UIKA: Fast Universal Head Avatar from Pose-Free Images

Supplementary Material

In this supplementary material, we first provide additional implementation details for our model, along with visualizations of components (Sec. A). We then present and evaluate our synthetic dataset (Sec. B). Next, we report additional comparative experiments and a user study, covering both self and cross reenactment on monocular and multi-view settings (Sec. C). We also include extended ablation studies, examining the impact of training data size as well as ablations of our method itself (Sec. D). We then provide more in-the-wild cases and applications, e.g., text-to-head-avatar generation (Sec. E). Finally, we discuss the limitations of our method (Sec. F) and its associated ethical implications (Sec. G). Additional dynamic results are provided in our supplementary video.

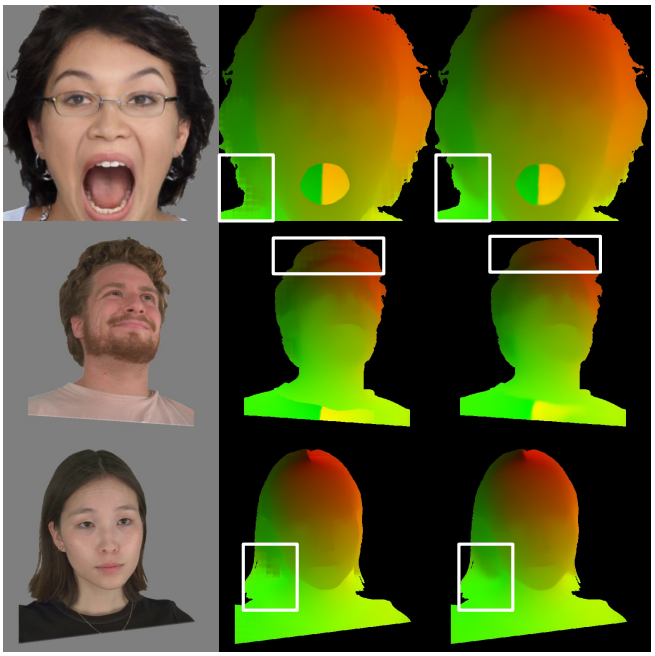
A. Additional Implementation Details

A.1. Facial Correspondence Estimator

As illustrated in Fig. S2, our facial correspondence estimator network consists of three main components: a frozen feature extractor, a trainable alternating attention module, and a trainable UV decoding head. We first extract patch-wise features using DINOv3 ViT-B/16 [14], which serves as a powerful pre-trained visual backbone. The extracted features are processed through *four* alternating attention layers. Following VGGT [16], *Frame Attention* first captures intra-frame spatial relationships within each individual image by computing *self attention* across patches of the same frame. *Global Attention* then establishes inter-frame correspondences by attending to tokens across all input frames simultaneously. This hierarchical attention design enables the network to jointly reason about local facial structures and global multi-view consistency. On top of it, we initialize a trainable DPT [12, 13] head to predict two-channel UV coordinates within the range $[0, 1]$. The predicted UV coordinates map is further multiplied by the input image mask to extract the valid foreground region of the human head.

A.2. UV coordinates map

We visualize the predicted UV coordinates map and compare them with Pixel3DMM [5]. As shown in Fig. S1, our approach produces significantly smoother results in the boundary regions, particularly around the hair. This smoothness is crucial for our subsequent reprojecting of screen-space color back into UV space, enabling more coherent and reliable reprojection results.



(a) Input (b) Pixel3DMM (c) Ours

Figure S1. **Visualization and Comparison of UV coordinates map.**

	Hyperparameter	Value
Input & Output	Input image resolution	512×512
	Train render resolution	512×512
Feature Extractor	DINOv3 version	vitl16
	DINOv3 patch size	16×16
	DINOv3 feature size	$\mathcal{N} \times 1024 \times 1024$
	DINOv3 intermediate layer	4, 11, 17, 23
MultiModal Transformer	Hidden dimension	1024
	Head numbers	16
	Self attention layers	12
	Learnable UV token size	$96 \times 96 \times 1024$
UV Gaussian Decoder	Gaussian attribute map size	384×384
	Aggregated UV map size	384×384
	UV DPT inner dimension	256
	MLP inner dimension	512
	MLP layers	3
	MLP activation	SiLU
Gaussian Settings	Offset max range	0.2
	Scaling clip range	0.01
	Init scaling	exp(-5.0)
	Init density	0.1

Table S1. **Hyperparameters used in our method.** $\mathcal{N}$ represents the number of input views.

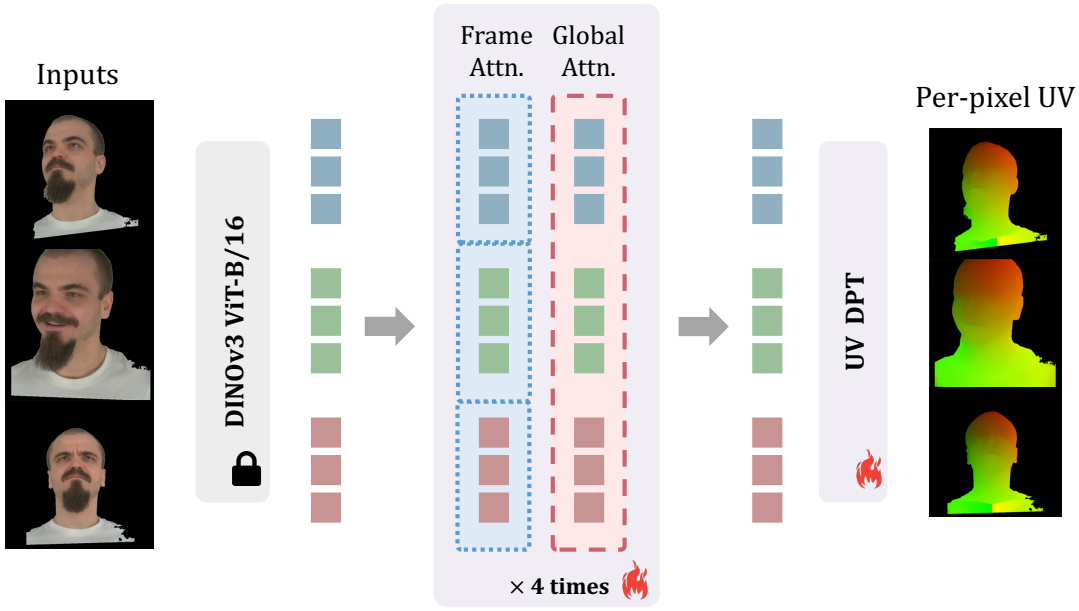


Figure S2. **Architecture of our facial correspondence estimator network.**

Input number	1	2	4	8	16	32
V-D Latency (s)	1.96	2.51	3.57	6.02	12.8	32.9

Table S2. **Latency analysis for view-dependent (V-D) modules.** We show running time for different number of input images.

Real / Synthetic Datasets	NeRSemble-v2	Ours	CAP4D
Spatial WE ($\times 10^{-2}$) ↓	2.377	4.252	10.45
Temporal WE ($\times 10^{-4}$) ↓	4.605	7.868	31.27

Table S3. **Data quality evaluation.** We use Warping Error(WE) as metric to evaluate spatial and temporal consistency of datasets.

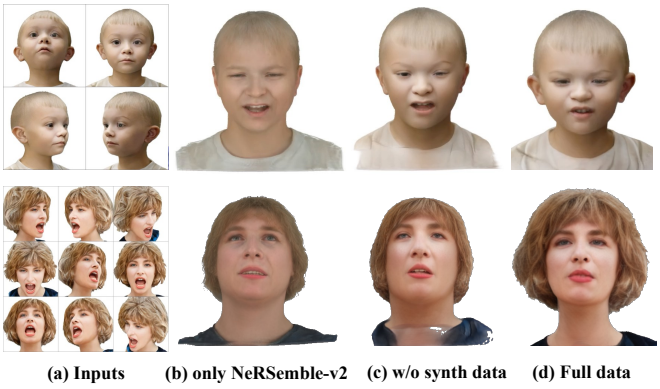


Figure S3. **Ablation study on identity robustness across different training data configurations.**

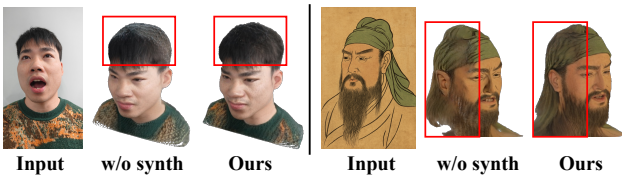


Figure S4. **Visual analysis of 3D consistency under varying training data settings.**

A.3. Hyperparameters

In Tab. S1, we provide additional detailed hyperparameters used in our model configuration.

A.4. Latency analysis.

In general, one pass inference consists of view-dependent (V-D) and view-independent (V-I) components. As shown in Tab. S2, the latency of V-D components (UV prediction, Transformer, and decoder) scales in $O\left(N^2\right)$ with the number of input images inherent to the self-attention mechanism. Once the canonical Gaussian avatar is obtained, the V-I module takes 3ms for LBS and 2ms for rendering. The total latency is in second level.

B. Synthetic Dataset

Visualization. In Sec. 3.4 of the main paper, we explain the curation process of our synthetic multi-view head dataset. In this section, we provide visualization results of

Method	InvertAvatar [18]	DiffusionRig [4]	GPAvatar [2]	LAM [6]	GAGAvatar [1]	Portrait4D-v2 [3]	Ours
Render Quality ↑	1.4	1.95	2.7	2.51	3.29	3.48	**4.37**
Motion Consistency ↑	1.85	2.05	2.73	2.93	3.45	3.4	**4.17**
Identity Preservation ↑	1.7	2.15	2.74	2.54	3.44	3.54	**4.23**

Table S4. **User study evaluation.** We ask users to rate the results in 1-5, the higher the better.

this dataset, as shown in Fig. S5, which illustrates the results of each identity under different camera viewpoints and expressions.

Quality assessment of our dataset. We evaluate our synthetic dataset in Tab. S3 by using warping error as in HuGe100K [19]. Our generated dataset achieves numeric results comparable to the real-captured dataset NeRSemble [7] and outperforms a synthetic dataset using CAP4D [15] in both spatial and temporal dimensions.

Our synthetic data achieves a well-balanced combination of identity diversity and expression richness, while maintaining multi-view and 3D consistency. Such a dataset contributes to training a more robust model. Please refer to our supplementary video for additional dynamic results.

C. Additional Comparison Results

Monocular Setting. In Fig. S6, we show more self and cross reenactment results on the VFHQ dataset and NeRSemble-v2 dataset.

Multi-view Setting. In Fig. S7, we show more results of self and cross reenactment on the NeRSemble-v2 dataset. Please refer to our supplementary video for additional dynamic results.

User Study. We have also included a human evaluation in Tab. S4 as an additional validation. Our method outperforms baselines in render quality, motion consistency, and identity preservation.

D. Additional Ablation Results

D.1. Ablation on training data

Thanks to the paradigm of our framework, the model can accept an arbitrary number of input images. Although the number of input views during training is limited to $1 \sim 16$ due to VRAM constraints, similar to VGGT [16], our model can take more than 16 input images during inference. This flexibility enables us to train on monocular video datasets, unlike methods such as Avat3r [8] that require a fixed set of four input views and therefore rely exclusively on multi-view datasets. The monocular video dataset VFHQ [17] contains approximately 7k identities, which is an order of magnitude larger than existing multi-view datasets such as NeRSemble-v2 [7], Ava-256 [10], and RenderMe-360 [11], each of which typically includes only a few hundred identities.

To evaluate the importance of high-quality training data, we prepare two ablated versions. One model is only trained on the NeRSemble-v2 dataset, as shown in Fig. S3 (b), which can hardly preserve the identity of input images. When using both NeRSemble-v2 and a rich-identity dataset VFHQ, the model generalizes better to novel identities, but would collapse in some extreme viewpoint in Fig. S3 (c). When including our multi-view synthetic data, our model demonstrates superior generalization capability of identity as shown in Fig. S3 (d). As shown in Fig. S4, our model achieves better 3D consistency using our synthetic dataset.

D.2. Ablation on our method

In this section, we provide additional visualizations of ablation studies on our method. Other than the ablated versions in the main paper, we further include an extra ablation on our self-adaptive fusion strategy, as shown in Fig. S8 (d). In our full model, the fusion weight for each Gaussian is predicted by the network as a per-Gaussian value in the range $[0, 1]$. In contrast, this ablated variant replaces the learned weight with a fixed value computed as 0.5 times the UV-domain confidence map described in Sec. 3.1. The results demonstrate that our proposed full model effectively leverages information from the input views, leading to higher-fidelity head avatar reconstruction. Please refer to our supplementary video for additional dynamic results.

E. Applications

In-the-wild Image Reenactment. We also demonstrate the reenactment results of our method on in-the-wild Internet cases, as shown in Fig. S9.

Text-to-Head-Avatar Generation. In addition, we visualize the pipeline for generating controllable head avatars from text prompts. Given a textual description, we employ advanced multimodal large models such as ChatGPT or Gemini to synthesize corresponding images, which are then fed into our model to produce a animatable head avatar. Detailed visualizations are provided in Fig. S10.

Such results show that our method generalizes well to a wide variety of visual styles, benefiting from both our proposed approach and the synthetic dataset. Please refer to our supplementary video for additional dynamic results.

F. Limitations

Despite its effectiveness, our approach has several limitations. First, the expressiveness of our reconstructed head avatars is inherently constrained by the FLAME [9] model used for both data tracking and avatar driving. As a result, fine-grained facial dynamics such as subtle wrinkles, micro-expressions, and tongue motions cannot be reliably captured or reproduced. Second, although our training includes both real and synthetic data, the combined dataset still exhibits certain demographic biases, which may lead to degraded performance or failure cases for underrepresented groups. Third, while our framework supports an arbitrary number of input images, the computational cost and memory consumption grow with the number of views, whereas the performance improvement saturates beyond a certain point. These limitations highlight important directions for future work, such as integrating more expressive parametric models, reducing data bias, and improving scalability for large-view inference.

G. Ethics

Our work focuses on feed-forward reconstruction of animatable head avatars from arbitrary numbers of input facial images. While the proposed method advances the efficiency and accessibility of personalized head avatar creation, it also raises several potential ethical concerns. First, the ability to reconstruct high-fidelity 3D human heads from sparse or casually captured images introduces risks of misuse, such as generating unauthorized digital replicas of individuals or producing manipulated content that may compromise privacy, consent, or identity integrity. Second, reconstructed avatars could be misappropriated for malicious applications, including impersonation, deepfake-style synthesis, or other forms of deceptive media generation.

To mitigate these risks, our research uses only publicly available datasets with established licenses and synthetic data generated in-house. We emphasize that our method is intended for legitimate applications such as virtual telepresence, animation, and human computer interaction. We strongly discourage any use of this technology for surveillance, non-consensual persona reproduction, or deceptive content creation. Future deployment of systems built upon our approach should incorporate suitable safeguards, such as perceptual watermarking, provenance tracking, or identity verification mechanisms, to ensure responsible and ethical use.

Figure S5. **Visualization of examples from our synthetic dataset.**

Figure S6. **Visualization of self and cross reenacted results on the VFHQ and NeRSemble-v2 datasets for the monocular input setting.**

Figure S7. **Visualization of self and cross reenacted results on the NeRSemble-v2 dataset for the multi-view setting.**

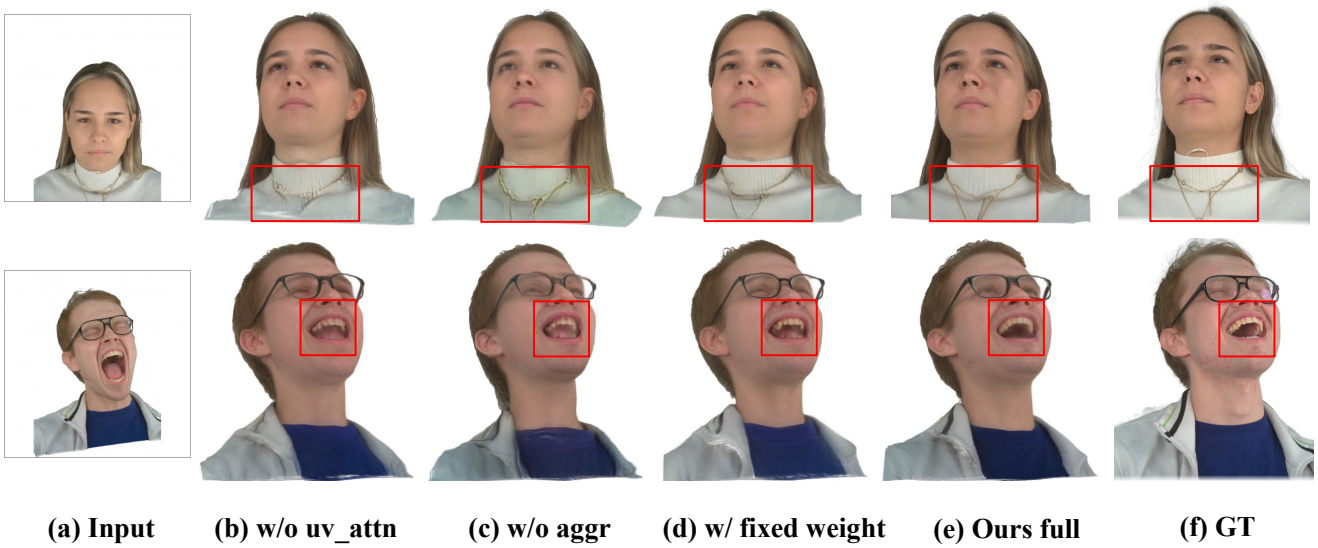


Figure S8. **Visualization of ablation study results of our method.**

Figure S9. **Visualization of in-the-wild cases.**

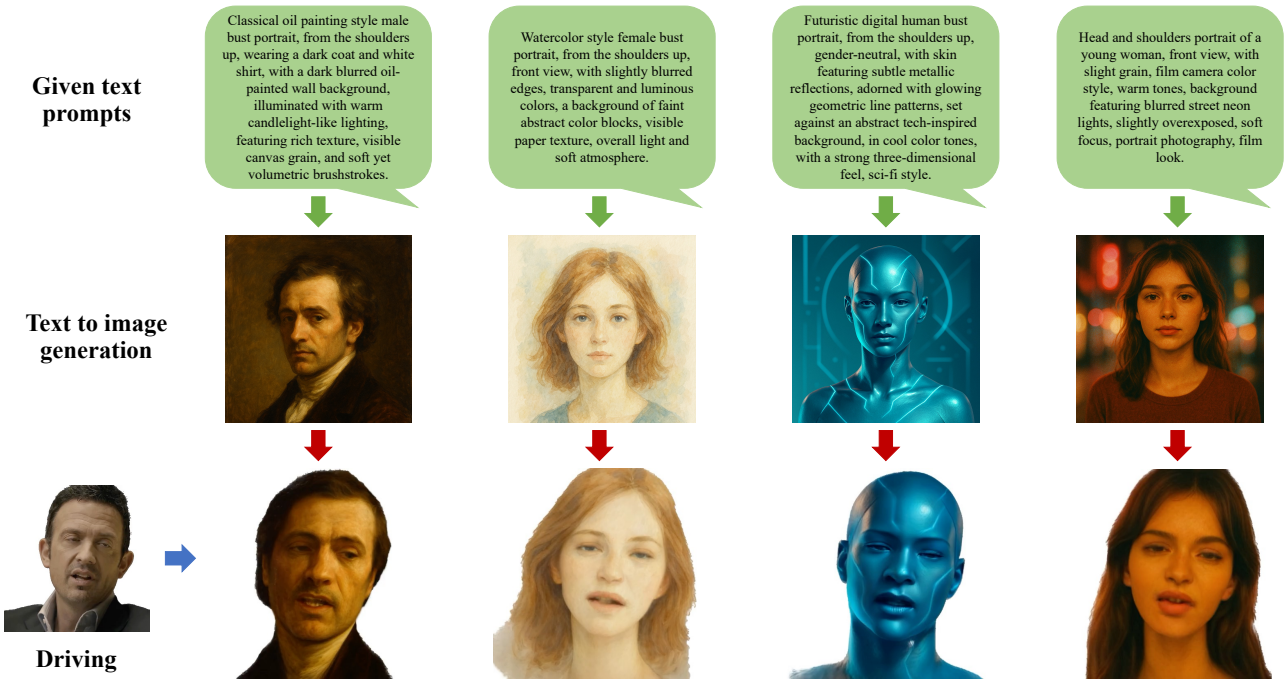


Figure S10. **Visualization of text-to-head-avatar generation.**

References

[1] Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 4

[2] 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. 4

[3] Yu Deng, Duomin Wang, and Baoyuan Wang. Portrait4d-v2: Pseudo multi-view data creates better 4d head synthesizer. *arXiv preprint arXiv:2403.13570*, 2024. 4

[4] Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12736–12746, 2023. 4

[5] Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Pixel3DMM: Versatile screen-space priors for single-image 3d face reconstruction. In *The Fourteenth International Conference on Learning Representations*, 2026. 2

[6] 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*, pages 1–13, 2025. 4

[7] 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), 2023. 4

[8] Tobias Kirschstein, Javier Romero, Artem Sevastopolsky, Matthias Nießner, and Shunsuke Saito. Avat3r: Large animatable gaussian reconstruction model for high-fidelity 3d head avatars. *arXiv preprint arXiv:2502.20220*, 2025. 4

[9] 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, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 5

[10] 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. 4

[11] Dongwei Pan, Long Zhuo, Jingtan Piao, Huiwen Luo, Wei Cheng, Yuxin Wang, Siming Fan, Shengqi Liu, Lei Yang, Bo Dai, et al. Renderme-360: a large digital asset library and benchmarks towards high-fidelity head avatars. *Advances in Neural Information Processing Systems*, 36, 2024. 4

[12] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 2

[13] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ArXiv preprint*, 2021. 2

[14] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. 2

[15] Felix Taubner, Ruihang Zhang, Mathieu Tuli, and David B. Lindell. CAP4D: Creating animatable 4D portrait avatars with morphable multi-view diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5318–5330, 2025. 4

[16] 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*, 2025. 2, 4

[17] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022. 4

[18] Xiaochen Zhao, Jingxiang Sun, Lizhen Wang, Jinli Suo, and Yebin Liu. Invertavatar: Incremental gan inversion for generalized head avatars. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–10, 2024. 4

[19] Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. Idol: Instant photorealistic 3d human creation from a single image. *arXiv preprint arXiv:2412.14963*, 2024. 4