudo-Multiview Dataset (EPPM-20K). Building on this supervision, we propose GlassHead, a feed-forward reconstructor that predicts a complete 3D Gaussian head from a single portrait through hierarchical coarse-to-fine decoding. To accommodate inconsistencies in generated views, we combine reliable frontal reconstruction with perceptual and camera-conditioned adversarial supervision on pseudo-multiview observations. Experiments demonstrate consistent improvements over recent single-image head reconstruction methods, with particularly clear gains in lateral eyeglass structures under novel viewpoints. At inference time, GlassHead requires only a single frontal portrait and reconstructs the 3D head in one forward pass within one second.

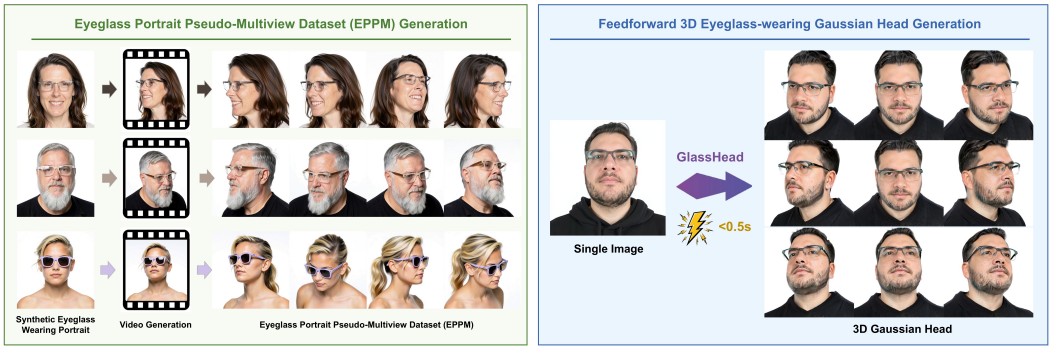

Figure 1: Overview of GlassHead. We construct the Eyeglass Portrait Pseudo-Multiview Dataset (EPPM) by generating view-diverse observations from synthetic eyeglass-wearing portraits, and train GlassHead with this pseudo-multiview supervision to reconstruct a complete 3D Gaussian head from a single frontal image in a feed-forward manner.

1 INTRODUCTION

Single-image 3D head reconstruction aims to recover a complete and novel-view-consistent head representation from a single RGB portrait. Recent feed-forward methods have made rapid progress by combining strong head priors with expressive neural representations such as 3D Gaussian Splatting Kerbl et al. (2023); Chu & Harada (2024); Lyu et al. (2025); Gao et al. (2026); Kirschstein

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al. (2026); Wu et al. (2026). Despite recent progress, single-image full-head reconstruction still struggles to recover regions that are weakly observed or invisible in the input, especially under large viewpoint changes. Consequently, reconstruction quality depends strongly on whether the training distribution provides sufficient geometric and appearance coverage for such regions.

Reconstructing eyeglass-wearing heads makes this limitation particularly evident. A frontal portrait clearly observes the rims and bridge, but provides little information about the temples and their attachment around the ears. Accurately recovering these structures therefore requires side-view supervision that is largely unavailable from the input image itself. However, existing high-quality multiview head datasets are typically captured using calibrated studio systems, which are costly to acquire and difficult to scale across diverse eyeglass stylesKirschstein et al. (2023); Martinez et al. (2024); Pan et al. (2024). Synthetic head data provides an alternative, but constructing realistic head assets with accurately aligned eyeglasses requires a complex rendering and asset-generation pipelineBuehler et al. (2024); Hewitt et al. (2024). As a result, the multiview supervision required by existing feed-forward 3D reconstruction methods is particularly limited for eyeglass portraits Wu et al. (2026); Kirschstein et al. (2026); Gao et al. (2026); Lyu et al. (2025); Kirschstein et al. (2025).

Recent advances in video generationWan Team (2025); Wu et al. (2025); MiniMax (2026) provide a promising way to overcome this supervision bottleneck. Given a single portrait of an eyeglass-wearing subject, modern video generation models can synthesize the same identity from substantially different viewpoints, revealing lateral structures that are not visible in the original image. This offers a scalable alternative to physical multiview capture system, and has the possibility that ordinary portrait collections can be transformed into much richer supervision for 3D reconstruction. However, generated views are not equivalent to real multiview observations. They may exhibit inconsistent geometry and appearance drift, making direct supervision prone to propagating generation artifacts into the reconstructed 3D representation.

In this work, we use generated videos to provide scalable pseudo-multiview supervision for single-image reconstruction of eyeglass-wearing heads. Starting from frontal eyeglass portraits, we generate view-diverse observations and recover their corresponding camera configurations, resulting in a large-scale pseudo-multiview training set. Based on this data, we train a feed-forward cascaded 3D Gaussian reconstructor that takes only the frontal portrait as input and predicts a high-quality 3D gaussian head representation. To account for the inevitable inconsistencies in generated views, we design the training objective to exploit their complementary viewpoint information without treating them as exact multiview ground truth. At inference time, the model directly reconstructs a 3D head from a single portrait without video generation or camera estimation (Figure 1).

Our contributions are summarized as follows:

- We introduce a scalable pipeline that generates diverse multiview observations from a single portrait using video generation models and recovers the corresponding camera poses to construct pseudo-multiview training data.
- Using this pipeline, we construct the Eyeglass Portrait Pseudo-Multiview Dataset (EPPM-20K), comprising nearly 20K eyeglass-wearing scenes with 320K views.
- We propose a feed-forward cascaded 3D Gaussian reconstruction framework that learns from imperfect pseudo-multiview supervision and reconstructs high-quality 3D eyeglass-wearing heads from a single portrait at inference time.

2 Related Work

2.1 Single-image 3D head reconstruction.

Single-image 3D head reconstruction has progressed from parametric face models to generalizable neural and Gaussian representations. Classical approaches recover low-dimensional 3DMM or FLAME parameters from monocular images (Blanz & Vetter, 1999; Li et al., 2017; Feng et al., 2021; Zielonka et al., 2022), providing stable facial geometry but limited representation of hair, accessories, and other structures beyond the predefined face model. Recent generalizable head reconstruction methods have substantially improved reconstruction from monocular or sparse observations (Chu et al., 2024; Chu & Harada, 2024; He et al., 2025; Kirschstein et al., 2026; Wu et al., 2026; Gao et al., 2026). Despite different representations and architectures, recovering geometry

that is not directly observed in the input relies heavily on priors learned from large-scale and diverse multiview training data. This requirement is especially important for eyeglasses, since the frontal rims and bridge are visible from a frontal portrait, whereas the temples and their interaction with the ears are largely unobserved and must be inferred from multiview supervision. However, acquiring such supervision at scale is difficult. Existing high-quality multiview head datasets are typically captured using synchronized and calibrated camera systems (Kirschstein et al., 2023; Martinez et al., 2024; Pan et al., 2024), which provide accurate observations but are expensive to collect and difficult to scale across diverse identities, hairstyles, and eyeglass styles. Synthetic data offers an alternative (Buehler et al., 2024; Hewitt et al., 2024; Gao et al., 2026), but constructing realistic 3D head assets together with accurately aligned accessories remains non-trivial. MEGANE explicitly models eyeglass geometry and its interaction with the face (Li et al., 2023), but relies on dedicated eyewear representations and controlled multiview acquisition. In contrast, our work focuses on obtaining scalable multiview supervision for eyeglass-wearing portraits without requiring specialized multiview capture.

Feed-forward head reconstruction. Recent feed-forward methods have enabled direct 3D head reconstruction from one or a few images without per-subject optimization. GPAvatar and GAGAvatar lift image observations into 3D representations with geometric guidance from parametric face or camera estimates (Chu et al., 2024; Chu & Harada, 2024), while LAM predicts Gaussian attributes from canonical FLAME queries (He et al., 2025). FlexAvatar learns a generalizable UV-based avatar representation from monocular and multiview supervision (Kirschstein et al., 2026), and more recent methods such as UIKA and Any3DAvatar further reduce the reliance on explicit input calibration (Wu et al., 2026; Gao et al., 2026). These methods generally decode the final 3D representation as a flat set of primitives. Interestingly, studies on 3D Gaussian generative modeling have shown that replacing flat prediction with a hierarchical organization can substantially improve generation quality and stability. GSGAN (Hyun & Heo, 2024) introduces a multi-scale Gaussian hierarchy in which finer Gaussians are progressively generated relative to coarser ones, and this design is later adopted and extended by CGS-GAN (Barthel et al., 2025). However, these hierarchical formulations are developed for GAN-based 3D generation rather than image-conditioned reconstruction. GlassHead brings this hierarchical principle into feed-forward single-image head reconstruction, progressively refining Gaussian geometry and appearance from coarse to fine.

3 METHOD

We study feed-forward 3D reconstruction of eyeglass-wearing portraits from a single image. Given a source portrait I^s, our goal is to predict a subject-specific Gaussian representation $\mathcal{G} = \{G_i\}_{i=1}^{N}$, where each Gaussian $G_i = (\boldsymbol{\mu}_i, \mathbf{s}_i, \mathbf{q}_i, \mathbf{c}_i, \alpha_i)$ is parameterized by its position, scale, rotation, color, and opacity.

We first construct EPPM-20K, a large-scale pseudo-multiview dataset from single frontal portraits by generating diverse views and recovering their corresponding camera parameters. Based on this dataset, we introduce GlassHead, a feed-forward hierarchical reconstructor that predicts a static 3D Gaussian representation from a single frontal image. During training, the model is supervised jointly by the frontal observation and the generated pseudo-multiview views, with objectives designed to tolerate imperfections in the generated data. At inference time, GlassHead requires only a single frontal portrait and reconstructs the complete 3D representation in one forward pass.

3.1 PSEUDO-MULTIVIEW EYEGLASS-WEARING PORTRAIT DATASET GENERATION

We construct the pseudo-multiview eyeglass-wearing portrait dataset (EPPM-20K) based on FFHQ (Karras et al., 2019). For each portrait, we first use FLUX.2 Klein (Black Forest Labs, 2026) to synthesize a frontal portrait with a white background while preserving identity and facial appearance (Figure 2 ①), and then apply a second editing step to add eyeglasses (Figure 2 ②), yielding the eyeglass-wearing source image I^s. Given I^s, we further employ the MiniMax H3 video generation model (MiniMax, 2026) to generate a short, continuous phone-scan-style video (Figure 2 ③). The subject is prompted to remain approximately static while the virtual camera moves around the face to generate diverse viewpoints with consistent identity, expression, eyeglasses, and background.

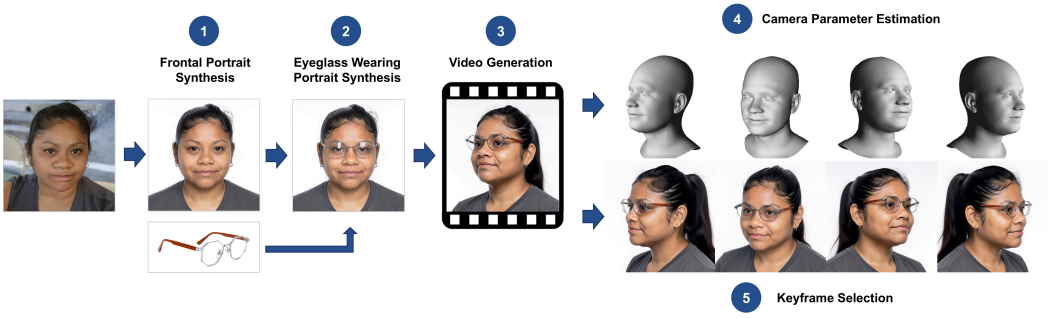

Figure 2: EPPM-20K dataset construction pipeline.

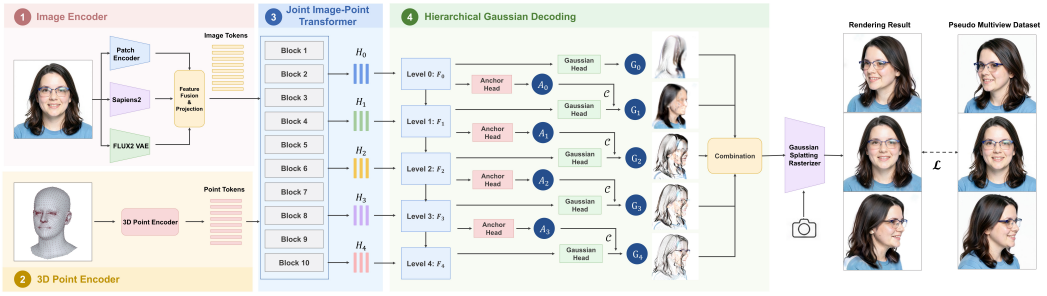

Figure 3: Model architecture of GlassHead.

The generated video provides multiview observations without explicit 3D geometry or camera information. To use these views for 3D supervision, we track the sequence in a shared FLAME canonical coordinate system and recover a relative camera for each view (Figure 2 ④). Specifically, 3DDFA-V2 (Guo et al., 2020) is used to estimate facial landmarks and head poses, while DECA (Feng et al., 2021) provides the initial identity parameters of the face model. We then jointly optimize the shared expression parameters, camera field of view (FoV), and per-view camera poses using a landmark reprojection loss with temporal regularization, followed by trajectory smoothing to obtain the final camera parameters.

Since neighboring frames are highly redundant and generated videos may contain motion blur or local synthesis artifacts, we estimate the head pose and image quality of each candidate frame using Hopenet (Ruiz et al., 2018) and TOPIQ (Chen et al., 2024), respectively. Frames are grouped according to their estimated head poses, and the highest-quality frame in each group is selected as a keyframe, yielding approximately 15–20 diverse pseudo-multiview observations for each source portrait (Figure 2

3.2 Feed-Forward Gaussian Reconstruction

Given a source portrait I^s, our reconstructor predicts a static 3D Gaussian representation $\mathcal{G}$ in a single forward pass. As illustrated in Figure 3, the architecture consists of four main components: an image encoder, a 3D point encoder, a joint image–point attention backbone, and a cascaded Gaussian decoder.

Image Encoder. Given a source portrait I^s with shape 512*512, we extract spatially aligned image features using a hybrid image encoder that combines trainable RGB features with pretrained semantic representations. Specifically, the RGB image is first partitioned into non-overlapping patches and mapped to a regular grid of learnable patch tokens. In parallel, we extract multi-level features from a frozen Sapiens2 backbone and latent features from a FLUX.2 VAE. Since all feature streams share the same spatial token layout, we concatenate them along the channel dimension and project the fused representation to the transformer backbone channel size:

$$\mathbf{X}^I = \phi_I\left(\left[\mathbf{F}^{\mathrm{rgb}}; \mathbf{F}^{\mathrm{sap}}; \mathbf{F}^{\mathrm{vae}}\right]\right), \tag{1}$$

where ϕ_I denotes a linear projection followed by normalization. This hybrid representation preserves local appearance cues from RGB while incorporating higher-level facial semantics and pretrained generative features.

3D Point Encoder. The 3D point encoder maps an initial set of learnable anchor coordinates $\mathbf{P} = \{\mathbf{p}_i\}_{i=1}^{K}$ into point tokens. We initialize the anchors from the canonical FLAME facial template. For each anchor, we first constrain its coordinate to a bounded canonical space, $\bar{\mathbf{p}}_i = \tanh(\mathbf{p}_i)$ and then encode each 3D anchor coordinate using a combination of learnable Fourier features (LFF) and learned positional embeddings. Specifically, given an anchor coordinate $\mathbf{p}_i$, we compute

$$\mathbf{x}_i^P = [\sin(\mathbf{W}_{\mathrm{LFF}}\mathbf{p}_i + \mathbf{b}_{\mathrm{LFF}})\,; \mathbf{e}(\mathbf{p}_i)]\,, \tag{2}$$

where $\mathbf{W}_{\mathrm{LFF}}$ and $\mathbf{b}_{\mathrm{LFF}}$ are learnable parameters and $\mathbf{e}(\cdot)$ denotes a learned positional embedding. The resulting tokens $\mathbf{X}^P$ encode the spatial locations of the initial 3D anchors and are subsequently used as queries for image–point interaction.

Joint Image–Point Transformer. Given the image tokens $\mathbf{X}^I$ and point tokens $\mathbf{X}^P$, we use a joint Transformer to exchange information between the 2D image and 3D point representations. In each block, the two streams are independently projected into queries, keys, and values, which are concatenated along the token dimension. A shared multi-head attention is then applied to the joint sequence, enabling direct interaction between image evidence and 3D point queries while preserving intra-modal communication. The attended features are subsequently split back into the image and point streams and updated through their respective residual and feed-forward layers. We retain point features $\{\mathbf{H}_\ell\}$ from multiple depths of the Transformer and use them as hierarchical representations for the subsequent Gaussian decoder. Specifically, our backbone consists of 10 Transformer blocks, from which five point representations are uniformly sampled at every two blocks.

Hierarchical Gaussian Decoding. Given the multi-level point representations $\{\mathbf{H}_\ell\}$, the decoder progressively constructs a coarse-to-fine Gaussian hierarchy. Each decoding stage combines features inherited from the previous level with level-specific details decoded from the corresponding Transformer representation. The root feature $\mathbf{F}_0$ is obtained directly from $\mathbf{H}_0$, while subsequent levels are constructed as

$$\mathbf{F}_\ell = \Phi_\ell(\mathcal{U}_\ell(\mathbf{F}_{\ell-1}), \Psi_\ell(\mathbf{H}_\ell)) \tag{3}$$

where $\mathcal{U}_\ell$ propagates and expands the features from the previous level, Ψ_ℓ decodes level-specific details from $\mathbf{H}_\ell$, and Φ_ℓ fuses the two streams. This parent–detail formulation preserves information accumulated at coarser levels while allowing each refinement stage to directly access image-conditioned features from the Transformer.

Each decoded feature $\mathbf{F}_\ell$ is processed by two parallel prediction heads. The *anchor head* predicts the parent state used to construct the next hierarchy level, whereas the *Gaussian head* predicts the renderable primitives at the current level. Both heads output residuals for position, scale, rotation, color, and opacity. Rather than directly regressing absolute Gaussian attributes, these residuals are composed with a parent anchor. Let the parent state be $(\mathbf{x}_p, \mathbf{s}_p, \mathbf{q}_p, \mathbf{c}_p, o_p)$, where $\mathbf{s}_p$ is represented in log-scale space, and let $(\mathbf{r}_x, \mathbf{r}_s, \mathbf{r}_q, \mathbf{r}_c, r_o)$ denote the predicted residuals. The child center is defined in the local coordinate frame of the parent as

$$\mathbf{x}_c = \mathbf{x}_p + \mathbf{R}(\mathrm{normalize}(\mathbf{q}_p))\,[\exp(\mathbf{s}_p) \odot \tanh(\mathbf{r}_x)]\,, \tag{4}$$

where the parent center defines the local origin, while its scale and orientation determine the spatial extent and direction of the child displacement.

The child scale is similarly refined relative to the parent in log-scale space,

$$\mathbf{s}_c = \mathrm{BoundScale}(\widetilde{\mathbf{s}}_p - \mathrm{softplus}(-\mathbf{r}_s))\,, \tag{5}$$

where $\widetilde{\mathbf{s}}_p$ denotes the hierarchically reduced parent scale. The non-negative $\mathrm{softplus}$ term further refines the scale without increasing it beyond the parent support. Rotation, color, and opacity are inherited from the parent and updated through bounded residuals,

$$\mathbf{q}_c = \mathrm{Bound}(\mathbf{q}_p + \mathbf{r}_q), \qquad \mathbf{c}_c = \mathrm{Bound}(\mathbf{c}_p + \mathbf{r}_c), \qquad o_c = \mathrm{Bound}(o_p + r_o). \tag{6}$$

Together, these transformations define a parent-relative composition operator $\mathcal{C}$, such that each child inherits the local geometry and appearance of its parent while learning only residual refinements.

At the root level, the anchor head transforms the initial 3D points into the first parent states $\mathbf{A}_0$ and Gaussians $\mathbf{G}_0$. At each subsequent level, the previous anchors are expanded to $\widehat{\mathbf{A}}_{\ell-1}$, and the anchor and Gaussian heads produce two sibling refinements from the same parent states:

$$\mathbf{A}_\ell = \mathcal{C}\Big(\widehat{\mathbf{A}}_{\ell-1}, h_\ell^A(\mathbf{F}_\ell)\Big), \qquad \mathbf{G}_\ell = \mathcal{C}\Big(\widehat{\mathbf{A}}_{\ell-1}, h_\ell^G(\mathbf{F}_\ell)\Big). \tag{7}$$

The anchor branch therefore propagates the spatial hierarchy to the next refinement stage, while the Gaussian branch emits the primitives used for rendering. At the finest level, no further anchor state is required and only the Gaussian branch is retained. Finally, the Gaussians produced across all hierarchy levels are combined into the static representation.

3.3 Learning from Imperfect Pseudo-Multiview Supervision

As metioned above, for each training scene, only the frontal portrait is provided as input. The reconstructor predicts a single static Gaussian representation $\mathcal{G}$, which is rendered under the recovered cameras: $\hat{I}^t = \mathcal{R}(\mathcal{G}, C^t)$, where $\mathcal{R}$ denotes differentiable Gaussian rasterization.

Since the frontal observation provides the most reliable supervision, we optimize it using both pixel-level and perceptual reconstruction losses:

$$\mathcal{L}_{\text{front}} = \|\hat{I}^s - I^s\|_1 + \lambda_{\text{vgg}}\mathcal{L}_{\text{VGG}}(\hat{I}^s, I^s). \tag{8}$$

For generated non-frontal views, which may contain local misalignment and synthesis artifacts, we avoid strict pixel-wise supervision and instead combine VGG perceptual supervision with a camera-conditioned adversarial loss. The perceptual term is defined as

$$\mathcal{L}_{\text{mv}} = \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_{\text{VGG}}(\hat{I}^t, I^t). \tag{9}$$

The discriminator $D(I, \psi(C))$ further enforces that the rendered image is realistic under the corresponding recovered camera. For each recovered camera C^t, we treat the corresponding pseudo-multiview observation I^t as the real sample and the rendered image $\hat{I}^t$ as the fake sample. Both are provided to the discriminator together with the same camera embedding $\psi(C^t)$, so that the discriminator learns not only image realism but also whether the appearance is compatible with the specified viewpoint. The discriminator is trained with

$$\mathcal{L}_D = -\mathbb{E}\log D(I^t, \psi(C^t)) - \mathbb{E}\log\Big(1 - D(\hat{I}^t, \psi(C^t))\Big), \tag{10}$$

while the reconstructor is optimized to fool the discriminator under the same camera condition:

$$\mathcal{L}_{\text{adv}} = -\mathbb{E}\log D(\hat{I}^t, \psi(C^t)). \tag{11}$$

The overall objective is

$$\mathcal{L}_G = \lambda_{\text{front}}\mathcal{L}_{\text{front}} + \lambda_{\text{mv}}\mathcal{L}_{\text{mv}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}. \tag{12}$$

4 Experiments

4.1 Experimental Setup

Data and evaluation protocol. We construct EPPM-20K, a large-scale pseudo-multiview dataset for eyeglass-aware portrait reconstruction. It contains 19,996 identities, 39,879 pseudo-multiview scenes, and 665,604 images with estimated camera parameters. For each identity, we generate two scene variants: one original portrait scene without eyeglasses and one eyeglass-wearing portrait scene. We adopt an identity-disjoint split, using 13,996 identities for training and reserving 6,000 identities for evaluation. The dataset will be publicly released to support future research.

We compare our method with recent publicly available head reconstruction methods, including UIKA (Wu et al., 2026), FlexAvatar (Kirschstein et al., 2026), GAGAvatar (Chu & Harada, 2024), FaceLift (Lyu et al., 2025), and LAM (He et al., 2025). We use their released implementations and evaluate all predictions under the corresponding target viewpoints, applying method-specific camera or pose conversions when necessary.

Evaluation metrics. We evaluate novel-view rendering fidelity using PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), and DreamSim (Fu et al., 2023). Identity preservation is measured by the cosine similarity between ArcFace embeddings (Deng et al., 2019), denoted as CSIM. To specifically evaluate eyewear appearance, we further compute feature-space cosine similarities on eye-and-eyeglass regions using DINOv2 (Oquab et al., 2024) and CLIP (Radford et al., 2021), denoted as Glass-DINO and Glass-CLIP. For each prediction–reference pair, eyeglass masks are obtained independently from both images using a pretrained glasses detector (Mantas, 2024). We then crop the corresponding eye-and-eyeglass regions according to the detected masks and compute feature similarities on these localized regions.

Implementation details. Our default model uses a five-level hierarchical Gaussian cascade with 4,096 FLAME-initialized points and predicts 126,976 Gaussians. The model takes a single frontal image as input and is trained with eight target views per scene at 512×512 resolution. All experiments are trained for 20,000 iterations on two NVIDIA A800 80GB GPUs. We optimize the network with AdamW using an initial learning rate of 1×10^{-4} and a cosine learning-rate schedule. At inference, GlassHead requires 4.75 GB of GPU memory and reconstructs a 3D head in under 1 s from a single frontal portrait.

4.2 Comparison with Existing Methods

Table 1 and Figure 4 compare GlassHead with existing single-image head reconstruction methods. The quantitative results show a consistent advantage of GlassHead across image fidelity, perceptual similarity, identity preservation, and eyeglass-specific metrics. The qualitative comparison reveals a clear difference under large viewpoint changes. While existing methods can generally preserve the frontal facial appearance, their oblique renderings often exhibit blurred or distorted facial details, unstable hair geometry, and incomplete eyeglass structures, particularly around the temples and ear-side regions. GlassHead produces substantially more coherent side views, with better preserved facial appearance and more complete eyeglass geometry. More visualization results can be found in supplementary webpage.

Table 1: Novel-view reconstruction on the 100 held-out eyeglass identities. Glass-DINO and Glass-CLIP are cosine similarities computed on independently detected eyeglass regions. Best results are shown in **bold** and second-best results are underlined.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	CSIM ↑	DreamSim ↓	Glass-DINO ↑	Glass-CLIP ↑
UIKA	11.41	0.601	0.402	0.536	0.143	0.738	0.908
FlexAvatar	<u>12.90</u>	0.573	0.463	<u>0.649</u>	<u>0.120</u>	<u>0.788</u>	0.912
GAGAvatar	12.25	<u>0.627</u>	<u>0.347</u>	0.554	0.141	0.779	<u>0.929</u>
FaceLift	10.18	0.563	0.478	0.560	0.149	0.736	0.897
LAM	10.06	0.607	0.398	0.395	0.164	0.777	0.911
Ours	**16.00**	**0.648**	**0.230**	**0.652**	**0.061**	**0.841**	**0.950**

4.3 Real-World Study

To evaluate the generalization of GlassHead beyond our training distribution, we select several eyeglass-wearing subjects from the NeRSemble dataset (Kirschstein et al., 2023) and perform reconstruction from a single frontal frame. As shown in Figure 5, GlassHead can produce consistent novel-view reconstructions from a single frontal image. We found that reconstruction from the raw NeRSemble frames is less stable because their appearance distribution differs from that of our training data. After processing the input with FLUX.2 Klein, the image appearance is better aligned with the training domain, leading to more coherent geometry and appearance across viewpoints, but may also slightly alter facial appearance and introduce mild identity changes. Additional real-world results are provided in Appendix E.

4.4 Ablation Studies

Effect of the GAN loss. Removing the GAN loss improves pixel-level metrics such as PSNR and SSIM, but noticeably degrades perceptual and eyeglass-specific metrics, including LPIPS, Dream-



Figure 4: Qualitative comparison of novel-view reconstruction on test split of EPPM-20K dataset.

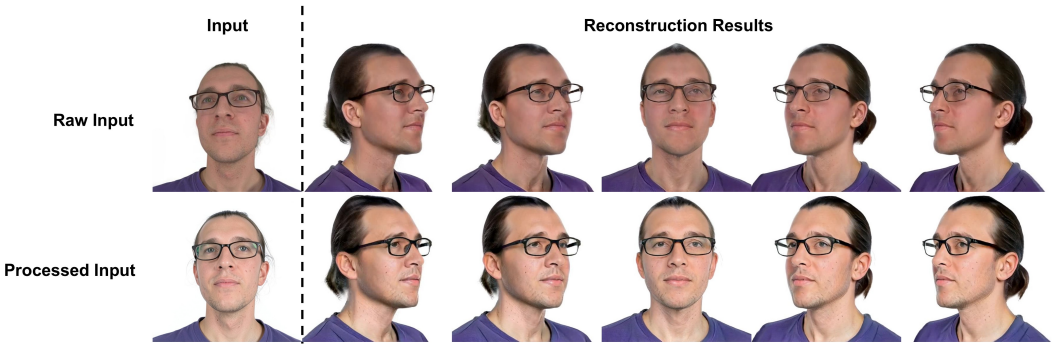

Figure 5: Real-world reconstruction on eyeglass-wearing subjects from NeRSemble.

Sim, Glass-DINO, and Glass-CLIP (Table 2). As illustrated in Figure 6, the model without adversarial supervision tends to lose fine appearance details in the novel views.

Effect of the hierarchical decoder. Replacing the hierarchical Gaussian decoder with a flat dense prediction (see Appendix C) substantially degrades both quantitative and qualitative performance. This highlights the importance of coarse-to-fine hierarchical decoding for stable novel-view reconstruction (Table 2, Figure 6).

Effect of hierarchy depth. We further vary the hierarchy depth from four to six levels while keeping the five-level design as the default. Both alternatives perform slightly worse than the five-level model (Table 2). A shallower hierarchy provides less refinement capacity, whereas increasing the depth does not yield further improvement despite using substantially more Gaussians (Figure 6).

Table 2: Ablation studies on eyeglass novel-view reconstruction.

Variant	# Gaussians	PSNR ↑	SSIM ↑	LPIPS ↓	CSIM ↑	DreamSim ↓	Glass-DINO ↑	Glass-CLIP ↑
w/o GAN	126,976	16.62	0.666	0.285	0.679	0.099	0.759	0.910
w/o Hierarchical Decoder	126,976	14.55	0.615	0.324	0.162	0.193	0.721	0.907
Hierarchical Decoder Depth = 4	61,440	15.66	0.643	0.244	0.592	0.072	0.831	0.946
Hierarchical Decoder Depth = 6	258,048	15.83	0.647	0.236	0.613	0.066	0.836	0.947
Default (Hierarchical Decoder Depth = 5)	126,976	16.00	0.648	0.230	0.652	0.061	0.841	0.950

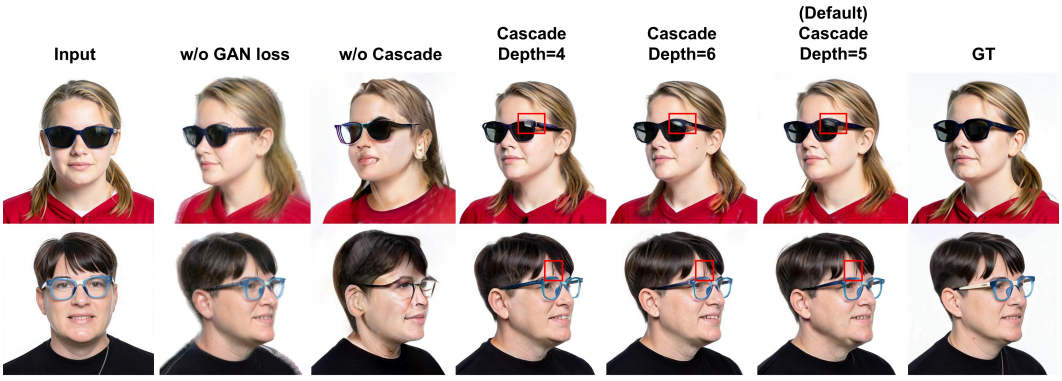

Figure 6: Qualitative results for ablation studies.

4.5 Limitations

Although GlassHead achieves consistent novel-view reconstruction, fine-scale details may not exactly match the target view, as illustrated in Figure 7. In particular, small facial details and local eyeglass structures can be altered. This is partly because our training relies on imperfect pseudo-multiview supervision, which encourages realistic appearance rather than strict pixel-level correspondence. Improving fine-grained geometric and texture consistency remains an important direction for future work.

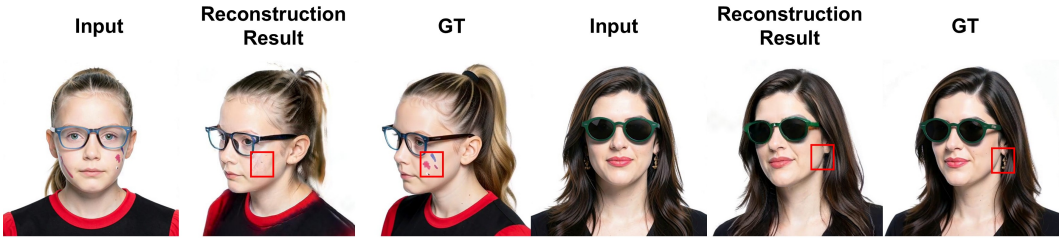

Figure 7: Limitations of GlassHead. GlassHead can reconstruct the overall novel-view structure consistently, but fine-grained facial and eyeglass details may not be fully preserved.

5 Conclusion

We presented GlassHead, a feed-forward framework for single-image 3D reconstruction of eyeglass-wearing heads. To address the lack of scalable supervision for poorly observed eyeglass structures, we introduced a video-generation-based pipeline for constructing pseudo-multiview training data and built the EPPM-20K dataset. GlassHead further employs a hierarchical Gaussian decoder to progressively recover head geometry and appearance from coarse to fine. Experiments demonstrate consistent improvements over existing single-image head reconstruction methods. Our results show that generated pseudo-multiview data can provide effective supervision for learning complete 3D eyeglass-aware head representations from a single portrait. The proposed pseudo-multiview data construction paradigm has the potential to be extended to other reconstruction settings where important structures are weakly observed and large-scale multiview data are difficult to acquire.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic portraits and pseudo-multiview videos as part of the proposed data construction pipeline, as described in the paper. We also used generative AI tools for language editing and improving the clarity of the manuscript. We did not use generative AI tools to generate or modify experimental results, citations, or scientific conclusions. All AI-generated training data were processed and filtered using the procedures described in the paper, and all AI-assisted text was reviewed and revised by the authors. References, experimental results, and technical claims were manually verified. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve direct interaction with human subjects or the collection of new personal data. EPPM-20K is constructed from publicly available portrait data and generative models, and will be released in accordance with the licenses and usage conditions of the underlying data and models. Since the data contain human faces, potential concerns include privacy, identity changes introduced by generation, and possible misuse of face reconstruction technology. We therefore limit the intended use to research and benign applications such as 3D portrait reconstruction and virtual try-on, and encourage responsible use consistent with applicable privacy and data protection requirements.

REFERENCES

Florian Barthel, Wieland Morgenstern, Paul Hinzer, Anna Hilsmann, and Peter Eisert. CGS-GAN: 3d consistent gaussian splatting gans for high resolution human head synthesis. *arXiv preprint arXiv:2505.17590*, 2025. doi: 10.48550/arXiv.2505.17590. URL https://arxiv.org/abs/2505.17590.

Black Forest Labs. FLUX.2 [klein]: Towards interactive visual intelligence. https://bfl.ai/blog/flux2-klein-towards-interactive-visual-intelligence, January 2026.

Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 187–194, 1999. doi: 10.1145/311535.311556.

Marcel C. Buehler, Gengyan Li, Erroll Wood, Leonhard Helminger, Xu Chen, Tanmay Shah, Daoye Wang, Stephan Garbin, Sergio Orts-Escolano, Otmar Hilliges, Dmitry Lagun, Jérémy Riviere, Paulo Gotardo, Thabo Beeler, Abhimitra Meka, and Kripasindhu Sarkar. Cafca: High-quality novel view synthesis of expressive faces from casual few-shot captures. In *ACM SIGGRAPH Asia 2024 Conference Paper*. 2024. doi: 10.1145/3680528.3687580. URL https://doi.org/10.1145/3680528.

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. TOPIQ: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024. doi: 10.1109/TIP.2024.3378466.

Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6a14c7f9fb3f42645cfa6bd5aa446819-Abstract-Conference.html.

Xuangeng Chu, Yu Li, Ailing Zeng, Tianyu Yang, Lijian Lin, Yunfei Liu, and Tatsuya Harada. Gpavatar: Generalizable and precise head avatar from image (s). *arXiv preprint arXiv:2401.10215*, 2024.

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. doi: 10.1109/CVPR.2019.00482.

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*, 40(4):1–13, 2021. doi: 10.1145/3450626.3459936.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, volume 36, pp. 50742–50768, 2023.

Yujie Gao, Yao Xiao, Xiangnan Zhu, Ya Li, Yiyi Zhang, Liqing Zhang, and Jianfu Zhang. Any3davatar: Fast and high-quality full-head 3d avatar reconstruction from single portrait image. *arXiv preprint arXiv:2604.13856*, 2026. doi: 10.48550/arXiv.2604.13856. URL https://arxiv.org/abs/2604.13856.

Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z. Li. Towards fast, accurate and stable 3d dense face alignment. In *European Conference on Computer Vision*, 2020. doi: 10.1007/978-3-030-58529-7_10. URL https://arxiv.org/abs/2009.09960.

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 *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–13, 2025.

Charlie Hewitt, Fatemeh Saleh, Sadegh Aliakbarian, Lohit Petikam, Shideh Rezaeifar, Louis Florentin, Zafiirah Hosenie, Thomas J Cashman, Julien Valentin, Darren Cosker, and Tadas Baltrušaitis. Look ma, no markers: holistic performance capture without the hassle. *ACM Transactions on Graphics (TOG)*, 43(6), 2024.

Sangeek Hyun and Jae-Pil Heo. Gsgan: Adversarial learning for hierarchical generation of 3d gaussian splats. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. doi: 10.1145/3592433.

Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics*, 42(4):1–14, 2023. doi: 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12089–12100, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Kirschstein_Avat3r_Large_Animatable_Gaussian_Reconstruction_Model_for_High-fidelity_3D_Head_ICCV_2025_paper.html.

Tobias Kirschstein, Simon Giebenhain, and Matthias Nießner. Flexavatar: Learning complete 3d head avatars with partial supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18193–18203, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Kirschstein_FlexAvatar_Learning_Complete_3D_Head_Avatars_with_Partial_Supervision_CVPR_2026_paper.pdf.

Junxuan Li, Shunsuke Saito, Tomas Simon, Stephen Lombardi, Hongdong Li, and Jason Saragih. MEGANE: Morphable eyeglass and avatar network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12769–12779, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Li_MEGANE_Morphable_Eyeglass_and_Avatar_Network_CVPR_2023_paper.html.

Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Transactions on Graphics*, 36(6):194:1–194:17, 2017. doi: 10.1145/3130800.3130813.

Weijie Lyu, Yi Zhou, Ming-Hsuan Yang, and Zhixin Shu. Facelift: Learning generalizable single image 3d face reconstruction from synthetic heads. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12691–12701, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Lyu_FaceLift_Learning_Generalizable_Single_Image_3D_Face_Reconstruction_from_Synthetic_ICCV_2025_paper.html.

Mantas. Glasses detector: Glasses detection, classification and segmentation. https://github.com/mantasu/glasses-detector, 2024. Accessed: 2026-09-24.

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.

MiniMax. MiniMax H3: An open model breaking the boundaries between tasks and modalities. https://www.minimax.io/blog/minimax-h3, July 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Dongwei Pan, Long Zhuo, Jingtan Piao, Huiwen Luo, Wei Cheng, Yuxin Wang, Siming Fan, Shengqi Liu, Lei Yang, Bo Dai, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, and Kwan-Yee Lin. Renderme-360: A large digital asset library and benchmarks towards high-fidelity head avatars. *Advances in Neural Information Processing Systems*, 36, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2074–2083, 2018. doi: 10.1109/CVPRW.2018.00281.

Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. Hunyuanvideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025.

Zijian Wu, Boyao Zhou, Liangxiao Hu, Hongyu Liu, Yuan Sun, Xuan Wang, Xun Cao, Yujun Shen, and Hao Zhu. UIKA: Fast universal head avatar from pose-free images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. doi: 10.48550/arXiv.2601.07603. URL https://arxiv.org/abs/2601.07603.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision*, pp. 250–269, 2022. doi: 10.1007/978-3-031-19778-9_15. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136730249.pdf.

A Details of the Dataset Generation Pipeline

This supplementary material provides additional details on the construction of EPPM-20K, camera recovery, the GlassHead architecture and training setup, and additional experiments.

A.1 Portrait Standardization and Eyeglass Synthesis

We construct paired original and eyeglass-wearing portraits in two stages. First, each source image is standardized into a clean frontal portrait using FLUX.2 Klein. The prompt used for portrait standardization is:

> *A 4k HD candid photo of the same person, centered composition, clean head-and-shoulders portrait, remove other people, pure white background, natural shoulder posture, only one person visible, no other people, no obstructions, highly realistic hairstyle, subtle and highly detailed hair strands, photorealistic, keep original facial features intact, maintain identity, maintain expression, enhance this person using super-resolution, natural skin, low saturation, no retouching, photographed with a DSLR, natural lighting, unstylized, high-resolution, true-to-life, studio-style portrait, remove unnatural edges.*

In the second stage, we synthesize the eyeglass-wearing portrait by using the standardized portrait as the immutable identity reference and an eyeglass product image as the accessory reference. The prompt is designed to restrict the edit to the eyewear region only, while preserving all other portrait content. The prompt used for eyeglass synthesis is:

> *STRICT SINGLE-EDIT RULE: Image 1 is immutable; Image 2 defines only the accessory. The only allowed change is adding that accessory at its anatomical location. Keep everything else in Image 1 exactly unchanged, including but not limited to: canvas, composition, crop, perspective, camera angle, background and every background detail; person count, identity, age, ethnicity, skin tone/texture; body build, shoulders, head/face shape and proportions; body pose and head yaw/pitch/roll; expression, emotion and gaze; eye size/shape/opening, eyelids, iris/pupil positions and catchlights; eyebrows, nose, cheeks, ears, lips, mouth/teeth and facial hair; hairstyle, hairline and visible hair; clothing, collar and neckline; lighting, exposure, colors and existing shadows. Never redraw, retouch, beautify, smooth, age, move, rotate, resize, reframe, relight or restyle the person/scene, and never alter a non-target object. Permit only accessory pixels, necessary local occlusion and a narrow contact shadow. Image 1 pixels override text. Never copy Image 2's background, stand, mannequin or camera view.*
> *TARGET: Copy Image 2's exact category/subtype, count, silhouette, geometry, proportions, construction, color, material, texture, pattern, decoration, hardware, transparency, gloss and distinguishing details. Change only its rotation/foreshortening to fit Image 1's fixed pose. Never invent, simplify, replace, recolor, duplicate, mirror or omit details. Image 2 appearance overrides its text description.*
> *GLASSES CONSTRAINTS: Fit them to the existing nose bridge/ears. Match lens outline/size, rim style/thickness, bridge, nose pads, hinges, end pieces, temples, frame color/material. Reproject glasses to the fixed head yaw/pitch/roll; never turn the face.*

Keep eyes, eyelids, irises, pupils, gaze, opening and catchlights pixel-faithful to Image 1. CLEAR-LENS RULE: White visible inside Image 2's lens outlines is only the white product-photo background, not lens color or reflection; never copy it onto the face. Treat lenses as visually empty: lens interiors equal Image 1 pixels exactly, with no tint, wash, glare, highlight or reflection. Add only frame, nose pads and tiny outer lens edges.

The output of this process is a pair of identity-consistent frontal portraits, one without eyeglasses and one with synthesized eyeglasses.

A.2 PHONE-SCAN VIDEO GENERATION

Starting from each frontal portrait, we generate a short phone-scan-style video using MiniMax-H3, a video generation model. Rather than asking the subject to rotate their head, we explicitly formulate the sequence as virtual camera motion while keeping the head approximately static, thereby capturing the subject from diverse viewpoints. The prompt used for phone-scan generation is:

For the target video, at 0.00 seconds into the target video, Picture 1 (from [Shot 1]) is fully referenced.
integrated_multimodal_description: [Shot 1] Live-action, realistic 7-second smartphone head-scan video in one continuous shot. The person shown in Picture 1 remains perfectly still throughout, preserving the exact identity, head pose, expression, gaze, hairstyle, clothing, lighting, and background. Only the camera moves. The camera follows a smooth, wide, front-facing orbit around the center of the face while maintaining approximately constant distance from the subject. It always points toward the facial center and keeps the head centered in frame. From 0.00 to 1.30 seconds, the camera moves from the front-center position to a left sideview. From 1.30 to 2.60 seconds, it moves diagonally upward to an upper-front view, approximately half a head height above eye level. From 2.60 to 3.90 seconds, it moves diagonally down-right to a right side view. From 3.90 to 5.20 seconds, it moves diagonally down-left to a lower-front view, approximately half a head height below eye level. From 5.20 to 6.50 seconds, it moves diagonally up-left back to the left side view. From 6.50 to 7.50 seconds, the camera holds goes back to center then to right side view. Keep the head centered with small pan and tilt corrections. The camera remains entirely within the front hemisphere, never passing behind the ears or moving to the rear. Allow natural three-quarter views and clear perspective changes, but never show a full profile or rear view. No head or body movement, blinking, expression change, zoom, cuts, camera shake, motion blur, distortion, identity drift, or background deformation. Keep background pure white.
$overall_soundscape : N/A.$
$non_diegetic_music : N/A.$

Examples of the generated videos are available on the supplementary webpage.

A.3 CAMERA PARAMETERS ESITMATION

Generated videos provide multiple views of the same subject but do not contain camera parameters. We therefore recover a temporally coherent camera trajectory by jointly fitting a shared FLAME model (Li et al., 2017) to each sequence. For every frame, we detect 68 facial landmarks using 3DDFA-V2 (Guo et al., 2020). The first frame is used to initialize the shared FLAME parameters: MICA (Zielonka et al., 2022) estimates the identity code $\boldsymbol{\beta}$, which is kept fixed during fitting, while DECA (Feng et al., 2021) initializes the expression code $\boldsymbol{\psi}_0$ and jaw pose $\boldsymbol{\omega}_{j,0}$. The shared global head rotation $\boldsymbol{\omega}_h$ is initialized to zero.

Since the generated subject remains approximately static, the identity, expression, global head rotation, jaw pose, and field of view are shared across the sequence. Each frame t is assigned an independent 6-DoF camera pose

$$\mathbf{p}_t = \left[\boldsymbol{\omega}_t^\top, \mathbf{t}_t^\top\right]^\top, \tag{13}$$

where $\boldsymbol{\omega}_t$ and $\mathbf{t}_t$ denote camera rotation and translation, respectively.

We jointly optimize the shared facial parameters, field of view, and per-frame camera poses using

$$\mathcal{L}_{\text{cam}} = \lambda_{\text{lm}}\mathcal{L}_{\text{lm}} + \mathcal{L}_{\text{vel}} + \mathcal{L}_{\text{acc}} + \mathcal{L}_{\text{prior}}. \tag{14}$$

Here, $\mathcal{L}_{\mathrm{lm}}$ aligns the projected FLAME landmarks with the detected 2D landmarks. $\mathcal{L}_{\mathrm{vel}}$ penalizes first-order differences between consecutive camera poses, while $\mathcal{L}_{\mathrm{acc}}$ penalizes second-order differences to suppress abrupt changes in the recovered trajectory. The prior term is defined as

$$\mathcal{L}_{\mathrm{prior}} = \lambda_{\psi} \left\|\boldsymbol{\psi} - \boldsymbol{\psi}_0\right\|_2^2 + \lambda_h \left\|\boldsymbol{\omega}_h - \boldsymbol{\omega}_{h,0}\right\|_2^2 + \lambda_j \left\|\boldsymbol{\omega}_j - \boldsymbol{\omega}_{j,0}\right\|_2^2 + \lambda_f (\Delta f)^2, \tag{15}$$

where $\boldsymbol{\psi}$, $\boldsymbol{\omega}_h$, and $\boldsymbol{\omega}_j$ denote the optimized shared expression, global head rotation, and jaw pose, respectively; Δf denotes the correction to the initial field of view. This regularization prevents the fitting procedure from explaining camera motion through implausible changes in the shared facial state.

We optimize each sequence for 1,000 iterations and constrain the field of view to $10°$–$60°$. Finally, the recovered camera trajectory is smoothed using a one-dimensional Gaussian filter with $\sigma = 0.7$ to remove residual temporal jitter.

B EPPM-20K Dataset Details

EPPM-20K contains 19,996 identities and 39,879 usable pseudo-multiview scenes, comprising 665,604 images with estimated camera parameters. Among them, 19,883 scenes correspond to eyeglass-wearing portraits, containing 320,219 views. Each scene consists of one frontal source portrait and multiple selected views with recovered cameras (Figure 8). As shown in Figure 9, the dataset covers a broad range of viewpoints.

C Model Details about GlassHead and its Variants

GlassHead takes a single frontal portrait and predicts a static 3D Gaussian head representation in one forward pass. The image encoder combines trainable RGB patch features with frozen Sapiens2-1B and FLUX.2 VAE features. These three feature streams are spatially aligned, concatenated, and projected to a shared feature dimension. For the DINO-based variant, we replace the Sapiens2-1B features extracted from blocks 9, 19, 29, and 39 with frozen DINOv3 ViT-L/16 features extracted from blocks 4, 11, 17, and 23, while keeping the remaining reconstruction architecture unchanged. The details of the GlassHead architecture are listed in Table 3.

To evaluate the effect of hierarchical decoding, we construct a dense prediction baseline that removes the parent–detail Gaussian hierarchy and directly predicts a flat set of 126,976 Gaussians. Unlike our hierarchical model, the dense variant does not maintain intermediate parent anchors or perform coarse-to-fine feature inheritance. Instead, the image-conditioned features are decoded directly into Gaussian attributes for the complete output set. We match the total number of output Gaussians to the five-level model so that the comparison primarily examines the effect of hierarchical organization rather than representation capacity.

D Results on Portrait Reconstruction without Eyeglasses

Although our method focuses on eyeglass-wearing portraits, we additionally evaluate reconstruction on the corresponding original portraits without eyeglasses (Figure 10). As shown in Table 4, GlassHead achieves the best PSNR, SSIM, LPIPS, and DreamSim among the evaluated methods, while FlexAvatar obtains the highest ArcFace identity similarity. These results indicate that GlassHead retains comparable reconstruction capability on general portraits without eyeglasses.

E More Results on Real-world Study

Figure 11 presents additional real-world reconstruction results on eyeglass-wearing subjects. Video visualizations of the reconstructed heads are provided on the supplementary website.



Figure 8: Examples of generated pseudo-multiview eyeglass portraits in EPPM-20K. The generated sequences cover a wide range of viewpoints while preserving identity, facial appearance, hairstyle, and eyeglass attributes.

F More Ablation Studies

F.1 Encoder Ablation

We compare FLUX2 VAE, DINOv3, DINOv3+VAE, Sapiens2 and our default Sapeins2+VAE image representations while keeping the reconstruction architecture unchanged. Adding VAE features improves perceptual reconstruction quality, while replacing DINOv3 with Sapiens2 further improves identity and eyeglass-region consistency (Table 5, Figure 12).

F.2 Initialization Strategy Ablation

We compare the default FLAME-based initialization with randomly sampled points on a sphere. As shown in Table 6 and Figure 13, the two strategies produce nearly identical reconstruction quality

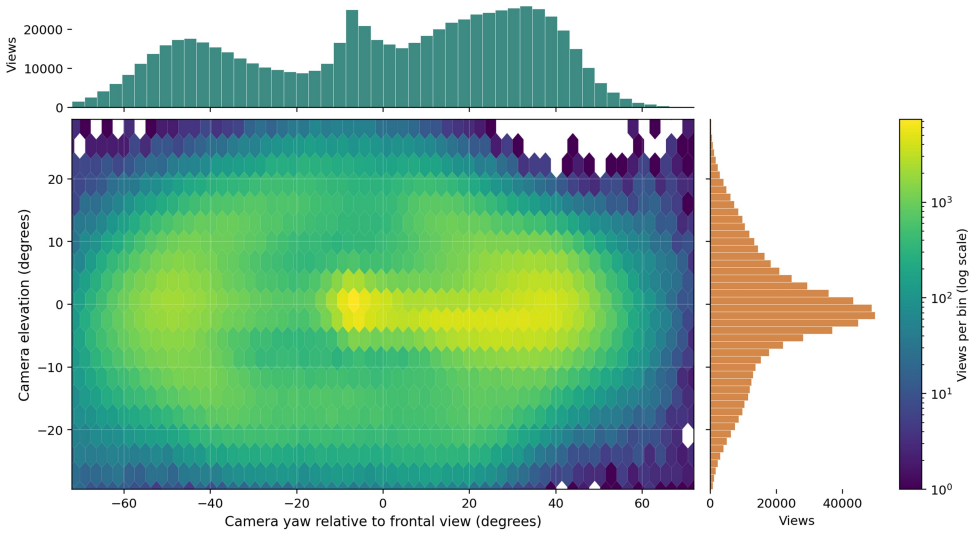

Figure 9: Viewpoint distribution of EPPM-20K.

Table 3: Hyperparameters of GlassHead architecture.

	Hyperparameter	**Value**
Input & Output	Input image resolution	512×512
	Render resolution	512×512
Image Encoder	RGB patch size	16×16
	RGB token number	$32 \times 32 = 1,024$
	Sapiens2 backbone	Sapiens2-1B
	Sapiens2 feature blocks	9, 19, 29, 39
	FLUX.2 VAE latent patch size	2×2
	Fused feature dimension	$6,784 \rightarrow 512$
3D Point Encoder	Initial point number	$4,096$
	Point feature dimension	512
Joint Transformer	Transformer blocks	10
	Hidden dimension	512
	Attention heads	8
	Feed-forward dimension	2,048
Gaussian Output	Total Gaussians	126,976

across all metrics and visual examples. This suggests that GlassHead is largely insensitive to the initialization strategy and does not depend strongly on the FLAME geometry prior.

Table 4: Quantitative comparison of novel-view reconstruction on test identities without eyeglasses. Best results are shown in **bold**, and second-best results are underlined.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	CSIM ↑	DreamSim ↓
UIKA	11.59	0.613	0.398	0.600	0.140
FlexAvatar	13.08	0.588	0.445	**0.733**	0.112
GAGAvatar	12.74	0.654	0.319	0.656	0.136
FaceLift	10.66	0.585	0.450	0.644	0.132
LAM	10.46	0.618	0.393	0.490	0.161
Ours	**17.33**	**0.680**	**0.198**	0.706	**0.055**



Figure 10: Qualitative comparison of novel-view reconstruction on portraits without eyeglasses.

Table 5: Comparison of different image encoder and latent feature combinations on novel-view reconstruction. Best results are shown in **bold**, and second-best results are underlined.

Variant	PSNR ↑	SSIM ↑	LPIPS ↓	CSIM ↑	DreamSim ↓	Glass-DINO ↑	Glass-CLIP ↑
FLUX2 VAE	14.87	0.627	0.276	0.386	0.097	0.815	0.941
DINOv3	15.74	0.641	0.241	0.566	0.069	0.838	0.946
DINOv3+VAE	15.89	0.646	0.233	0.621	0.063	0.837	0.948
Sapiens2	15.97	0.647	**0.229**	**0.669**	**0.060**	0.838	0.948
Sapiens2+VAE(Default)	**16.00**	**0.648**	0.230	0.652	0.061	**0.841**	**0.950**

Table 6: Comparison of FLAME-based and random point initialization on novel-view reconstruction. Best results are shown in **bold**.

Variant	PSNR ↑	SSIM ↑	LPIPS ↓	CSIM ↑	DreamSim ↓	Glass-DINO ↑	Glass-CLIP ↑
FLAME init (Default)	16.00	0.648	0.2296	0.652	**0.061**	**0.841**	**0.950**
Random init	**16.01**	**0.652**	**0.2286**	**0.654**	0.062	0.840	0.949

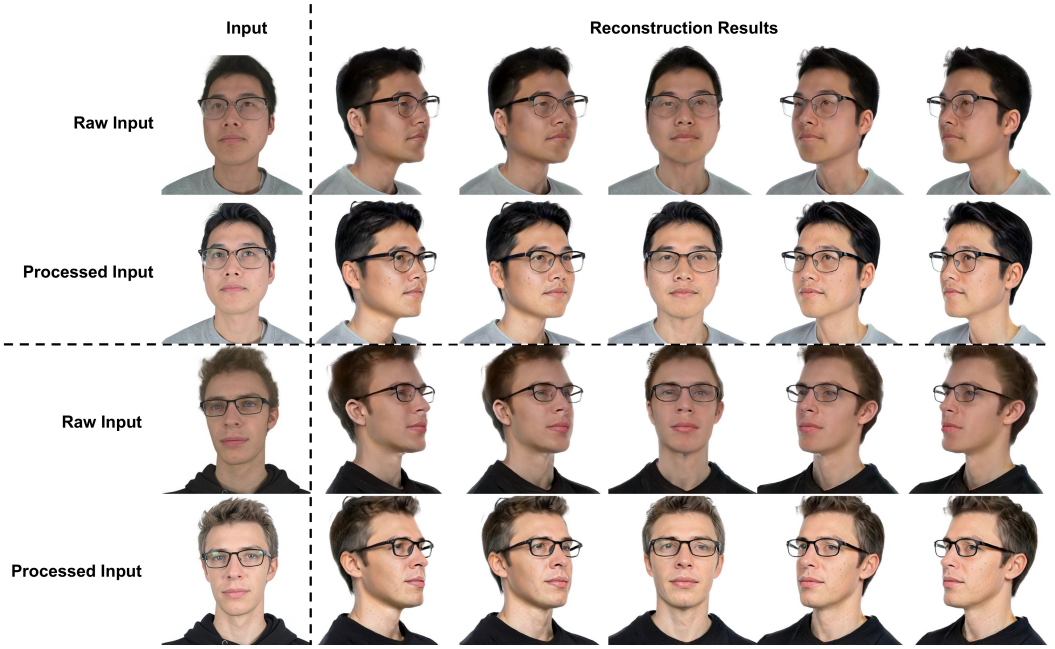

Figure 11: Real-world reconstruction on eyeglass-wearing subjects from NeRSemble.



Figure 12: Qualitative comparison of different image encoder variants.

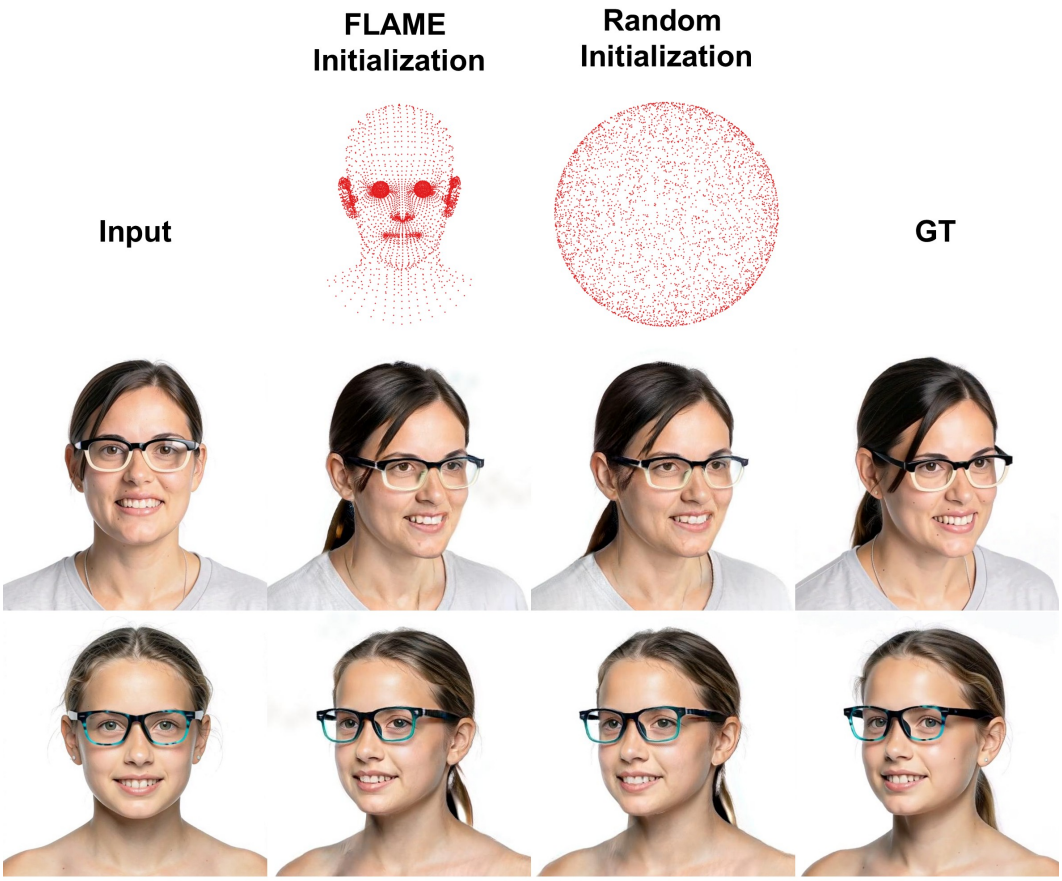

Figure 13: Qualitative comparison of different initialization strategy.