IM-Portrait: Learning 3D-aware Video Diffusion for Photorealistic Talking Heads from Monocular Videos

Supplementary Material

In this supplementary material, we provide more information about implementation details (Sec. A) and additional results, including more qualitative and quantitative comparisons with other methods (Sec. B, Sec. C, Sec. D), evaluations on out-of-distribution data (Sec. E), evaluations on pseudo ground truth images (Sec. F), evaluations on side view renderings (Sec. G) and comparison with "2D diffusion + depth" (Sec. H). Additionally, we add further discussions on limitations (Sec. I).

A. Implementation Details

Network Architecture We build our model architecture based on Space-Time U-Net [1]. We first downsample 512×512 images into 128×128 feature maps using one convolution layer. Then the feature map is downsampled into 64×64 by another convolution layer. In the color branch, both downsample blocks and upsample blocks consist of 2 Convolution-based Inflation blocks for resolution 64 and 3 Convolution-based Inflation blocks for resolution 32 and 16. We also add spatial and temporal self-attention layers inside each block with resolution 16. In the geometry branch, each resolution only has one Convolution-based Inflation block. We apply cross-attention layers between the source image features and every block with 32 resolution in the color branch. The output feature map is firstly upsampled by pixel shuffle from 64×64 into 128×128 and processed by a convolution layer. In the color branch, the result 128×128 feature map is then upsampled into outputs of 512×512 by another pixel shuffle and convolution layer. In the geometry branch, 128×128 feature map is firstly upsampled into 256×256 and then upsampled into 512×512 by pixel shuffle and convolutions. The color branch predicts the estimated frontal/residual RGB image of the Multiplane Images (MPIs), while the geometry branch predicts alpha images of all planes. Since the geometry branch does not take reference image as condition, reference image features are passed through zero-convolutions between the color branch and the geometry branch.

Training Details In the data preparation stage, the reference and target pair of images are randomly selected from the video following [7, 11], where the camera poses are obtained through 3DMM fitting.

During the model bootstrapping described in main paper Sec. 3.3, we only train the color branch with loss $\mathcal{L}_\theta^{\text{mpi}}$ Eq. (6). After bootstrapping, we freeze the color branch except the output layer. Then, we train the geometry branch, the output layer of the color branch, and all the zero convolution layers with both losses $\mathcal{L}_\theta^{\text{mpi}}$ Eq. (6) and $\mathcal{L}_\theta^{\text{side}}$ Eq. (7). In practice, in each iteration, we randomly select one loss from the two to compute the gradients for updating the model. We select the loss $\mathcal{L}_\theta^{\text{mpi}}$ (Eq. (6)) with a probability of 0.8, where MPIs are constructed using target cameras. We select the loss $\mathcal{L}_\theta^{\text{side}}$ (Eq. (7)) with a probability of 0.2, where MPIs are constructed using reference cameras, and target cameras serve as side-view cameras.

During training, the weight of LPIPS loss is set to 0.1 while the mask loss has a weight of 0.01. The mask loss is formulated as L2 loss since the matting mask has soft boundary. The depth smoothing loss is calculated by a Laplacian kernel and has a weight of 0.01. The disparity loss has a weight of 0.001. The drop rate of first frame is 0.7 while the drop rate of the reference portrait is 0.1.

Rendering Details Following [13], we choose Multiplane Images (MPIs) as our scene representation. We set near and far planes of MPIs adaptively based on the distance r from the MPI frontal camera to the head joint of the parametric model. We place the near plane at $r - 0.15$ while the far plane at $r + 0.05$.

During inference, we handle camera via two methods: 1) Rendering the generated MPIs into explicit cameras $\{\phi^{\text{side}}\}$; 2) Rasterizing the 3DMM UV coordinate maps into freely selected MPI frontal cameras as the diffusion controlling signals $\{\mathbf{C}\}$. In our experiments, all the side view renderings, stereo renderings and rendering speed measurements are conducted through the first method. When we render novel views through the first method, novel camera views must stay within a specifiv range of the MPI frontal camera to avoid pose deviations affecting rendering.

Long Video Classifiers-Free-Guidance (CFG) During inference, we assign two different CFG scales to first frame condition and reference portrait condition following Brooks *et al*. [2]. The guidance scale of the reference portrait is 1.5. The guidance scale of the first frame is linearly decreased from 1.0 to 0.0 from the beginning to the middle of the video clip while the remaining 16 frames in the clip shares a scale of 0.0.

We show the ablation on first frame's classifier-free guidance in Fig.A. We compare our scheduled scale with constant guidance scale of 1.5. As show in Fig.A, artifacts become more pronounced as additional frames are generated.

Method	VFHQ				Self-Collected Dataset			
	L1 ↓	LPIPS ↓	FID ↓	FVD ↓	L1 ↓	LPIPS ↓	FID ↓	FVD ↓
Face-V2V	0.051	0.290	71.58	226.42	0.050	0.208	47.13	405.34
EmoPortrait	0.064	0.251	48.96	292.01	0.071	0.236	52.85	506.70
Portrait4d-v2	-	-	54.26	329.58	-	-	54.46	553.42
Follow-your-emoji	0.057	0.198	32.40	214.29	0.060	0.181	34.94	364.36
X-Portrait	0.061	0.204	26.22	211.62	0.062	0.185	32.77	499.67
Ours	0.054	0.204	33.10	201.41	0.048	0.174	36.98	342.78

Table A. **More comparisons on VFHQ and self-collected dataset.** We further compare our method with baselines on both VFHQ and self-collected dataset. Our method achieves comparable image quality and the best video quality (FVD), which is consistent with quantitative comparison in the main paper.

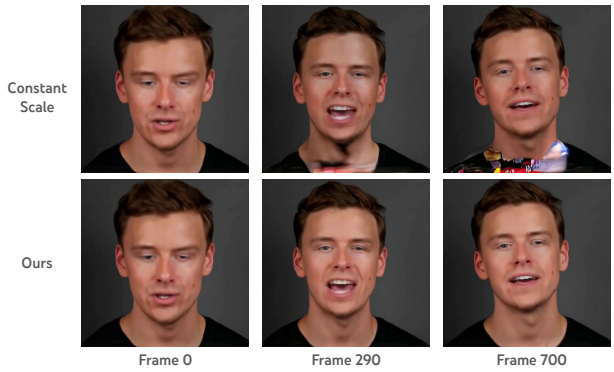

Figure A. **Ablation on first frame's classifier-free guidance.** We compare the generated results using a constant classifier-free guidance scale for the first frame versus our scheduled guidance scale.

However, using our scheduled scale helps prevent the accumulation of errors.

B. Additional Qualitative Results

We show more qualitative results on HDTF and Talkinghead1kh datasets in Fig.C. Face-V2V [9] shows good consistency against the driving signal and the reference portrait, while losing details due to its low output resolution. EmoPortrait [5] gives sharp renderings but suffers from identity shifts. Portrait4D-v2 [4] shows better identity consistency, but losses subtle facial movements. XPortrait [11] and Follow-your-emoji [7] generates sharp and high quality frames but suffers from expression and head pose misalignment against the driving signal. Our method generates sharp results that faithfully following driving signals while keeping the identity of generated avatar aligned with the reference portrait. For more comparisons on video quality, please refer to our project page in supplementary material.

Before evaluation, we fix the cropping of all the evaluation data with a tight cropping around head as show in Fig.B.

C. Additional Quantitative Results

We show more qualitative results on VFHQ [10] and self-collected dataset in Tab.A. For VFHQ dataset, we sample

Figure B. **Fixed cropping.** We preprocess some portraits with a tighter cropping around head.

Method	HDTF		Talkinghead1kh	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Face-V2V	27.00	0.865	24.17	0.823
EmoPortrait	22.35	0.794	19.59	0.731
Follow-your-emoji	24.16	0.811	21.50	0.759
X-Portrait	24.54	0.820	21.82	0.767
Ours	24.83	0.833	22.43	0.777

Table B. **PSNR and SSIM on HDTF and Talkinghead1kh datasets.** PSNR and SSIM rankings align with the L1 score rankings in Table 1 of the main paper.

the first 100 frames from all evaluation identities. The self-collected dataset has 50 identities, while each has 32 frames. Consistent with the other two datasets in the main paper, ours achieves the best video quality (FVD) while being competitive in other image-based metrics on VFHQ dataset. On self-collected dataset, ours achieves the best L1, LPIPS and FVD scores. The cross-dataset evaluations following [3] reinforces the generation quality and generalizability of our model.

D. Further Explanations on Comparisons

In Tab.1 in main paper, we report L1 scores instead of PSNR and SSIM, as these latter metrics are more suitable for pixel-wise alignment in neural rendering tasks, while LPIPS better aligns with perceptual image quality and tolerates some misalignment [8]. As shown in Tab. B, PSNR and SSIM closely correlate with L1 scores reported in Tab.1, where our method performs worse than non-generative methods like Face-V2V but outperforms others. Due to the gener-

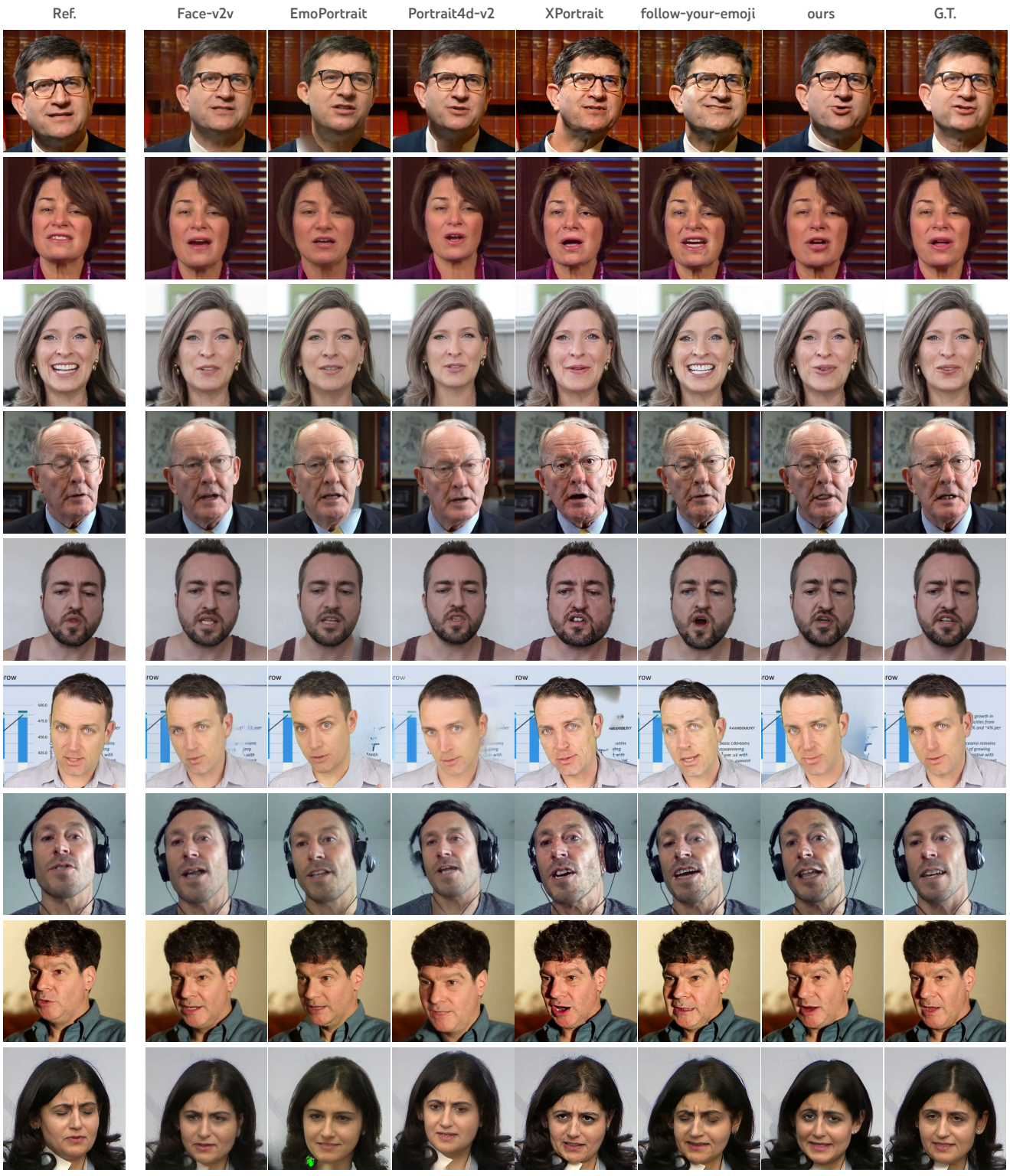

Figure C. **Talking head results.** We show more results from previous work and our method.

Method	FVD ↓	FID ↓
Face-V2V [9]	134.92	23.29
EmoPortrait [5]	197.87	34.54
Portrait4d-v2 [4]	185.42	28.86
Follow-your-emoji [7]	154.14	20.88
X-Portrait [11]	199.27	21.03
Ours	107.90	18.00

Table C. **Evaluation of rendering quality on cross-identity reenactment.**

Method	FID ↓
View Point	Random within $\pm 5^\circ$
Face-V2V [9]	22.27
EmoPortrait [5]	27.71
Portrait4d-v2 [4]	27.83
Ours 2D + Depth [12]	19.72
Ours	18.12

Table D. **Evaluation of side view rendering quality**. Our method outperforms all the baseline on rendering quality.

ative nature of diffusion models, our method has slightly lower L1 scores but achieves better LPIPS, as it tolerates misalignments. As shown in both Tab.1 and Tab.B, while sharing network capacity for geometry and learning without 3D data results in a slightly lower FID compared to purely 2D video methods (e.g., Follow-your-emoji, X-Portrait), we still achieve competitive results across all image quality metrics and the best FVD score, which evaluates both image and temporal quality.

E. Evaluation on Out-of-distribution Data

We further test our model on out-of-distribution data. Despite trained on real world talking-head videos, our model still generalizes to stylized portraits generated by Stable-Diffusion 3 [6]. Results are shown in Fig.F.

We also evaluate our model on large head pose movements between reference portraits and target expressions through the second method mentioned in **Rendering Details** of Sec.A. As shown in Fig.D, our model generalize to large head pose movements. However, in some extreme cases, e.g. reference portrait only includes side view of the head, our model may struggle to generate aligned results with ground truth.

F. Evaluation on Pseudo Ground Truth Images

In Fig.E, we show some examples of pseudo ground truth images mentioned in Reference-Target Alternating Training section. The generated pseudo ground truth images aligns with target expressions. Note that the pseudo images are only used in training when noise is large, so they don't need to be sharp and clear as long as the global structure is correct.

G. Evaluation on Side Views

We evaluate our method on side view renderings with HDTF dataset. We select Face-V2V [9], EmoPortrait [5], Portrait4D-v2 [4] as baselines since they have 3D representations or implicit 3D feature volumes that support explicit camera control. On the other hand, the diffusion-based baselines [7, 11] only take in images or eyes and mouth landmarks, thus do not enable explicit camera control. More specifically, we render several frames by applying a random horizontal viewpoint change (rotate the camera around a fixed look-at point) uniformly sampled from $\pm 5^\circ$, and compute the FID score. All the evaluations are performed in a self-reenactment way. The evaluation results are shown in Tab.D, where our method outperforms all the baselines on FID score. For more visualizations of our method's side view renderings, please refer to the project page in supplementary material.

H. Compare With 2D Diffusion and Depth Prediction

We also build up a baseline with the 2D video diffusion variant of our model and a monocular depth predictor [12] to lift the images into 3D. To be specific, we firstly run depth predictor on all the generated frames. The predicted monocular depth is then aligned with depth of 3DMM mesh by calculating scale and shift. Since the 3DMM mesh doesn't have hair and cloth geometry, we sample depth values on detected facial landmarks as inputs for alignment. Before rendering, we first project aligned depth into 3D then connect 3D points of adjacent pixels to create triangles. We also use foreground matting mask to filter out background vertices. During rendering, we use the same background image used by our method.

We compare our method with this baseline and show FID comparisons in Tab.D. We also show qualitative comparisons in Fig.G. Using monocular depth estimator to lift 2D diffusion to 3D suffers from blurry results in occluded regions. Moreover, this baseline also fails to render detailed geometry such as hair as shown in the last row of Fig.G.

I. Additional Discussions on Limitations

As discussed in the main paper, our model mainly focuses on realistic foreground human rendering, while background is generated by an inpainting network following [5]. Thus, the generated results do not include animation in the background. To achieve more realistic generated results for background, future works could train an auxiliary network to animate the inpainted background. Furthermore,

Figure D. **Evaluation on large head pose movements.** We show generated results in the MPI frontal camera where the target head poses deviate largely from reference portrait's head poses. In the last row, generated result doesn't align with ground truth since the reference portrait doesn't contain any information about the other side of the character.

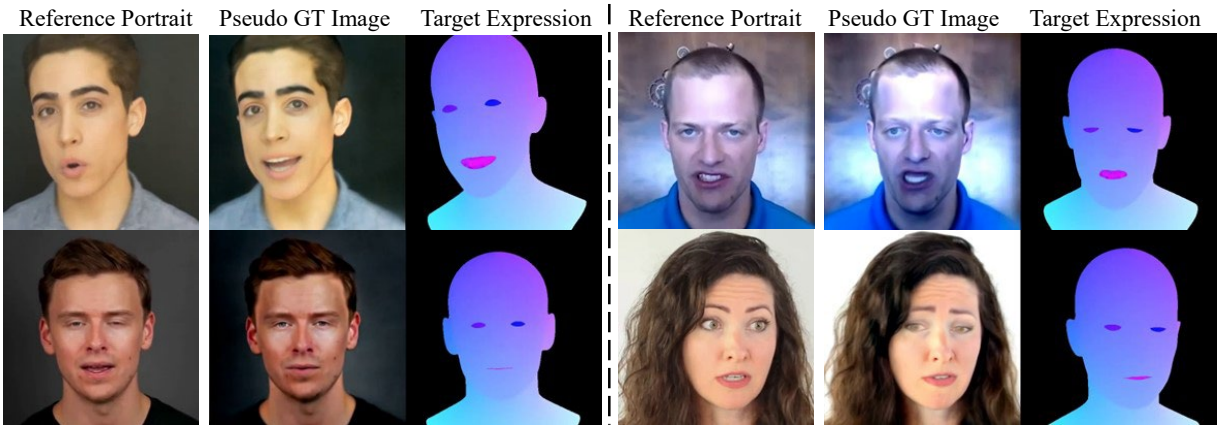

Figure E. **Evaluation on pseudo ground truth images.** We show pseudo ground truth images used in the **Reference-Target Alternating Training**. Pseudo images are only used when large noise is sampled. Though there exist drifting in the color tone, the global structure of generated result aligns with target expressions.

Figure F. **Evaluation on synthetic images.** We show generated results when taking synthetic images as reference images. Though trained on a real-world distribution, our model is able to generalize to synthetic images even with a huge gap between the color distribution.

as shown in Fig.D, large self-occlusions in the reference portrait may lead to degraded results in the unseen facial region. Future work could focus on introducing facial symmetry priors to mitigate this issue.

References

[1] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024. 1

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1

[3] Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and

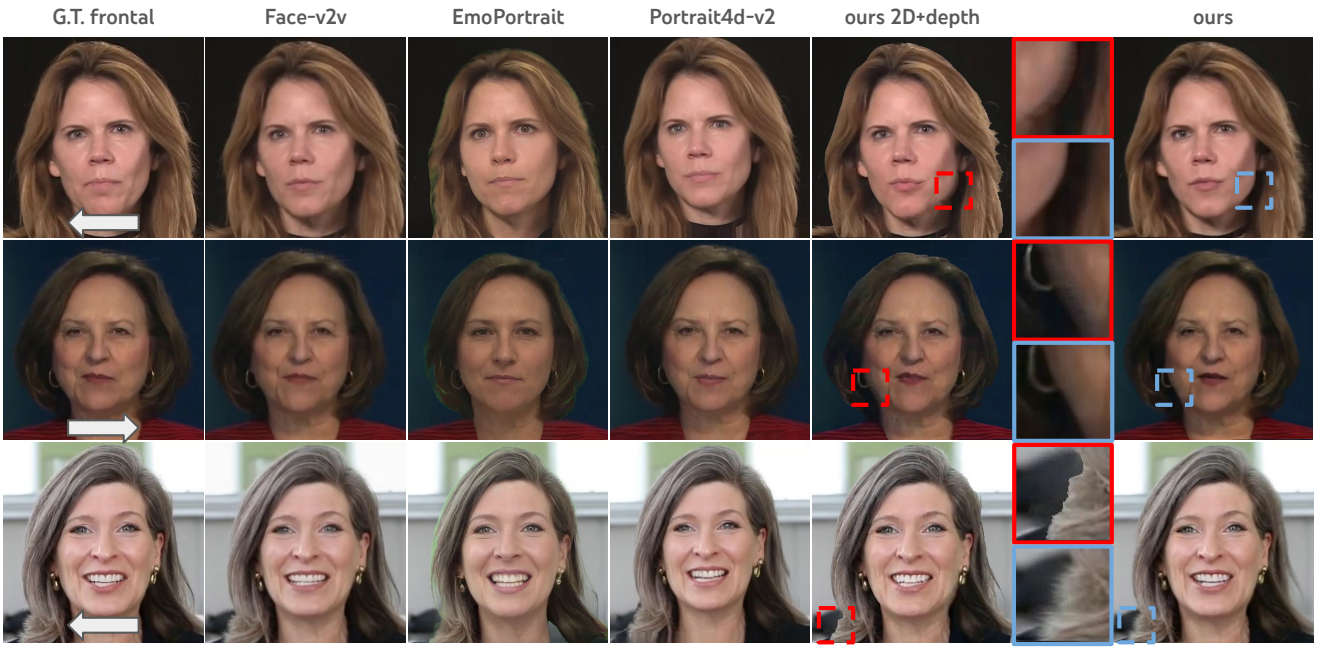

Figure G. **Side view rendering comparisons.** We annotate rotation directions by arrows in the ground truth frontal images.

facial expressions retargeting with identity-aware diffusion. In *ICML*, 2024. 2

[4] Yu Deng, Duomin Wang, and baoyuan Wang. Portrait4d-v2: Pseudo multi-view data creates better 4d head synthesizer. *arXiv*, 2024. 2, 4

[5] Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, and Maja Pantic. Emoportraits: Emotion-enhanced multimodal one-shot head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8498–8507, 2024. 2, 4

[6] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 4

[7] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. *arXiv preprint arXiv:2406.01900*, 2024. 1, 2, 4

[8] Alex Trevithick, Matthew Chan, Michael Stengel, Eric R. Chan, Chao Liu, Zhiding Yu, Sameh Khamis, Manmohan Chandraker, Ravi Ramamoorthi, and Koki Nagano. Real-time radiance fields for single-image portrait view synthesis. In *SIGGRAPH*, 2023. 2

[9] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021. 2, 4

[10] 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. 2

[11] You Xie, Hongyi Xu, Guoxian Song, Chao Wang, Yichun Shi, and Linjie Luo. X-portrait: Expressive portrait animation with hierarchical motion attention. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 1, 2, 4

[12] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 4

[13] Xiaoming Zhao, Fangchang Ma, David Güera, Zhile Ren, Alexander G Schwing, and Alex Colburn. Generative multiplane images: Making a 2d gan 3d-aware. In *European conference on computer vision*, pages 18–35. Springer, 2022. 1